Debiasing thinking among non-WEIRD reasoners

Esther Boissin*[1], Mathilde Josserand[2], Wim De Neys[1,3], Serge Caparos[4,5]

[1] Université Paris Cité, LaPsyDÉ, Paris, France

[2] Université Lumière Lyon 2, Laboratoire Dynamique Du Langage, Lyon, France

[3] CNRS, Paris, France

[4] Université Paris 8, DysCo lab, Saint-Denis, France

[5] Institut Universitaire de France, Paris, France

Abstract

Human reasoning has been shown to be biased in a variety of situations. While most studies have focused on samples of WEIRD participants (from Western Educated Industrialized Rich and Democratic societies), the sparse non-WEIRD data on the topic suggest an even stronger propensity for biased reasoning. This could be explained by a competence issue (people lack the ability to integrate logical knowledge into their reasoning) or a performance issue (people possess the logical knowledge but do not know it is relevant). We addressed this question using a debiasing paradigm with the base-rate task on a sample of non-industrialized people, the Himba of Namibia. After a short training, most participants were debiased, lending credence to the performance account. Debiasing was however to some extent boosted by schooling and living environment suggesting that competence also plays a role (in that more acquired knowledge allows for more training benefits). Results imply that debias interventions can be successfully employed to boost sound reasoning around the world.

Keywords: reasoning; heuristics & biases; debiasing paradigm; Himba of Namibia

Introduction

Decades of research on decision-making have shown that we do not always reason correctly. For instance, if in a sample of 1000 people, with 995 females and 5 males, a randomly selected person is described as liking beer and football, many people would assume that this person is male based on stereotypical associations cued by the descriptive information (“men like beer and football”). If we only consider the description, that may be a reasonable assumption. In general, there might be more males than females who like beer and football. However, some females also like this, and there were far more females (995) than males (only 5) in the sample. The base-rate information should thus push the scale to the “female” side. Unfortunately, people often ignore base-rate information and over rely on their stereotypical prior beliefs (e.g., Kahneman & Tversky, 1973).

This tendency to favour personal beliefs over logical and mathematical principles (e.g., base rates) has been shown to bias reasoning in various situations (Evans & Over, 1996; Kahneman, 2011; Morewedge & Kahneman, 2010). However, as in many fields, most previous work has been conducted on samples of WEIRD (Western Educated Industrialized Rich and Democratic societies, Henrich et al., 2010) participants, representing only about 15% of the world population (Arnett, 2008). Interestingly, the sparse data regarding non-WEIRDs' reasoning performance often suggest an even stronger propensity to base their reasoning on personal beliefs (Luria, 1971, 1976; Scribner, 1977; Trémolière et al., 2022, but see also Maddux et al., 2010). An iconic example comes from Luria's work in which illiterate Uzbekistan people showed very poor performance on deductive reasoning problems such as "In the Far North, where there is snow, all bears are white. Novaya Zemlya is in the Far North and there is always snow there. What color are bears there?" (e.g., one participant stated: "I've seen a black bear, I've never seen any others", Luria, 1976, p. 109). Similar poor accuracies were observed with African and central American illiterate people (Cole & Bruner, 1971; Scribner, 1977; Sharp & Cole, 1975).

Although these studies typically indicate that non-WEIRD groups show more biased reasoning, the precise reason is less clear. One possibility is that non-WEIRD people lack the ability to integrate logic into their reasoning. As Luria (1971, 1976) famously argued, logical or abstract reasoning could be intrinsically linked with literacy. In the absence of literacy, people's responses would be necessarily driven by their immediate practical experiences; They might be incapable of decontextualizing the information contained in the problem and integrate more abstract, logical principles (Luria, 1971; Nisbett et al., 2001; Scribner & Cole, 1981).

Alternatively, the more biased responding in non-WEIRD reasoners might be explained by a performance, rather than a competence deficit. That is, while illiterate people from non-WEIRD societies may possess the necessary logical knowledge and may have the ability to integrate it into their reasoning, they would simply not realize that it is useful for solving the problem at hand, particularly for unfamiliar tasks.

Interestingly, indirect evidence from studies with WEIRD samples supports the performance hypothesis. Recent debias studies have shown that single-shot interventions, consisting of a brief reminder of the correct solution, can easily boost WEIRD participants' reasoning (e.g., Boissin et al., 2021, 2022; Claidière et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014; see also Isler & Yilmaz, 2022; Isler et al., 2020; Yilmaz & Saribay, 2017a). Given that these interventions are extremely short—often lasting no more than 5 minutes—the finding that this nevertheless suffices to remediate participant's reasoning suggests that the bias does not result from a lack of knowledge or so-called "mind" or "storage" gap (e.g., De Neys & Bonnefon, 2013;

Stanovich, 2011). Clearly, it is unlikely that a simple reminder suffices to learn the underlying logical principles ex nihilo. That is, if the short explanation works and allows people to respond correctly, it might be because the critical knowledge was already acquired (Boissin et al., 2021, 2022; Mata, 2020). Its relevancy simply needed to be stressed.

These debiasing results might imply that a similar performance account may also be at play in non-WEIRD samples. Some non-WEIRD studies even provide initial support for this interpretation (e.g., Dias et al., 2005 and see Mercier, 2011). Notably, one study among indigenous Mayan groups showed that non-WEIRD reasoners can accurately perform adapted probabilistic evaluations (Fontanari et al., 2014), suggesting that they possess the basic probabilistic knowledge essential for correct responding in various reasoning tasks (e.g., the base-rate task). Moreover, Castelain et al. (2016) demonstrated that non-WEIRD reasoners can perform better when core logical principles are explicitly brought to their attention. Nonetheless, evidence is still limited and further validation of the performance hypothesis within non-WEIRD samples is required.

In the present study, we tested the performance-competence question concerning non-WEIRD samples' reported poorer reasoning performance directly, by adopting a debiasing paradigm similar to Boissin et al.'s (2022) study with a non-WEIRD sample: The Himba of Northern Namibia (Africa). The Himba are semi-nomadic herders with limited exposure to western culture, including limited participation in western-style education (schooling) and low levels of industrialization. They predominantly speak the "Otjiherero" dialect, often live in small, collectivist villages (up to 40 adults), and do not use written language (Trémolière et al., 2022) or monetary exchanges (Cameron, 2013). Figure 1 illustrates a typical Himba village.

We tested the reasoning of Himba people using the base-rate neglect task (e.g., see the introductory example) which has often been used to study biased thinking in WEIRD samples. We used adapted, familiar material in which typical stereotypical Himba beliefs were pitted against logical base-rate information. We measured the debiasing intervention effect by contrasting pre- and post-intervention performance. The intervention was structured as in previous WEIRD studies in which the role of base-rates on likelihood estimation was briefly explained. In line with previous non-WEIRD studies, we expected accuracy to be very low before training. If a performance deficit explains the reasoning errors, we expect the Himba to benefit from the intervention and show higher post-training accuracy (just as in WEIRD samples). On the contrary, if illiterate Himba individuals do not have the basic competence to use logical considerations to overrule personal beliefs, the short intervention should not help them, and they should show stable (low) performance across the study.

Furthermore, by taking advantage of natural ecological variation in the Himba region we could also test factors that may modulate the efficiency of the training. That is, although most adult Himba are illiterate and have received no formal schooling, schooling is not completely absent in the region. The constitution of Namibia makes primary education mandatory and compels the state to establish and maintain free schools. This led to the implementation of standardised curricula in which basic reading, writing and maths skills are taught (see the 2017 Namibian Ministry-of-Education report). Although not always instantiated in practice, through this initiative at least some Himba villagers have received some minimal formal teaching. Relatedly, in the vicinity of the traditional Himba villages, the town of Opuwo (~30,000 inhabitants; Mwinga et al., 2022) is home to some Himba individuals. In contrast with the Himba villages, Himba urban settlements in Opuwo can be equipped with some modern supplies (electricity, running water, mobile phones, and basic furniture). There is also access to public services (e.g., health centre, town hall) and private ones (e.g., food stores, pharmacies, banks). The Himba living in Opuwo regularly engage in the market economy (Cameron, 2013), by buying food and selling jewellery, which implies some familiarization with numerical processing (e.g., Carraher et al., 1985, 1987; Nunes et al., 1993; Saxe, 1988). By testing both Himba participants living in traditional villages and in the town of Opuwo—and recording their schooling record—we could *a posteriori* explore the impact of living environment (i.e., an industrialization proxy) and schooling on the debiasing effect, within the same country, language, and ethnic population.

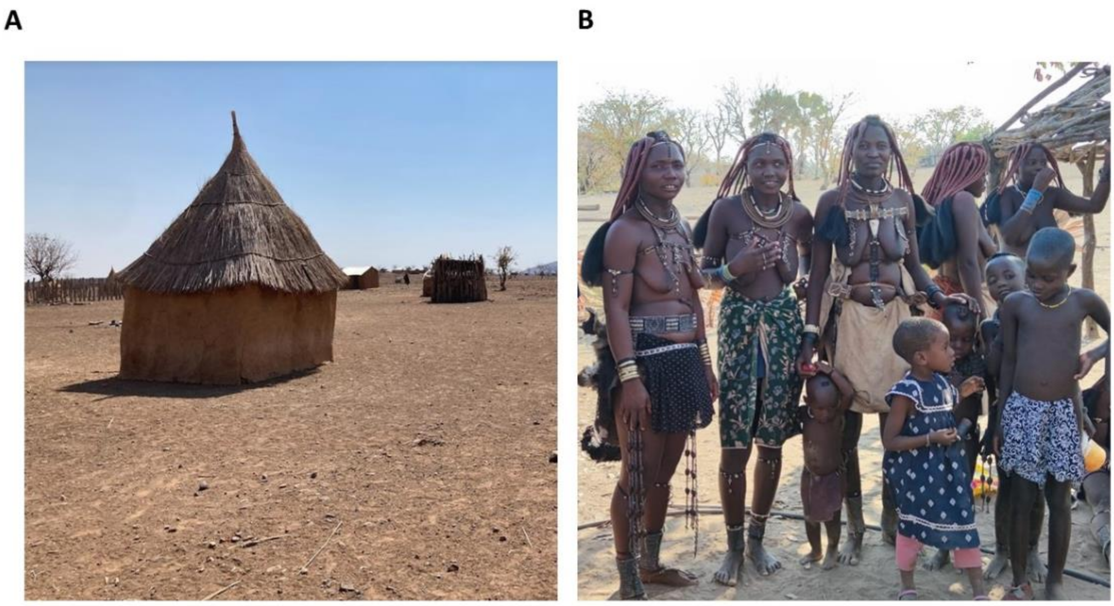

Figure 1. A house in a typical Himba village (A) and Himba women and children (B).

Methods

Participants

A total of 84 Himba individuals (50 females, mean estimated age = 26.6 years, SEM = 1.0, range from 16 to 55 years, see Supplementary Material Section A for the age-estimation method) were recruited during a 2-month research visit to Namibia in October and November 2021. To our knowledge, none of the participants had ever participated in experimental research.

Among the 84 participants, 49 were living in traditional Himba villages and 35 were living in town (i.e., Opuwo). The villages were located at about 60 km from the town of Opuwo, in the area of Otuazuma. Of the tested villagers, 23 (47%) had gone to school (average number of years of schooling = 5.3 years, SEM = 0.6, range between 1 and 9 years). Among the Himba who lived in town, 17 (49%) had gone to school (average number of years of schooling = 6.9 years, SEM = 0.7, range between 1 and 9 years).

In total, 74 participants were assigned to the critical training group. Practical on-site testing and recruitment complications prevented testing a larger sample. Our primary goal was to test a sample of approximately 50 villagers in the training group—equivalent to the sample size in the WEIRD debias studies of Boissin et al. (2022). Near the end of the data collection, we also allocated 10 participants to a control group in which the participants received no training intervention (4 lived in town; 6 lived in the villages). Note that given the non-random group allocation and small numbers of participants in the control group, results should be interpreted with caution. However, our rationale was that the control group would still allow us to explore whether mere habituation and test repetition could lead to debiasing effects.

All participants received compensation for their participation (they could choose 3 gift items among the following: Maize meal, sugar, vaseline or soap—roughly equivalent to a 5-euro value). The study was approved by the CER-Paris Cité Ethics Committee (Project ID: 2021-88-********).

Materials

The study consisted of three blocks presented in the following order: A pre-intervention, an intervention, and a post-intervention block. In total, each participant had to solve 10 problems during the pre-intervention block (four conflict, four no-conflict, and two neutral problems, see below). In the post-intervention block the same number of problems, but with different content, were used along with two extra control heuristic problems (see further). At the end of the main study, we asked participants to rate the material to check that it cued the intended stereotypes and beliefs. All material was translated from English to Otjiherero by two different native translators who discussed and agreed upon each translation. All instructions and task problems were orally presented in Otjiherero by one translator. All the problems are presented in the Supplementary Material Section B.

Base rate problems. In total, 19 base-rate problems were created, based on the Bago and De Neys (2017) presentation format. Participants always received base-rate information about the composition of a new village (e.g., "In this new village, there are 8 men and 2 women") followed by the description of one villager which was designed to cue a stereotypical association (e.g., "This randomly chosen person looks after children"). As shown in Figure 2, the problem was presented with a photo illustrating the composition of the village. For the example above, we presented 8 different male portraits and 2 different female portraits. The base rate was consistently 8/2. The portraits, the base rates, and the description were presented together, on a tablet screen. Participants' task was to indicate to which group the random villager most likely belonged to (e.g., men or women in the example above). Participants gave their answer by touching one of two target photos on the touchscreen, which evoked the possible answers (e.g., to distinguish between men and women, we used pictures of male and female torsos). The position of each response photo was counterbalanced between participants and items. Figure 2 illustrates the full problem format.

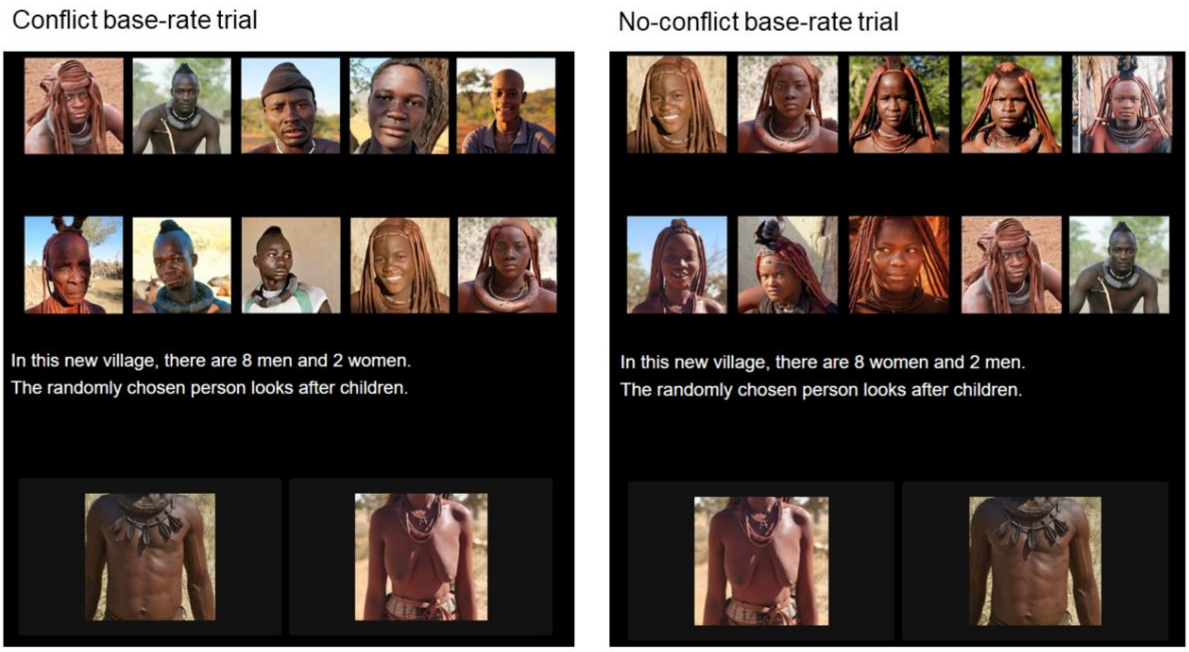

Figure 2. Illustration of a conflict and no-conflict base-rate trial.

In the pre- and post-intervention blocks, we presented four critical "conflict" items, and four "no-conflict" items. In the conflict items, the base rate probabilities and the stereotypical information cued conflicting responses (see example above). In the no-conflict items, they both cued the same response (i.e., the description triggered a stereotypical trait of a member of the largest group, see Figure 2). In the above example, the correct response to the conflict item is “a male”, while for the no-conflict items, the correct response is “a female”. These no-conflict problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy to no-conflict problems (Bago & De Neys, 2019). In addition, the no-conflict items allow us to control for a “reversed heuristic” account (Boissin et al., 2022). That is, it is possible that our training does not succeed in getting participants to give more weight to the base rates per se but simply have them assume that they are being presented with a counter-intuitive task in which the right answer is always the opposite of the cued heuristic/stereotypical response (e.g., “pick what you don’t think it is”). This would lead to selection of the correct response on conflict problems. However, it should also result in floored performance on the no-conflict problems (on which the heuristic response is also correct).

Two sets of 16 unique items (8 pre-intervention and 8 post-intervention block items) were used for counterbalancing purposes. For each block, the conflict problems in one set were the no-conflict problems in the other, and vice-versa (i.e., the base-rates were reversed). Participants were randomly assigned to one of the two sets. Consequently, none of the pre- and post-intervention problem contents was repeated within-subjects (i.e., participants saw a total of 16 different items each with a unique stereotypical association).

Neutral problems. We also presented four neutral base-rate problems. These problems were designed such that they did not cue any stereotypical association (i.e., the descriptive information was not diagnostic of a particular group). Here is an example of a neutral base-rate problem:

In this new village, there are 8 women and 2 men.
The randomly chosen person has two hands.
Is the randomly chosen most likely to be:
- *a woman*
- *a man*

The neutral base-rate items are traditionally used to track people's knowledge of the underlying logical principles or “mindware” (Stanovich, 2011). Usually, reasoners have little trouble solving them (e.g., De Neys & Glumicic, 2008; Frey & De Neys, 2017). Here the neutral problems allowed us to explore whether a potential learning effect on conflict base-rate problems in which the reasoner needs to discard a conflicting stereotypical association, leads to a more general performance boost on other untrained base-rate problems (e.g., Boissin et al., 2022). Boissin et al. (2022) did not observe such transfer in their debias studies with WEIRD samples.

Control heuristic problems. We also presented two control heuristic problems during the post-intervention block. The control heuristic problems aimed to check that reasoners did not apply a new heuristic strategy such as "always pick the largest group" and no longer regard the description after the training intervention. This strategy would lead reasoners to systematically give a correct response for conflict and no-conflict problems, but it would also result in the systematic neglect of the descriptive information, which could be suboptimal. Ultimately, we want reasoners to be able to combine both base rate and stereotypical description information to make a correct inference rather than to simply neglect the description in favour of the base rates. Indeed, some descriptions may be extremely diagnostic of a particular group of people and should take precedence over the use of probabilities. For example, the description "the randomly selected person is pregnant" is extremely diagnostic of the group "female", regardless of the base-rates in the sample. The control items exploited this feature. Here is a full example:

In this new village, there are 8 men and 2 women.
The randomly chosen person is pregnant.
Is the randomly chosen most likely to be:
- *a man*
- *a woman*

Thus, here, "the randomly selected person is pregnant" will always be a woman, although there are more men in the sample. If training leads to a simple description neglect, performance on these problems should be floored.

Intervention block. During the intervention block, the participants tried to solve one additional conflict base-rate problem (i.e., "In this new village, there are 8 women and 2 men. The randomly chosen person digs wells"), twice[1]. In the training group, participants were explained the correct solution after responded. Participants in the control group received no such explanation. The following example illustrates the explanation.

> *"The correct answer to the problem is that the randomly chosen person is most likely a woman.*
>
> *Many people think it is a man, but this answer is wrong. Most people base their answer simply on the given description of the randomly chosen person. If this were all information you got, this answer would be correct, as it is likely that there are more men than women who dig wells.*
>
> *However, in the problem, you also have information about the composition of the village from where the person has been randomly chosen. You were informed that the chosen person was drawn randomly from a new village with 8 women and 2 men. Since there are more women than men, it becomes more likely that the randomly chosen person is a woman than a man. After all, although men might in general dig wells more than women, there are also women who dig wells. If you combine this with the vastly larger number of women in the village, it will be more plausible that you're dealing with a woman who digs wells."*

[1] During the intervention, the same problem was presented (and then explained to the training group) twice. This was due to practical constraints, as we had limited pre-tested materials available.

The explanation was based on the same general principles that were adopted by Boissin et al. (2021, 2022): The explanation was as brief and simple as possible to prevent fatigue or disengagement from the task, and explicitly stated both the correct response and the typical incorrect response. No personal performance feedback (e.g., “Your answer was wrong”) was given in order to avoid promoting feelings of judgment (Trouche et al., 2014). Also, to avoid inducing mathematical anxiety, the explanation never mentioned a formal algebraic equation (Hoover & Healy, 2017). Note that various debiasing interventions have been explored and have been shown to be effective at debiasing WEIRD reasoners (e.g., Adame, 2016; Hirt & Markman, 1995; Hogarth, 2008; Isler et al., 2020; Isler & Yilmaz, 2022; Mata, 2020; Mata et al., 2013; Yilmaz & Saribay, 2017a, 2017b). We opted for the Boissin et al. format for mere practical reasons (see also Discussion).

Rating task. At the end of the main study, participants were asked to rate the content to ensure it was familiar and that it elicited the intended stereotypes and beliefs. Participants were asked to indicate to which degree they agreed with statements on a 5-points-Likert-scale ranging from “Don’t agree at all” to “Absolutely agree” (Trémolière et al., 2022). Similar to Raoelison et al. (2021), two statements were presented at the same time, which were associating each group with a description (e.g.,” Women dig wells” and ”Men dig wells”). In total, 25 pairs of statements were rated, corresponding to the 16 pairs of conflict and no-conflict items in the reasoning task, 4 pairs of neutral items, 1 intervention item pair, 2 pairs of practice items, and 2 pairs of control heuristic items.

For 23 pairs out of a total of 25 problems, the expected stereotypical association was consistently rated higher than the contrasting association for all problems except for neutral problems for which both associations were at similar rate (for a detailed overview, see Supplementary Material Section C). We discarded 2 problems which did not respect the latter pattern from all our reported analyses.

Procedure

The experiment was run through the Qualtrics (www.qualtrics.com) offline app on tactile tablets and was part of a larger study that focused on evaluating cognitive processing in the Namibian population. The current training study was always preceded by two unrelated experiments.

In the villages, the experiment took place outdoors, while in the town participants were tested indoors. Participants were always sitting next to a research assistant. All participants were tested in Otjiherero (i.e., “Himba language”) and were placed in front of a tablet.

Participants were instructed that the experiment would demand their full concentration. A general description of the task was presented in which participants were instructed that the translator would read reasoning problems, for which they would have to provide a response.

In order to familiarise the participants with the procedure and to ensure that they understood the base-rate task, the translator explained that the task would initially consist of introducing the inhabitants of a village. This was accompanied by the presentation of paper cards showing the portraits of these villagers. Then the translator explained to the participants that we would draw one person from this village at random. To illustrate this point, the translator collected all the paper cards and put them in an opaque bag, which he shuffled in full view of the participant. He then drew a random card out of the bag. The translator read aloud the description associated with the person, without showing the card to the participant, and asked the participant to select which one of the two groups of villagers the selected person was more likely to belong to. To give an answer, the participant was presented with two pictures representing a characteristic of each group and asked to touch the picture corresponding to their response. The translator did not provide any feedback. Then the previous steps were repeated on the tablet. The participants then saw two practice problems. After answering these problems, the translator asked the participants if they understood the task and whether they needed to practice again. None of the participants answered that they had not understood the task and needed a second practice.

As stated, the pre- and post-intervention blocks were run on a tablet. For the intervention block, the translator again used the paper cards and the bag to illustrate the explanations. Unlike the practice trials, the translator systematically showed the participant the selected card from the bag to accompany his words. The translator selected the response given by the participant during the intervention on the tablet, whereas in the pre- and post-intervention blocks, the participant was encouraged to do it themselves.

At the end of the study, participants in the control group were also presented with the explanations about how to solve the base-rate problems. All participants were asked to rate the material content and demographic information was collected, including gender, age, number of years at school, and literacy.

Results

Training group VS Control group

We first focused on the overall (i.e., irrespective of schooling and living environment) potential training effect among Himba participants by contrasting performance in the training and control group before and after the intervention. For each participant, we calculated the average proportion of correct responses for conflict, no-conflict, and neutral problems, in the pre- and post-intervention blocks. We analysed the data using a Linear Mixed-Effects Model (LMM) on average conflict, no-conflict and neutral accuracies with block (pre- VS post-intervention) and group (training VS control) as fixed factors and participants as random effect intercept. We tested the fixed-effects using the Kenward-Roger method that has been shown sufficiently conservative and robust for small sample sizes (Luke, 2017). Furthermore, we selectively applied Bayes Factor analysis to evaluate the evidence for null effects. This approach was specifically used when conventional frequentist statistical tests failed to reveal statistically significant results.

As Figure 3 shows, participants performed very poorly on the critical conflict problems before the intervention (respectively, M = 8.3%, SEM = 4.3, and M = 3.0%, SEM = 1.2, in the control and training groups). However, after the intervention there was a sharp increase in the training group (n = 74, M = 38.0%, SEM = 3.7) whereas this was not the case in the control group (n = 10, M = 0.0%, SEM = 0.0). The block x group interaction was significant, $F(1,82) = 15.84$, $p < .001$. These findings indicate that the training can boost conflict problem performance among the Himba—just as it has been previously observed in western samples (Boissin et al., 2022)[2].

On the no-conflict problems, performance was overall very high. The high no-conflict performance established that participants could perform our task and refrained from simple random responding. Although there tended to be an overall slight accuracy decrease in the post-intervention block (M pre-intervention = 96.2%, SEM = 1.2, M post-intervention = 80.8%, SEM = 2.5), $F(1,82) = 13.61$, $p < .001$, this was observed both in the control (n = 10) and training (n = 74) groups. Hence, there was no indication of a block x group interaction, $F(1,82) = 0.02$, $p = .964$. A Bayes factor analysis designed to contrast the effect of the block x group interaction on no-conflict items accuracy, reveals that the model without interaction is moderately more likely than the model with interaction ($BF_{01} = 2.90$, 95% CI ± 1.96%). This argues against an alternative "reversed heuristic" account (Boissin et al., 2022) in which training would lead participants to generally distrust their (stereotypical) intuition and blindly select the response that goes against it. Such a mechanism should have resulted in a specific decreased and floored post-intervention performance in the training group only (see also methods).

We also included neutral problems in our test set that did not cue a stereotypical response (e.g., "this person has two hands") and were used to explore whether the training effect on conflict problems transferred to a general increased selection of the base-rate response in the absence of conflicting information. Boissin et al. (2022) did not observe such an effect in their WEIRD studies. Although there was a visual trend toward an overall post-intervention performance increase in the

[2] For comparison, Boissin et al. (2022) observed a 33 points base-rate training increase, compared to the present 35 points increase.

current study, (M pre-intervention = 45.2%, SEM = 4.1, M post-intervention = 61.5%, SEM = 4.2), the main effect of block failed to reach significance, $F(1,82) = 1.38$, $p = .240$, as did the block x group interaction, $F(1,82) = 0.05$, $p = .816$. A Bayes factor analysis designed to contrast the effect of the block x group interaction on neutral items accuracy, reveals that the model without interaction is moderately more likely than the model with interaction ($BF_{01} = 2.30$, 95% CI ± 23.91%). Hence, as in Boissin et al. (2022), we did not observe a transfer training effect.

Finally, to ensure that the training did not lead reasoners to use a new heuristic such as "always pick the largest group", our post-intervention block included control heuristic items in which the description was completely diagnostic (e.g., "the person is pregnant", see methods). Here the correct response is to go against the base-rates. Performance on these control items was very high both for the trained reasoners (n = 74, M = 89.9%, SEM = 2.5) and for reasoners of the control group (n = 10, M = 100.0%, SEM = 0.0). Although there was a slightly lower performance in the training group, $t(83) = 3.99$, $p < .001$, $d = 0.66$, the training group performance remained close to ceiling and clearly differed from chance, $t(73) = 15.68$, $p < .001$, $d = 1.82$, or zero, $t(73) = 35.36$, $p < .001$, $d = 4.11$. Thus, trained reasoners seem to grasp the problem logic and do not start blindly selecting the base-rate information at the expense of the descriptive information. Overall, these training results among a Himba population are highly consistent with the training results observed in Western samples.

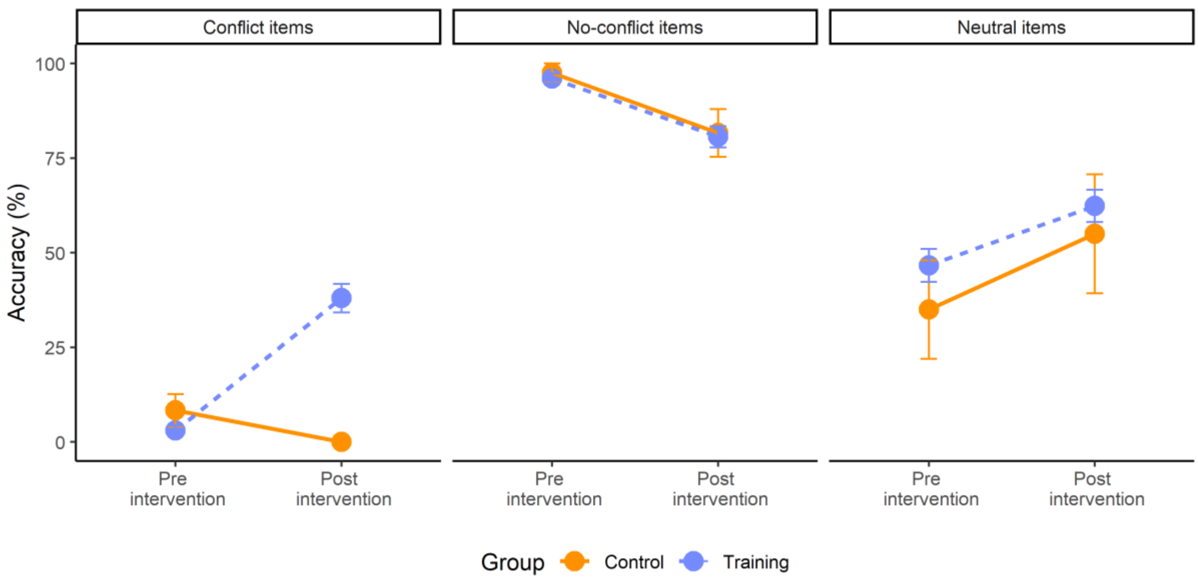

Figure 3. Average accuracy on conflict, no-conflict and neutral problems across block (pre- VS post-intervention) and group (training VS control). Error bars represent standard errors of the mean (SEM).

<u>Formal education and living environment effect on training performance</u>

Next, we explored the impact of formal education and living environment on the training effect. We analysed the data using a LMM on average conflict items of the trained participants with block (pre- VS post-intervention), formal education (school VS no school) or living environment (town VS village) as fixed factors and participants as random effect intercept. We tested the fixed effects using the Kenward-Roger method. Bayes Factor analysis was used to evaluate the evidence for null effects.

Regarding formal education, most reasoners failed to solve conflict problems before the intervention whether they went to school (n = 36, M = 2.08%, SEM = 1.17) or not (n = 38, M = 3.95%, SEM = 2.10). Both groups improved after the intervention, main effect of block: $F(1,72) = 80.83$, $p < .001$, but the post-intervention increase tended to be slightly more pronounced for the school group (M = 43.9%, SEM = 31.09) compared to the no-school group (M = 32.46%, SEM = 5.30). The block x formal education interaction reached marginal significance, $F(1,72) = 2.88$, $p = .094$. A Bayes factor analysis aimed at contrasting the effect of the block x group interaction reveals that the model with interaction is only slightly more likely than the model without it ($BF_{10} = 0.85$, 95% CI ± 3.36%). Therefore, this suggests that the evidence for the interaction is indeed weak.

Regarding living environment, both towners (n = 31) and villagers (n = 43) showed very poor accuracy before the training (respectively, M = 2.96%, SEM = 2.27, and M = 3.10%, SEM = 1.33). Both groups improved after the training, main effect of block: $F(1,72) = 95.02$, $p < .001$. However, people in town (M = 51.45%, SEM = 5.60) benefitted more from the training compared to villagers (M = 28.30%, SEM = 4.52), the block x living environment interaction was significant, $F(1,72) = 9.50$, $p = .003$.

In sum, our data show that both schooling and living in an industrialized environment boost the training effect. However, as Figure 4 shows, even among people who lived in rural Himba villages and had not received any formal schooling (n = 22) we still see a clear training benefit, with a performance increasing from 3.8% (SEM = 3.1) to 23.1% (SEM = 6.5). Bonferroni post-hoc analysis revealed a significant effect $z(21) = 2.82$, $p = .005$.

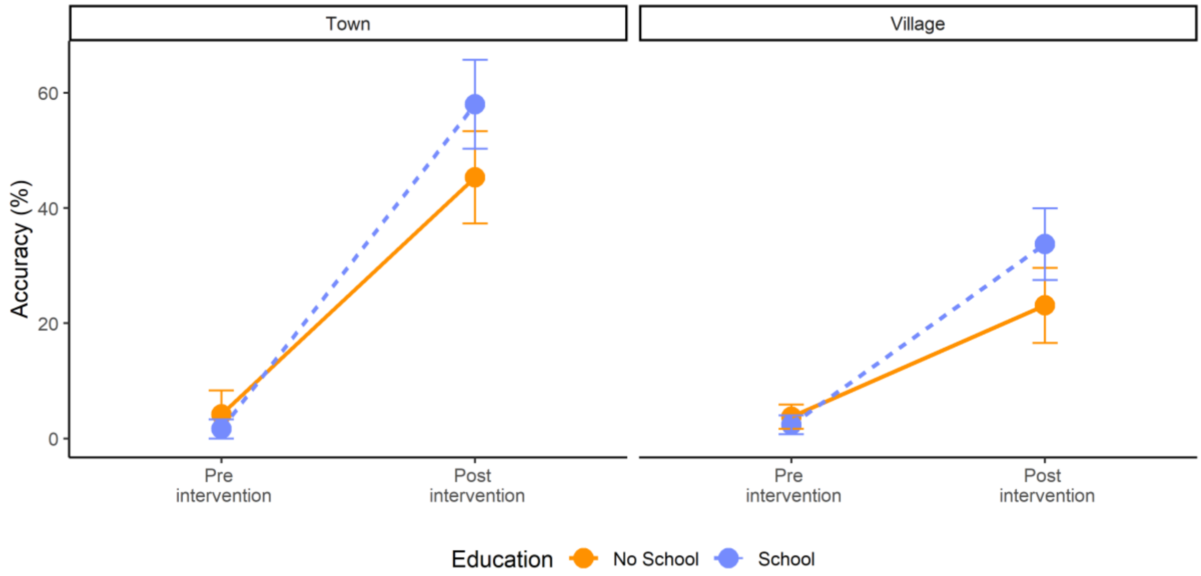

Figure 4. Average accuracy on conflict problems in the training group, as a function of living environment (town VS village) and formal education (school VS no school), before and after the intervention. Error bars represent standard errors of the mean (SEM).

<u>Individual level analysis across formal education and living environment</u>

To explore further how participants benefited from the training (or not), we classified reasoners according to an individual level conflict accuracy analysis for each trained participant. We distinguished three different categories: Participants who consistently gave incorrect responses throughout the study ("Biased" participants; i.e., participants with 0% accuracy both before and after the intervention), participants who improved after the intervention ("Improved" participants; i.e., who showed a higher proportion of correct answers after than before the intervention), and participants who showed an inconsistent response pattern ("Other"; i.e., participants with neither the 'biased' nor the 'improved' pattern, for instance showing a decrease in accuracy after the intervention). Table 1 shows the distribution. It is clear that, although in all conditions there are some participants who remain biased after the training intervention, the majority of reasoners does benefit from the intervention. Corroborating our overall analysis, there are more improved reasoners among those who live in town or went to school but critically, even among Himba villagers who never went to school, up to 45.5% of the sample does benefit from the training.

Table 1. Proportion of type of reasoners (biased, improved VS others) based on an individual level accuracy classification as a function of formal education (school VS no-school) and living environment (town VS village) in the training group.

		Biased	Improved	Other
Town	School	13.3%	86.7%	-
(n = 31)	(n = 15)	(n = 2)	(n = 13)	
	No-school	12.5%	87.5%	-
	(n = 16)	(n = 2)	(n = 14)	
Village	School	23.8%	71.4%	4.8%
(n = 43)	(n = 21)	(n = 5)	(n = 15)	(n = 1)
	No-school	40.9%	45.5%	13.6%
	(n = 22)	(n = 9)	(n = 10)	(n = 3)

Finally, we also wanted to check whether there were any further post-intervention performance differences between the improved reasoners in our different conditions. We therefore calculated the average accuracy of the improved reasoners group on our different item types (i.e., conflict, no-conflict, neutral and control heuristic items) as a function of living environment (i.e., town and village) and formal education (i.e., school and no-school). As Table 2 indicates, the average post-intervention accuracy of improved reasoners on conflict problems is fairly comparable, as is their performance on control, no-conflict, neutral, and control heuristic problems. A set of LMMs with formal education (school VS no school) and living environment (town VS village) as between-subjects factors on improved reasoners' average accuracies and participants as random intercept on conflict, $F(1,48.01) = 1.38$, $p = .25$, no-conflict, $F(1,48.00) = 0.13$, $p = .72$, neutral, $F(1,48.01) = 0.64$, $p = .43$, and control heuristic problem accuracy, $F(1,48.01) = 1.15$, $p = .29$, did not detect significant differences (note that we tested fixed effects using the Kenward-Roger method). Bayes factor analyses, based on the comparison of the models with the education x living environment interaction to the models without the interaction, led to the same conclusions (conflict items: $BF_{01} = 1.72$, 95% CI± 2.74%, no-conflict items: $BF_{01} = 2.58$, 95% CI± 3.64%, neutral items: $BF_{01} = 2.28$, 95% CI± 4.62%, control items: $BF_{01} = 1.87$, 95% CI± 1.66%). In sum, although schooling and urban living boost the number of individuals who benefit from training, those who do benefit, benefit about equally, even when they live in a rural village and had no schooling whatsoever.

Table 2. Conflict, No-conflict, Neutral and Control heuristic problem mean post-intervention accuracy (standard errors of the mean, SEM) of the improved reasoners from the training group as a function of formal education (school VS no-school) and living environment (town VS village).

		Conflict	No-conflict	Neutral	Control heuristic
Town	School	64.7% (7.0)	76.9% (7.5)	76.9% (10.8)	84.6% (8.7)
	No-school	51.8% (7.7)	78.0% (6.2)	71.4% (6.9)	92.9% (4.9)
Village	School	47.2% (5.6)	82.2% (5.7)	70.0% (6.5)	86.7% (5.9)
	No-school	50.8% (7.8)	78.3% (6.9)	50.0% (12.9)	80.0% (8.2)

Discussion

In the present study we used a debiasing paradigm with Himba participants in rural Namibia to explore the nature of biased reasoning among non-WEIRD samples. Our results indicate that a short debiasing intervention boosted reasoning performance of Himba participants on problems in which personal beliefs conflict with logical (i.e., base-rate) considerations. Although pre-training performance was floored and strongly drew up on personal beliefs (consistent with previous studies; Luria, 1971; Scribner, 1977; Trémolière et al., 2022), it increased dramatically after training and became overall comparable to that reported in WEIRD samples (e.g., Boissin et al., 2022).

We also observed that living in an industrialized environment (as opposed to a rural village) and having received formal education boosted the training effect. More individuals benefitted from the training among Himba who had received formal education or lived in town. A substantial number of rural Himba villagers who received no formal education (about 50%) nevertheless showed a performance increase after training. Moreover, the latter villagers who benefitted from the training benefitted to the same extent as towners or schooled individuals with no indication that the accuracy boost on conflict problems was driven by an increased tendency to guess or a suboptimal heuristic strategy (i.e., general intuition mistrust or blind discarding of the description).

The present findings help answer the performance-competence question regarding non-WEIRD participants' alleged failure to decontextualize problems, and their tendency to let personal beliefs prevail over logical principles (Luria, 1971, 1976). Given that our debiasing intervention was extremely short (10 minutes) and highlighted the logico-mathematical base-rate principle with just one example, the finding that this can nevertheless suffice to remediate even unschooled, rural Himba villagers' reasoning, suggests that the bias does not result from a lack of knowledge (Stanovich, 2011). That is, the fact that the short explanation allowed a large number of the participants to respond correctly suggests that the critical knowledge concerning the role of base-rates on probability judgments was already implicitly available. It was simply less instantiated (Stanovich, 2018; Raoelison et al., 2021) and its relevancy needed to be clarified. Hence, these debiasing results lend credence to the idea that, as in WEIRD samples (Boissin et al., 2021, 2022), the previously reported pronounced biased thinking of non-WEIRD participants (e.g., Luria, 1971; Scribner, 1977; Trémolière et al., 2022) primarily results from a performance rather than an ability issue. At the same time, given that not all unschooled villagers benefited from the training, we cannot exclude that some illiterate participants (just like schooled WEIRD participants) may suffer from a competence failure. At the very least, it is safe to conclude that the poor reasoning performance previously reported among non-WEIRD participants does not necessarily result from a competence failure.

Our debiasing results may have important practical and methodological implications. Although we adopted a task format, task material, and task familiarization procedure that was especially designed and optimized for testing among a specific non-WEIRD sample (the Himba), our participants' untrained performance was extremely low (i.e., accuracy between 3% and 8%). To put this in perspective, with similar base-rate problems, WEIRD adults score around 20% to 60% (Bago & De Neys, 2017; Boissin et al., 2022; De Neys & Glumicic, 2008; Pennycook et al., 2015) and even WEIRD pre-schoolers tend to show better performance than our untrained adult Himba participants (from 20% to 40%, e.g., De Neys & Vanderputte, 2011). In a standard no-training paradigm, we may have concluded from the low reasoning performance that non-WEIRD reasoners are dramatically poor at decontextualizing and indeed do not manage to apply logico-mathematical principles. This suggests that the assumption that non-WEIRDs lack the necessary logical knowledge for sound reasoning always has to be evaluated with caution. We advise to integrate the current brief training approach in cross-cultural studies aimed at measuring the reasoning capacities of non-WEIRD samples.

We believe that the present study may serve as a proof-of-principle for the potential of debias training in non-WEIRD studies. However, it is important to note that non-WEIRD studies remain rare in reasoning bias research and that the present study also faced some limitations. For example, the study was not preregistered, the sample size remained limited and unbalanced across conditions. Although some of these limitations underscore the intrinsic challenges in testing non-WEIRD groups (e.g., in-person testing in remote villages) this implies that the results remain to be interpreted with caution and that further testing and validation is required.

With respect to follow-up work it is also important to stress that we only explored one specific type of intervention. Within the field of heuristics and biases, various other approaches to debiasing WEIRD reasoners have been investigated (e.g., see Adame, 2016; Mata, 2020; Mata et al., 2013; Reis et al., 2023). For instance, Isler and Yilmaz (2022) and Isler et al. (2020) demonstrated that various debiasing techniques such as asking participants to provide justifications for their responses alongside explanations, as well as monetary incentives, were also effective at boosting reasoners' performance. Future studies should investigate the impact of these interventions and compare their effects within non-WEIRD populations. That is, it may be possible that other approaches result in similar or even stronger debias effects among non-WEIRD reasoners. The present proof-of-principle can clearly serve as a precursor for such extended non-WEIRD testing in debias studies.

In sum, our results reinforce the idea that biased human reasoning often results from a performance rather than a competence issue (De Neys & Bonnefon, 2013; Stanovich, 2011). They show that a short training intervention is sufficient to help remote non-WEIRD participants to improve their reasoning and get them to favour numerical probabilities over stereotypical beliefs. Overall, this study sheds new light on the universality of both the flaws and strengths of human thinking and suggests that simple interventions can be employed to boost sound reasoning all around the world.

Declaration of competing interest

The authors report no conflict of interest.

CRediT author statement

Conceptualization: EB, SC, WDN. **Methodology:** EB, SC, WDN. **Software:** EB. **Formal analysis:** EB. **Investigation:** EB, MJ, SC. **Resources:** SC, WDN. **Data curation:** EB. **Writing – Original Draft:** EB. **Writing – Review & Editing:** EB, MJ, WDN, SC. **Visualization:** EB. **Supervision:** WDN, SC. **Funding acquisition:** SC, WDN.

Acknowledgments

This work was funded by an IUF (2019) grant.

References

Adame, B. J. (2016). Training in the mitigation of anchoring bias: A test of the consider theopposite strategy. *Learning and Motivation*, 53, 36-48. https://doi.org/10.1016/j.lmot.2015.11.002

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63, 602–614. https://doi.org/10.1037/14805-008

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90-109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257-299.

https://doi.org/10.1080/13546783.2018.1507949

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing system 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making, 17,* 646-690.

Cameron, A. S. (2013). *The influence of media on Himba conceptions of dress, ancestral and cattle worship, and the implications for culture change*. [Master's thesis, Brigham Young University]. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=4426&context=etd

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology*, *3*(1), 21-29. https://doi.org/10.1111/j.2044-835X.1985.tb00951.x

Carraher, T. N., Carraher, D. W., Schliemann, A. D., & Schliemann, A. D. (1987). Written and Oral Mathematics. *Journal for Research in Mathematics Education*, *18*(2), 83. https://doi.org/10.2307/749244

Castelain, T., Girotto, V., Jamet, F., & Mercier, H. (2016). Evidence for benefits of argumentation in a Mayan indigenous population. *Evolution and Human Behavior*, 37(5), 337-342 https://doi.org/10.1016/j.evolhumbehav.2016.02.002

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052-1066. https://doi.org/10.1037/xge0000323

Cole, M., & Bruner, J. S. (1971). Cultural differences and inferences about psychological processes. *American Psychologist*, *26*(10), 867.

De Neys, W., & Bonnefon, J. F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172-178. https://doi.org/10.1016/j.tics.2013.02.001

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248-1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Vanderputte, K. (2011). When less is not always more: Stereotype knowledge and reasoning development. *Developmental Psychology*, *47*(2), 432-441. https://doi.org/10.1037/a0021313

Dias, M., Roazzi, A., & Harris, P. (2005). Reasoning from unfamiliar premises a study with unschooled adults. *Psychological Science*, *16*, 550–554. https://doi.org/10.1111/j.0956-7976.2005.01573.x

Evans, J. S. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103(2)*, 356–363. https://doi.org/ 10.1037/0033-295X.103.2.356

Fontanari, L., Gonzalez, M., Vallortigara, G., & Girotto, V. (2014). Probabilistic cognition in two indigenous Mayan groups. Proceedings of the National Academy of Sciences, 111(48), 17075-17080. https://doi.org/10.1073/pnas.1410583111

Frey, D., & De Neys, W. (2017). Is conflict detection in reasoning domain general? Proceedings of the annual meeting of the cognitive science society (pp. 391–396).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–135. https://doi.org/10.1017/S0140525X0999152X

Hirt, E.R., & Markman, K.D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. Journal of Personality and Social Psychology, 69(6), 1069.

Hogarth, R.M. (2008). On the learning of intuition. In Plessner, H., Betsch, C., & Betsch, T. (Eds.), Intuition in judgment and decision making (pp. 91–105). Lawrence Erlbaum Associates Publishers

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants : Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, *24*(6), 1922-1928. https://doi.org/10.3758/s13423-017-1241-8

Isler, O., & Yilmaz, O. (2022). How to activate intuitive and reflective thinking in behavior research? A comprehensive examination of experimental techniques. *Behavior Research Methods,* 1-20. https://doi.org/10.3758/s13428-022-01984-4

Isler, O., Yilmaz, O., & Dogruyol, B. (2020). Activating reflective thinking with decision justification and debiasing training. *Judgment and Decision making*, *15*(6), 926-938.

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Strauss, Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80(4),* 237–251. https://doi.org/10.1037/h0034747

Luke, S.G. (2017). Evaluating significance in linear mixed-effects models in R. Behavior Research Methods, 49(4), 1494–1502. doi:10.3758/s13428-016-0809-y.

Luria, A. R. (1971). Towards the problem of the historical nature of psychological processes 1. *International Journal of Psychology*, *6*(4), 259-272. https://doi.org/10.1080/00207597108246692

Luria, A. R. (1976). The neuropsychology of memory. Washington, DC: V. H. Winston.

Maddux, W. W., Yang, H., Falk, C., Adam, H., Adair, W., Endo, Y., Carmon, Z., & Heine, S. J. (2010). For whom is parting with possessions more painful?: Cultural differences in the endowment effect. *Psychological Science*, 21(12), 1910–1917. https://doi.org/10.1177/0956797610388818

Mata, A. (2020). An easy fix for reasoning errors : Attention capturers improve reasoning performance. *Quarterly Journal of Experimental Psychology*, *73*(10), 1695-1702. https://doi.org/10.1177/1747021820931499

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, 49(3), 486-491. https://doi.org/10.1016/j.jesp.2013.01.010

Mercier, H. (2011). On the universality of argumentative reasoning. *Journal of Cognition and Culture*, *11*(1–2), 85–113. https://doi.org/10.1163/156853711X568707

Ministry of Education, Arts and Culture of Republic of Namibia. (2017). *A public expenditure review of the basic education sector in Namibia.* https://www.unicef.org/esa/sites/unicef.org.esa/files/2019-04/PER-of-Education-in-Nambia-%282017%29.pdf

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. Trends in Cognitive Sciences, 14(10), 435–440. https://doi.org/10.1016/j.tics.2010.07. 004

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129-140. https://doi.org/10.1177/2372732215600886

Mwinga, M., Kavezuva, C., Shifdi, A., & Simasiku, P. (2022). *Opuwo economic profil*. First Capital Namibia. https://firstcapitalnam.com/wp-content/uploads/2022/10/Opuwo-Economy-Profile-Report.pdf

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310. https://doi.org/10.1037/0033-295X.108.2.291

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge University Press.

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2020). Domain-specific experience and dual-process thinking. *Thinking & Reasoning, 27* (2), 239-267. https://doi.org/10.1080/13546783.2020.1793813

Raoelison, M., Boissin, E., Borst, G., & Neys, W. D. (2021). From slow to fast logic: The development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599-622. https://doi.org/10.1080/13546783.2021.1885488

Reis, J., Ferreira, M. B., Mata, A., Seruti, A., & Garcia-Marques, L. (2023). Anchoring in a social context: How the possibility of being misinformed by others impacts one's judgment. *Social Cognition*, *41*(1), 67-87. https://doi.org/10.1521/soco.2023.41.1.67

Saxe, G. B. (1988). Candy selling and math learning. *Educational researcher*, *17*(6), 14-21.

Scribner, S. (1977). Modes of thinking and ways of speaking: Culture and logic reconsidered. *Thinking: Readings in cognitive science*, 483-500.

Scribner, S., & Cole, M. (1981). Unpackaging literacy. *Writing: The nature, development, and teaching of written communication*, *1*, 71-87.

Sharp, D. W., & Cole, M. (1975). The influence of educational experience on the development of cognitive skills as measured in formal tests and experiments. Rockfeller University, New York, NY.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444. https://doi.org/10.1080/13546783.2018.1459314

Trémolière, B., Davidoff, J., & Caparos, S. (2022). A 21st century cognitive portrait of the Himba, a remote people of Namibia. *British Journal of Psychology*, *113*(2), 508-530. https://doi.org/10.1111/bjop.12539

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958-1971. https://doi.org/10.1037/a0037099

Yilmaz, O., & Saribay, S. A. (2017a). Analytic thought training promotes liberalism on contextualized (but not stable) political opinions. *Social Psychological and Personality Science*, *8*(7), 789–795. https://doi.org/10.1177/1948550616687092

Yilmaz, O., & Saribay, S. A. (2017b). Activating analytic thinking enhances the value given to individualizing moral foundations. *Cognition*, *165*, 88-96. http://dx.doi.org/10.1016/j.cognition.2017.05.009

Supplementary Material

Age-estimation method

We assessed age using an approximate method given that Himba villagers do not record their age and are not familiar with the numerical age concept. Our local guide asked whether the participant knew their age, whether they had children and grand-children (if yes, how many), and evaluated physical appearance. Given on all this information, the guide estimated the participant's age.

Problems used in the Study (in English version)

N°	No-conflict version	Conflict version
1	In this new village, there are 8 women and 2 men. The randomly chosen person looks after children.	In this new village, there are 2 women and 8 men. The randomly chosen person looks after children.
2	In this new village, there are 2 men and 8 women. The randomly chosen person prepares the meals.	In this new village, there are 8 men and 2 women. The randomly chosen person prepares the meals.
3	In this new village, there are 8 young people and 2 old people. The randomly chosen person went to school.	In this new village, there are 2 young people and 8 old people. The randomly chosen person went to school.
4	In this new village, there are 2 old people and 8 young people. The randomly chosen person fetches water.	In this new village, there are 8 old people and 2 young people. The randomly chosen person fetches water.
5	In this new village, there are 8 men and 2 women. The randomly chosen person slaughters goats.	In this new village, there are 2 men and 8 women. The randomly chosen person slaughters goats.
6	In this new village, there are 2 women and 8 men. The randomly chosen person went outside the village.	In this new village, there are 8 women and 2 men. The randomly chosen person went outside the village.
7	In this new village, there are 2 young people and 8 old people. The randomly chosen person knows the history of the ancestors.	In this new village, there are 8 young people and 2 old people. The randomly chosen person knows the history of the ancestors.
8	In this new village, there are 8 young people and 2 old people. The randomly chosen person searches for lost animals.	In this new village, there are 2 young people and 8 old people. The randomly chosen person searches for lost animals.
9 (Intervention item)	In this new village, there are 2 men and 8 women. The randomly chosen person digs wells.	In this new village, there are 8 men and 2 women. The randomly chosen person digs wells.

10	In this new village, there are 8 men and 2 women. The randomly chosen person smokes cigarets.	In this new village, there are 2 men and 8 women. The randomly chosen person smokes cigarets.
11	In this new village, there are 2 women and 8 men. The randomly chosen person takes care of the cows.	In this new village, there are 8 women and 2 men. The randomly chosen person takes care of the cows.
12	In this new village, there are 2 old people and 8 young people. The randomly chosen person has the most cows.	In this new village, there are 8 old people and 2 young people. The randomly chosen person has the most cows.
13	In this new village, there are 8 young people and 2 old people. The randomly chosen person is responsible for trading.	In this new village, there are 2 young people and 8 old people. The randomly chosen person is responsible for trading.
14	In this new village, there are 2 young people and 8 old people. The randomly chosen person looks after the cattle.	In this new village, there are 8 young people and 2 old people. The randomly chosen person looks after the cattle.
15	In this new village, there are 8 old people and 2 young people. The randomly chosen person is a messenger.	In this new village, there are 2 old people and 8 young people. The randomly chosen person is a messenger.
16	In this new village, there are 8 women and 2 men. The randomly chosen person milks goats.	In this new village, there are 2 women and 8 men. The randomly chosen person milks goats.
17	In this new village, there are 2 men and 8 women. The randomly chosen person collects firewood.	In this new village, there are 8 men and 2 women. The randomly chosen person collects firewood.

N°	Neutral version	N°	Control version
18	In this new village, there are 8 women and 2 men. The randomly chosen person has two hands.	24	In this new village, there are 8 men and 2 women. The randomly chosen person is pregnant.

19	In this new village, there are 8 young people and 2 old people. The randomly chosen person has two legs.	25	In this new village, there are 8 young people and 2 old people. The randomly chosen person has kids.
20	In this new village, there are 8 men and 2 women. The randomly chosen person has two eyes.		
21	In this new village, there are 8 old people and 2 young people. The randomly chosen person has two feet.		

N°	**Practice trials – No conflict version**	**Practice trials – Conflict version**
22	In this new village, there are 8 women and 2 men. The randomly chosen person wears jewel.	In this new village, there are 2 women and 8 men. The randomly chosen person wears jewel.
23	In this new village, there are 8 men and 2 women. The randomly chosen person drives cars.	In this new village, there are 2 men and 8 women. The randomly chosen person drives cars.

Rating task of base-rate trials

Figure S1. Example of one rating item question. Reasoners have to decide for each statement, how much they agree with. The first sentence is congruent since the selected group (i.e., "Women") fits with the general description (i.e., "Prepare the meals"). The second sentence is incongruent since usually Men do not prepare meals. The order presentation of congruent and incongruent sentence was counterbalanced for each subject.

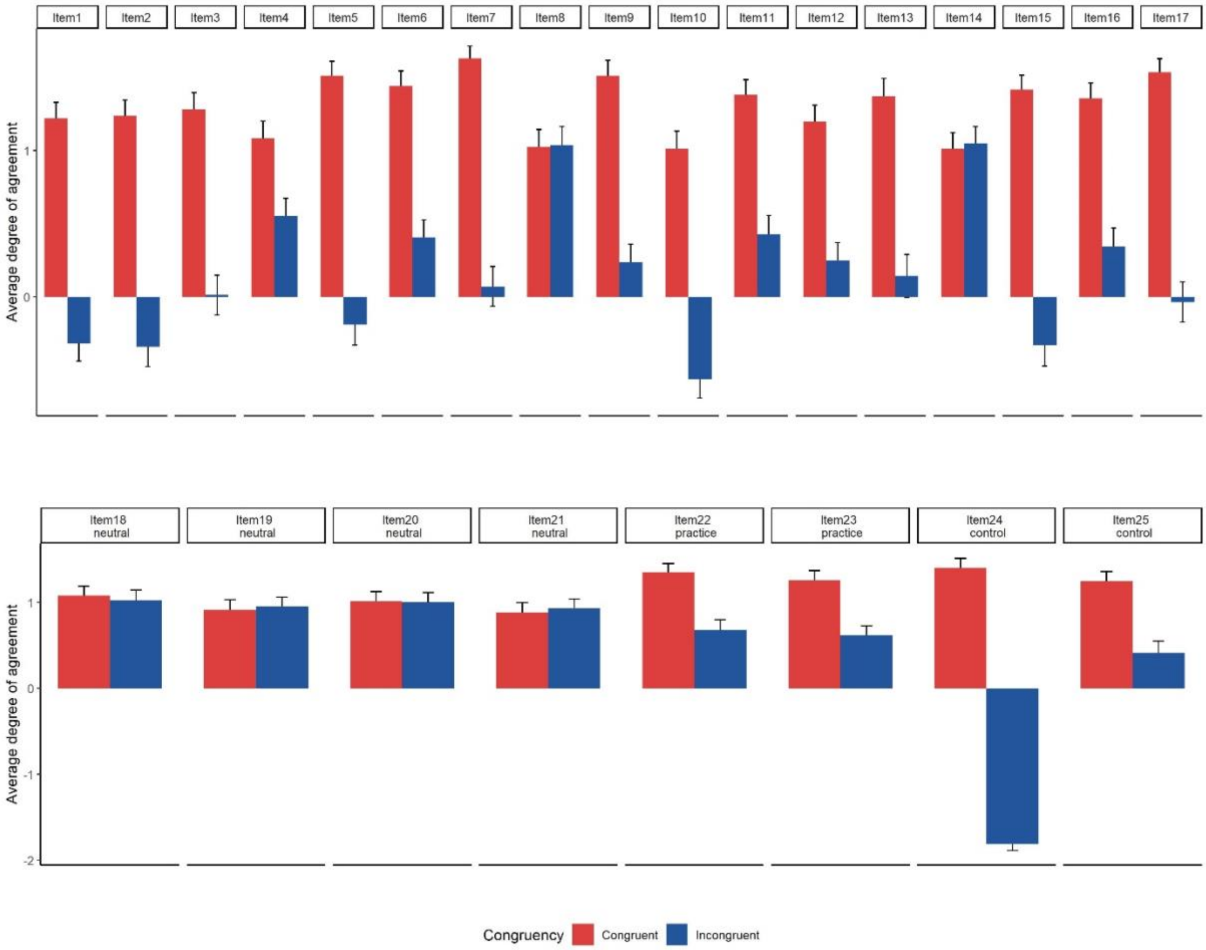

Figure S2. Average degree of agreement of each base-rate trials as a function of congruency (congruent vs incongruent). Bar errors are errors standard of the mean (SEM). Validated items are those for which the score of congruent statements is higher than incongruent statements, except in the case of neutral items for which each statement must have similar rates. Thus, Items 8 and 14 were removed from our analyses.