Comparing Individual and Collective Performance in Deductive Reasoning across Content Types

Alba Massolo (alba.massolo@unc.edu.ar)
Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba
Pabellón Agustín Tosco, 1° Piso, Ciudad Universitaria, C.P. 5000, Córdoba, Argentina

Facultad de Ciencias de la Salud, Universidad Católica de Córdoba
Obispo Trejo 323, C.P. 5000, Córdoba, Argentina

Mariel Traversi (1936359@ucc.edu.ar)
Facultad de Ciencias de la Salud, Universidad Católica de Córdoba
Obispo Trejo 323, C.P. 5000, Córdoba, Argentina

Abstract

This research analyzes whether collective performance surpasses individual performance in solving two different deductive reasoning tasks. One of the tasks contains arguments with factual content, while the other includes arguments with ideological content. Additionally, we seek to determine whether the truth-wins model accurately represents the social combination process involved in the collective resolution of both deductive tasks. We designed and conducted two studies. Study 1 (N = 115) employed a within-subjects design with three conditions: individual, collective, and post-collective individual resolutions. Participants evaluated syllogisms with factual content. Study 2 (N = 111) followed a similar design but used syllogisms with ideologically controversial content. Results from both studies indicate that collective performance surpassed individual performance. Furthermore, while the truth-wins model best describes the collaborative decision-making process in Study 1, the majority model was more accurate for Study 2. These findings align with theories advocating for the social origins of human reasoning.

Keywords: deductive reasoning; collaborative reasoning; content effects; social combination models; cognitive biases

Introduction

The psychology of reasoning has amply demonstrated over the years that human deductive performance is poor. Numerous empirical studies have shown human errors in solving diverse deductive reasoning problems (Evans, 2002; Stanovich, 2021; Wason, 1966). However, when these deductive tasks are resolved in collaborative group settings, the results show a marked improvement (Moshman, 2021; Trouche, Sander & Mercier, 2014). These findings suggest that deductive performance in collaborative settings is quite good, unlike individual reasoning.

These empirical findings align with philosophical theories that advocate for a dialogical origin of deduction. It has been argued that humans are better deductive reasoners when they exchange reasons and arguments in a dialogical setting (Dutilh-Novaes, 2021). Furthermore, evolutionary theories about the social origins of human reasoning offer a phylogenetic explanation for the emergence of these deductive skills. These theories propose that humans developed cognitive abilities for collective agency as environmental pressures required cooperation among individuals for successful foraging (Tomasello, 2022). In a similar line of research, it has been suggested that humans, building upon a genetic starter kit, developed in their cultural practices a series of cognitive gadgets that were culturally transmitted (Heyes, 2018). These tools likely facilitated the emergence of argumentative practices, which, in turn, laid the groundwork for the development of deductive reasoning (Dutilh-Novaes, 2021).

The Wason selection task, introduced by Peter Wason in 1966, is a well-known test for analyzing deductive reasoning. Research indicates that only about 9% of individuals manage to solve this task correctly (Johnson-Laird & Wason, 1970). However, when this task is performed in a group setting, the percentage of correct responses significantly increases to 75% (Moshman & Geil, 1998). Subsequent research on group problem-solving consistently confirms this enhanced performance, both for the Wason selection task (Boku, Yama, & Mercier, 2018; Mercier et al., 2016) and other deductive tasks like the Levesque problem (1986) (Castelain et al., 2016; Trouche et al., 2014). These findings suggest that the dialogical exchange of arguments during deductive reasoning significantly improves human cognitive performance.

In a series of studies on group problem-solving, Patrick Laughlin (2011) proposed a continuum of problem types with two distinct extremes: intellective and judgmental problems. Intellective problems refer to tasks with demonstrably correct solutions within mathematical, logical, scientific, or conceptual frameworks. In contrast, judgmental problems are evaluative tasks or aesthetic judgments without a demonstrably correct solution. In this continuum, deductive reasoning problems constitute a paradigmatic example of an intellective problem since they possess a unique correct solution within a formal logical system.

To explain how individual responses are combined into a collective decision, Laughlin proposed five models of social combination, with research showing that the number of individual responses needed to produce a collective response is inversely proportional to the demonstrability of the problem (Laughlin & Ellis, 1986). Thus, the most suitable social combination model for a typical judgmental problem

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

is the majority model: Two-thirds of coinciding individual decisions are necessary and sufficient to provide a collective response. Whereas for an intellective problem, the most accurate social combination model is the truth-wins model, where only one correct individual response is necessary and sufficient for the collective response to be correct (Laughlin & Adamopoulos, 1980).

In the proposed models of social combination, the majority model and the truth-wins model represent opposite extremes regarding the number of individual responses needed to reach a collective decision. In decreasing order of individual responses required, the majority model is followed by the proportionality model, where a simple majority of responses is necessary and sufficient for the group to arrive at a collective response. Next is the equiprobability model, in which the collective response reflects the same probability as any response given by at least one group member. Before reaching the truth-wins model, there is the truth-supported-wins model, which requires two correct individual responses for a correct collective response (Laughlin, 2011).

A key topic in the study of human reasoning is the influence of content. Research indicates that while the content of a task can sometimes facilitate reasoning, it can also introduce biases that distort cognitive processes (Dominowski, 2019). For instance, factual information about the world can influence reasoning by leading individuals to judge the validity of an argument based on the believability of its conclusion rather than its logical structure. People are more likely to accept an argument as valid if they believe in its conclusion, and to reject it if the conclusion seems false. This effect, known as belief bias, has been widely documented (Ball & Thompson, 2018; Boissin, Caparos & De Neys, 2023; de Chantal et al., 2020; Evans, 2017; Šrol, 2022; Unsworth & Medin, 2005). In addition to belief bias, reasoning is also shaped by prior attitudes, opinions, and ideological tendencies. Several studies have shown that individuals tend to evaluate arguments based on their agreement with the conclusion. When they support the conclusion, they are more likely to accept the argument, whereas disagreement generally results in its rejection. Consequently, this phenomenon is referred to as myside bias (Baron, 1995; Čavojová, Šrol & Adamus, 2018; Stanovich, West & Toplak, 2013).

It is crucial to highlight that belief bias and myside bias are distinct. The former occurs when reasoning is influenced by factual information about the world, while the latter arises when reasoning is shaped by the desire to reach a conclusion that aligns with one's pre-existing ideological attitudes (Stanovich, 2021). Robert Abelson (1986) proposed a conceptual distinction between testable and distal beliefs. Testable beliefs are related to the real world and can be confirmed or refuted empirically. In contrast, distal beliefs encompass worldviews, values, attitudes, opinions, and ideologies, which are not amenable to empirical verification or refutation. Therefore, myside bias predominantly relates to distal beliefs that are not testable, while belief bias primarily involves testable beliefs (Stanovich, 2021).

The goal of the current study is to analyze the individual and the collective resolution of two different deductive reasoning tasks. One task contains arguments with factual content, while the other involves arguments with ideological content. Additionally, we seek to determine whether the truth-wins model is the most accurate representation of the social combination process involved in the collective resolution of both deductive tasks. To achieve these aims, we designed and conducted two studies. Study 1 employed a within-subjects design with three conditions: individual resolution, collective resolution, and post-collective individual resolution. In this study, participants were tasked with evaluating syllogisms with factual content. Study 2 followed a similar design but used syllogisms with ideologically controversial content. Based on the reviewed literature, we propose two main hypotheses:

Hypothesis 1: Collaborative performance will be significantly superior to individual performance in both deductive tasks.

Hypothesis 2: The truth-wins model will best represent the social combination process underlying the collaborative resolution of both reasoning tasks.

Study 1

We asked participants to determine the validity of a series of syllogisms through three conditions: (1) individual resolution, (2) collective resolution, and (3) post-collective individual resolution. We presented eight categorical syllogisms, each containing factual information about the world. These syllogisms fell into four categories: valid believable (VB), valid unbelievable (VU), invalid believable (IB), and invalid unbelievable (IU). Each category included two syllogisms.

Method

Participants We employed G*Power software to estimate an appropriate sample size of 42 participants, with an effect size of $f = .25$, a significance level of $\alpha = .05$, and a statistical power of .80. Our study analyzed a non-randomized sample of 115 undergraduate students from Argentina, aged between 18 and 54, with a mean age of 19.70 ($SD = 3.69$). Most participants (77.4%) identified as female, while 20.9% identified as male, and 1.7% indicated belonging to another gender. These students were enrolled in undergraduate psychology programs and volunteered to participate without any payment or compensation. Data collection took place in their classrooms, where we explained the research objectives and assured participants that their data would remain confidential.

Design We used a within-subjects design with three conditions: (1) individual resolution, (2) collective resolution, and (3) post-collective individual resolution. These three conditions were treated as levels of a within-group factor. Accuracy (correct or incorrect response) and

type of syllogism (VB, VU, IB, IU) were used as dependent measures.

Materials An argument evaluation task was administered in Spanish, the sample's native language. This task included eight categorical syllogisms containing factual content about the world. Both single-model and multiple-model syllogisms were employed. Participants were asked to determine whether each argument was valid or invalid. An English version of the task is available at: https://osf.io/ga6q7/?view_only=da65b8667c6f49bc8e71771da13385d1

Before conducting the study, we calculated the mean scores for the believability of the syllogism conclusions based on survey results from a group of 47 participants who did not take part in the current study. Each conclusion was rated on a Likert scale ranging from 1 ("I am totally sure that it is false") to 7 ("I am totally sure that it is true"). The results showed a noticeable difference in scores: Believable conclusions had a mean score of 5.97 (SD = 1.13), while unbelievable conclusions had a mean score of 1.7 (SD = 0.94).

Procedure Before beginning the reasoning task, all participants provided informed consent. They received general instructions, including a brief explanation of logical validity and examples of valid and invalid arguments. The study was conducted in three sequential phases. In the first phase, participants completed the argument evaluation task individually. They had to choose between two options to express their judgment on whether the conclusion followed logically: "Yes, it does logically follow" or "No, it does not logically follow." This phase corresponds to Condition 1, representing individual resolution. In the second phase, participants were randomly assigned to groups of four or five individuals and were asked to solve the task collectively. They were relocated to different classrooms for group discussions and were asked to reach an agreement on the eight arguments. This phase corresponds to Condition 2. In the third and final phase, participants completed the argument evaluation task individually once again. This final phase corresponds to Condition 3. The same syllogism evaluation task was used in all three phases. Besides, participants had no time constraints to solve the task in any of the three phases.

Results and Discussion

The collected data was analyzed using SPSS v.25.0 software. The raw data set is available at: https://osf.io/ga6q7/?view_only=da65b8667c6f49bc8e71771da13385d1. To assess global accuracy scores, we conducted a within-subject repeated measures ANOVA to compare the three conditions. The analysis revealed a main effect, $F(1.49) = 4.60$, $p = .02$, $\eta^2 = .167$. Specifically, the mean accuracy score was significantly higher in Condition 2 ($M = .78$, $SD = .14$) compared to Condition 1 ($M = .71$, $SD = .16$). No statistically significant differences were observed for Condition 3 ($M = .76$, $SD = .15$).

Furthermore, we conducted within-subject repeated measures ANOVAs to compare accuracy scores for each type of syllogism and found significant differences only for valid syllogisms. Participants accurately responded to more VB syllogisms in Condition 2 ($F(1.72) = 7.00$, $p = .004$, $\eta^2 = .233$, $M = .98$, $SD = .14$) than in Condition 1 ($M = .83$, $SD = .38$). Similarly, more VU syllogisms were correctly responded to in Condition 2 ($F(1.25) = 7.14$, $p = .008$, $\eta^2 = .237$, $M = .88$, $SD = .33$) than in Condition 1 ($M = .63$, $SD = .48$), or Condition 3 ($M = .87$, $SD = .33$). Regarding IB and IU syllogisms, no differences were found between the mean accuracy scores. However, it is noteworthy that Condition 1 showed a higher number of accurate responses for IB syllogisms ($M = .51$, $SD = .50$) compared to Condition 2 ($M = .44$, $SD = .50$). The same pattern was observed for IU syllogisms, with Condition 1 ($M = .88$, $SD = .32$) having a slightly higher accuracy than Condition 2 ($M = .83$, $SD = .38$). Figure 1 shows the accuracy score for the four syllogism types across the three conditions.

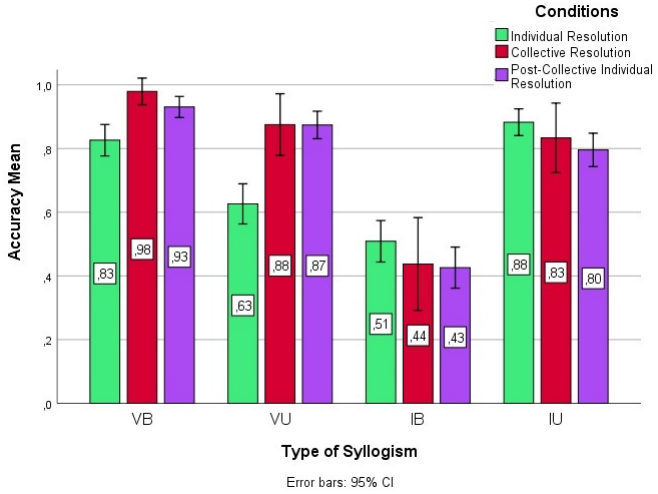

Figure 1: Accuracy scores for each type of syllogism across the three conditions of Study 1.

The overall results of this study show that collective performance surpasses individual performance in reasoning tasks with factual content, supporting Hypothesis 1. However, performance varies depending on the syllogism type: While improvement in collective reasoning was observed for valid arguments, there were no significant differences for invalid arguments.

In Condition 2, 24 groups were formed, 19 of which included five participants, while the remaining five included four participants. The initial investigation focused on the proportions of correct and incorrect responses within each group for each syllogism to determine whether the truth-wins model is the social combination model that best represents the collective decision-making process. Possible responses were classified as correct (C) or incorrect (I). For groups of five members, there are six possible classifications: five correct and no incorrect responses (5-0), four correct and one

incorrect (4-1), and so on. Similarly, for groups of four members, five classifications are possible. Each group was classified for each of the eight syllogisms, resulting in 192 group responses, of which 150 were correct (78.1%).

The pattern of correct group responses aligned with the truth-wins model, as all correct group responses included at least one member who had responded correctly in Condition 1. However, in the 1-4 and 1-3 group classifications, where only one member had responded correctly in Condition 1, the proportion of correct responses was 27%. Notably, in these cases, 75% of the incorrect responses were recorded in one of the IB syllogisms. In fact, the majority model best represents the response pattern in IB syllogisms: In 56.25% of cases, at least two-thirds of coincident individual decisions were necessary and sufficient to provide a collective response. In general, IB syllogisms showed the highest rates of incorrect responses, as noted by Stupple et al. (2011).

The majority model was observed in 69.27% of group responses, while the truth-supported-wins model was met in 77.08% of group responses. The truth-wins model applies to the greatest number of group responses (78.64%), thus confirming Hypothesis 2 for the factual content deductive task. The equiprobability model was not considered, as there are only two possible response options (valid or invalid), and the probability of either occurring is already 50%.

Study 2

We asked participants to evaluate the validity of a series of syllogisms through the same three conditions as in Study 1. In this task, we presented eight categorical syllogisms, each containing ideologically controversial content. Specifically, the conclusions of these syllogisms advocated either in favor of or against abortion. We chose this controversial topic because abortion has been heatedly debated in Argentina in recent years, especially following its legalization in 2020. The syllogism evaluation task included four different types of arguments: two valid pro-abortion (VPro), two valid anti-abortion (VAnti), two invalid pro-abortion (IPro), and two invalid anti-abortion (IAnti) arguments.

Method

Participants To determine an appropriate sample size, we used G*Power software, estimating a required sample of 45 participants based on an effect size of $f = .25$, a significance level of $\alpha = .05$, and a statistical power of .08. Our non-randomized sample consisted of 111 undergraduate psychology students from Argentina, aged between 18 and 47 ($M = 19.65$, $SD = 3.45$). Most participants (81.3%) identified as female, while 18.7% identified as male. Data collection occurred in their classrooms after obtaining informed consent. During this process, we explained the research objectives and assured participants of the confidentiality of their responses.

Design As in Study 1, we employed a within-subjects design in which the three conditions were treated as levels of a within-group factor. The dependent measures included accuracy (correct or incorrect response), type of syllogism (VB, VU, IB, IU), and stance on abortion.

Materials Likert-style Questionnaire: Prior to the argument evaluation task, participants filled out a six-point Likert-style questionnaire to assess their stance on abortion. Based on the scores from this opinion survey, we established a cut-off point. As a result, 8% of participants held an anti-abortion stance, while 92% supported pro-abortion views.

Deductive Reasoning Task: The argument evaluation task included eight syllogisms with ideological content related to abortion, maintaining the same syllogism logical forms used in Study 1. Among these syllogisms, four were valid, and four were invalid, with half leading to pro-abortion conclusions and the other half to anti-abortion conclusions. As in Study 1, participants had two response options: “Yes, it follows logically” or “No, it does not follow logically.”

An English translation of these materials is available at: https://osf.io/ga6q7/?view_only=da65b8667c6f49bc8e71771da13385d1

Procedure At the beginning of the study, all participants were informed about the procedure and objectives and agreed to participate by providing their consent. Subsequently, they completed a series of demographic questions followed by the Likert-style questionnaire. Afterward, participants started to solve the syllogism evaluation task, for which they were given an instruction sheet outlining the concept of logical validity, along with examples of valid and invalid arguments. The instructions, response options, and the three sequential phases (corresponding to the three conditions) were the same as in Study 1. Additionally, participants were not subject to any time constraints when solving the task in any of the three phases.

Results and Discussion

The collected data was analyzed using SPSS v.25.0 software. The raw data set is available at: https://osf.io/ga6q7/?view_only=da65b8667c6f49bc8e71771da13385d1

The eight syllogisms were categorized as believable or unbelievable based on the participants' expressed ideological stance in the Likert-style questionnaire. Subsequently, the ideologically controversial syllogisms were reorganized based on this reclassification. If a participant was in favor of abortion and the syllogism had a valid pro-abortion conclusion, it was categorized as valid believable (VB). Conversely, if a participant expressed an anti-abortion attitude, the syllogism was classified as valid unbelievable (VU). Similarly, an invalid syllogism with an anti-abortion conclusion was reclassified as invalid unbelievable (IU) for participants with pro-abortion ideological stances, and as invalid believable (IB) for participants with anti-abortion ideological stances.

Concerning the global accuracy scores, we performed a within-subject repeated measures ANOVA to compare the

different conditions. The analysis showed a main effect: $F(1.84) = 8.53$, $p < .001$, $\eta^2 = .073$. In particular, the mean accuracy score was significantly lower in Condition 1 ($M = .67$, $SD = .47$) compared to Condition 2 ($M = .77$, $SD = .42$, $p = .001$), and Condition 3 ($M = .73$, $SD = .45$, $p = .03$).

We conducted within-subject repeated measures ANOVAs to evaluate the accuracy scores for each type of syllogism. For the VB syllogisms, participants provided more correct responses in Condition 2 ($M = .92$, $SD = .28$) than in Condition 1 ($M = .66$, $SD = .47$), with a significant effect: $F(1.56) = 29.6$, $p < .001$, $\eta^2 = .214$. Additionally, Condition 3 also had a higher accuracy score compared to Condition 1, with more correct responses observed ($M = .86$, $SD = .34$, $p < .001$). In the case of VU syllogisms, Condition 2 once again resulted in more correct answers ($M = .75$, $SD = .44$) compared to Condition 1 ($M = .65$, $SD = .48$), with a significant difference observed ($F(1.81) = 9.76$, $p < .001$, $\eta^2 = .082$). Condition 3 also outperformed Condition 1, showing improved accuracy ($M = .78$, $SD = .41$, $p = .001$). For IB syllogisms, participants were more accurate in Condition 2 ($M = .75$, $SD = .44$) than in Condition 1 ($M = .62$, $SD = .49$), a statistically significant finding ($F(1.84) = 5.72$, $p = .005$, $\eta^2 = .050$). Condition 3 ($M = .65$, $SD = .48$) also showed fewer correct responses compared to Condition 2 with $p = .003$. However, conversely, IU syllogisms yielded the highest number of correct answers in Condition 1 ($M = .77$, $SD = .42$). This was significantly different from Conditions 2 ($M = .65$, $SD = .48$, $p = .001$) and 3 ($M = .61$, $SD = .49$, $p = .002$), with the results indicating a significant effect: $F(2) = 9.62$, $p < .001$, $\eta^2 = .081$. Figure 2 illustrates the accuracy scores for all syllogism types across the three conditions.

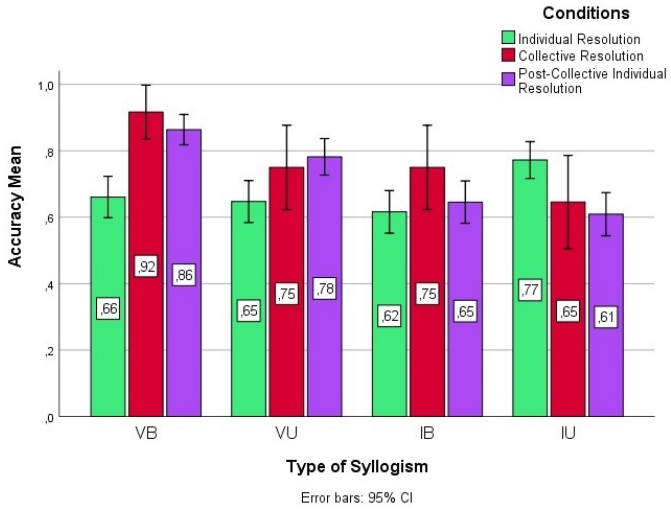

Figure 2: Accuracy scores for each type of syllogism across the three conditions of Study 2.

The overall accuracy results support Hypothesis 1, showing that collective performance surpasses individual performance in the reasoning task involving ideologically controversial arguments. However, performance varies depending on the syllogism type: While improvement in collective reasoning was observed for VB, VU, and IB syllogisms, individual performance was superior for IU syllogisms.

During Condition 2, a total of 25 groups were formed. The groups were composed as follows: 16 groups consisted of five participants, five groups had four participants, and the remaining groups consisted of three participants. Among these groups, 32% had a mixed ideological composition, with at least one member expressing an anti-abortion stance. The remaining 68% had a homogeneous pro-abortion composition. Initially, the proportions of correct and incorrect responses per group were investigated for each syllogism. The aim was to determine whether the truth-wins model could be applied to the collective decision-making process in solving this deductive task.

As in Study 1, the possible responses were two: correct (C) or incorrect (I). There are six possible classifications for groups of five members: five correct and no incorrect responses (5-0), four correct and one incorrect (4-1), and so on. Similarly, for groups of four members, the possible classifications are five: 4-0, 3-1, 2-2, 1-3, and 0-4; and for groups of three members, there are four: 3-0, 2-1, 1-2, and 0-3. Each group was classified for each of the eight syllogisms, resulting in 200 group responses, of which 153 were correct (76.5%).

Of the total correct responses given by the groups, on 17 occasions (11%) only one participant in the group responded correctly in Condition 1. Of these 17 responses, 7 (41%) turned out to be correct. There were 5 cases (3%) where the group responded correctly without any member having responded correctly during the individual resolution phase, that is, cases of group classification 0-5, 0-4, or 0-3. While the majority model was observed in 48.37% of correct group responses, the truth-wins model was only observed in 20.91%. Figure 3 compares these two social combination models across the two studies.

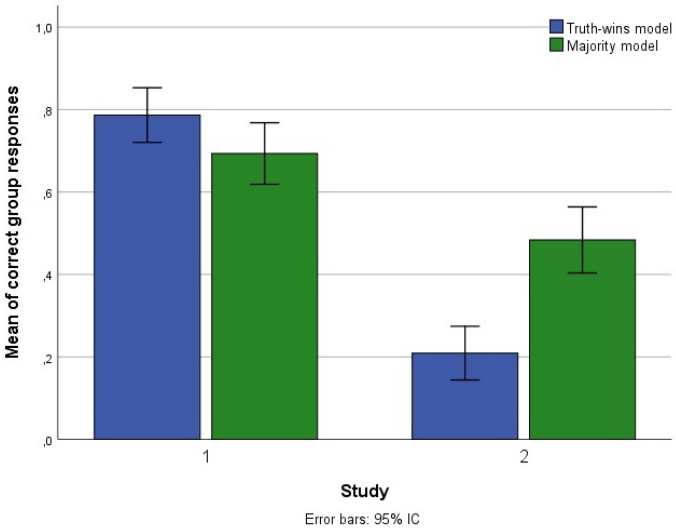

Figure 3: Truth-wins and majority social combination models across Study 1 and Study 2.

Thus, Hypothesis 2 is not supported in this second study, as the truth-wins model does not accurately represent the social combination process underlying the resolution of the deductive task with ideologically controversial content. Conversely, the results indicate that the majority model provides a better description of the collective decision-making process for this reasoning task.

General Discussion

In our two studies, collective resolution outperformed individual resolution, confirming our first hypothesis. These results are consistent with previous research on deductive reasoning that compares individual and collective performance. Therefore, our findings support the social accounts of reasoning advocating for a dialogical origin of deduction. If deductive reasoning is a social activity, reasoning is expected to benefit from a dialogical setting. These are precisely the findings that our two studies reveal.

However, there were differences in performance in the two studies. Study 1 shows superior collective performance on valid arguments, but not on invalid ones. This contrast is particularly noticeable with conflict syllogisms involving arguments where belief and logic conflict, such as VU and IB syllogisms. Conversely, Study 2 reported improvement in collective reasoning for both valid syllogisms (VB and VU) and IB syllogisms. As a result, collective performance surpasses individual performance for both valid and conflict syllogisms when reasoning with ideologically controversial content. These findings align with previous research that emphasizes the distinctions between belief bias and myside bias, as well as the influence of content on human reasoning.

These findings could be explained within dual-process theories (Evans, 2017). Arguably, collective reasoning enhances the intervention of type 2 processes in deductive tasks. This enhancement is evident in the improved accuracy in responses to valid arguments during the collective resolution process in Study 1 and to valid and conflict arguments in Study 2. The decline in accuracy for IB syllogisms in Study 1, related to collective performance, may be attributed to more reasoners searching for a model that supports believable conclusions, as all reasoners consider those conclusions believable. However, in Study 2, where reasoners hold different views on abortion, not all of them regard those conclusions as believable. Consequently, enhancements in collective resolution may stem from some reasoners seeking models that refute unbelievable conclusions.

As regards Hypothesis 2, the obtained data indicates that the truth-wins model is the most suitable social combination model to account for the group resolution of the syllogism evaluation task of Study 1. It is important to emphasize that while the correct response of at least one group member was necessary for the entire group to respond correctly, it was not a sufficient condition. These results support Hypothesis 2 for the case of deductive tasks with factual content. Nevertheless, the majority model most accurately describes the collective decision-making process in resolving the deductive task from Study 2. These results challenge our second hypothesis and differ from previous studies, which show that the truth-wins model best describes the resolution of intellective tasks (Laughlin, 2011). The underlying idea of the truth-wins model suggests that the group member who best understands the logical structure of the deductive problem can convince the rest of the group members to accept their viewpoint. However, we did not observe this social combination model in our second study. It could be hypothesized that since the deductive task used in our second study involves ideologically controversial content, such as the stance on abortion, this may have influenced how the group approached its collective resolution. Thus, given the controversial content of the intellective task, the ideological stance of most of the group members could have played a central role in shaping the collective decision-making process. Hence, instead of convincing their partners of the correct response to the problem, group members could have tried to convince them of their ideological stance. This effect was not observed in the deductive task of Study 1, involving factual content, as every group member shared the same testable beliefs. It is important to note that none of the previous studies incorporated ideologically controversial content in their discussion of collaborative reasoning. This possible explanation, however, is highly speculative and would require further studies specifically designed to test this hypothesis.

Unlike the results obtained in the studies of Boku et al. (2018) and Moshman and Geil (1998) regarding the Wason selection task, in our first study, there were no cases where the group responded correctly if no member had responded correctly during the individual resolution phase (Condition 1), specifically in the cases of the 0-5 or 0-4 group classification. This finding underscores the importance of at least one group member possessing the correct response so that the group can arrive at the correct answer. In the case of Study 2, there were five instances where no individual participant provided the correct answer. However, during the collaborative reasoning phase, the groups formed by these participants were able to reach the correct answer. In contrast with the results obtained in Study 1, data from Study 2 aligns with the results of Boku et al. (2018) and Moshman and Geil (1998).

Something worth noting is that one limitation of our study is the use of a convenience sample, which has resulted in a predominantly female test group. However, the major constraint of our research can be observed in Study 2, where 92% of participants held a pro-abortion stance, leading to a lack of variability in the sample regarding perspectives on that topic. To address this, future research should aim for a more diverse sample.

Acknowledgments

This research was funded by ANPYCT (Argentinian Agency for the Promotion of Scientific Research and Technological Development), PICT 2021 N°I-INVI-00722 “Aspectos normativos del razonamiento humano: un abordaje filosófico

y cognitivo", and by UCC (Universidad Católica de Córdoba), Research Project: "Lógica y psicología del razonamiento: un análisis filosófico y empírico sobre sesgos cognitivos".

References

Abelson, R. (1986). Beliefs are like possessions. *Journal for the Theory of Social Behaviour*, 16 (3), 223–250.

Ball, L. J., & Thompson, V. A. (2018). Belief bias and reasoning. In L. Ball & V. Thompson (Eds.), *The Routledge international handbook of thinking and reasoning*. London: Routledge.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning*, 1, 221-135.

Boissin, E., Caparos, S., & De Neys, W. (2023). No easy fix for belief bias during syllogistic reasoning? *Journal of Cognitive Psychology*, 35(4), 401–421.

Boku, M., Yama, H. & Mercier, H. (2018). Robust improvements on reasoning performance following discussion in Japan. *Japanese Psychological Research*, 60(1), 47-53.

Castelain, T., Bernard, S., Van der Henst, J.-B. and Mercier, H. (2016). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science*, 19(6), 957-966.

Čavojová, V., Šrol, J., & Adamus, M. (2018). My point is valid, yours is not: Myside bias in reasoning about abortion. *Journal of Cognitive Psychology*, 30(7), 656–669.

de Chantal, P., Newman, I., Thompson, V. & Markovits, H. (2020). Who resists belief-biased inferences? The role of individual differences in reasoning strategies, working memory, and attentional focus. *Memory & Cognition*, 48 (4), 655–671.

Dominowski, R. (2019). Content Effects in Wason Selection Task. In S. Newstead & J. Evans (Eds.), *Perspectives on thinking and reasoning. Essays in honour of Peter Wason*. London: Routledge.

Dutilh-Novaes, C. (2021). *The dialogical roots of deduction*. Cambridge: Cambridge University Press.

Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128(6), 978–996.

Evans, J. (2017). Belief bias in deductive reasoning. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory*. Abingdon: Routledge.

Heyes, C. (2018). *Cognitive Gadgets. The cultural evolution of thinking.* Cambridge, MA: The Belknap Press of Harvard University Press.

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1(2), 134–148.

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press.

Laughlin, P. R., & Adamopoulos, J. (1980). Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology*, 38(6), 941–947.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, 22(3), 177–189.

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, 30(1), 81-108.

Mercier, H., Deguchi, M., Van der Henst, J.-B., & Yama, H. (2016). The benefits of argumentation are cross-culturally robust: The case of Japan. *Thinking & Reasoning*, 22(1), 1–15.

Moshman, D. (2021). *Reasoning, argumentation, and deliberative democracy*. New York: Routledge.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, 4(3), 231–248.

Šrol, J. (2022). Individual differences in epistemically suspect beliefs: The role of analytic thinking and susceptibility to cognitive biases. *Thinking & Reasoning*, 28(1), 125-162.

Stanovich, K. (2021). *The bias that divide us. The science and politics of myside thinking.* Cambridge, MA: The MIT Press.

Stanovich, K., West, R. & Toplak, M. (2013). Myside bias, rational thinking and intelligence. *Current Directions in Psychological Science*, 22(4), 259-264.

Stupple, E. J., Ball, L. J., Evans, J. S. B., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, 23(8), 931-941.

Tomasello, M. (2022). *The evolution of agency. Behavioral organization from lizards to humans*. Cambridge, MA: The MIT Press.

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.

Unsworth, S. & Medin, D. (2005). Cultural Differences in Belief Bias Associated with Deductive Reasoning? Cognitive Science, 29(4), 525-529.

Wason, P. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Vol. 1. Harmondsworth: Penguin.