Who Gains More: Experts or Novices? The Benefits of Interaction under Numerical Uncertainty

Francesco Sella[a,*], Robert Blakey[a,*], Dan Bang[abcd], Bahador Bahrami[ef], & Roi Cohen Kadosh[a]

[a]Department of Experimental Psychology, University of Oxford, Oxford, UK
[b]Calleva Research Centre for Evolution and Human Sciences, Magdalen College, Oxford, UK.
[c]Interacting Minds Centre, Aarhus University, Aarhus, Denmark.
[d]Wellcome Trust Centre for Neuroimaging, University College London, London, UK.
[e]Institute of Cognitive Neuroscience, University College London, London, UK
[f]Faculty of Psychology and Educational Sciences, Ludwig Maximilian University, Munich, Germany

* These authors equally contributed.

Word count: 5997

Corresponding author:

Francesco Sella
sella.francesco@gmail.com
francesco.sella@psy.ox.ac.uk
Department of Experimental Psychology
University of Oxford
Office D154 - Tinbergen Building
9 South Parks Road
OX1 3UD - Oxford, UK

Abstract

Interacting to reach a shared decision is an omnipresent component of human collaboration. We explored the interaction between dyads of individuals with different levels of expertise. The members of the dyads completed a number line task privately, jointly and privately again. In the joint condition, dyad members shared their private estimates and then negotiated a joint estimate. Both dyad members averaged their private individual estimates to determine joint estimates, thereby showing a strong equality bias. Their performance in the joint condition exceeded the performance of the dyad's best estimator, demonstrating interaction benefit, only when the dyad members had similar levels of expertise and when the averaged dyad performance was sufficiently accurate. At the end of the task, participants rated their and their partner's level of competence. Participants were accurate in classifying themselves as the expert or the novice within the dyad. Nevertheless, novices tended to overestimate their ability as they admitted to being less competent but only slightly worse than their expert partner. Experts, instead, believed themselves to be more competent but were humble and considered their performance only marginally better than their partner. Overall, these results have important implications for settings in which people with different levels of expertise interact.

Keywords: Decision making, Numerical cognition, Interactive minds, Expertise, Number line.

Statement of the public significance:

Interaction to reach a shared decision is an omnipresent component of human collaboration, and is presented in different scenarios from the classroom, to the workplace, and the army. Among pairs of individuals (dyads) with different levels of expertise, it is desired that novices will profit the most from collaborating with an expert. We showed that the two members of the dyad, regardless of their level of expertise, greatly adopted an averaging strategy assigning an equal weight to their estimates. Interaction was beneficial when the dyad members had similar levels of expertise and when, on their own, they were sufficiently accurate. Moreover, novices tended to overestimate their ability as they admitted to being less competent but only slightly worse than their partner. Experts, instead, believed themselves to be more competent but were humble and considered their performance only marginally better than their novice partner.

Acknowledgments. We are grateful to the Editor and three anonymous reviewers for their helpful and constructive comments, which helped us to improve the manuscript. We also thank Mihaela Duta for programming the number line task. This work has been supported by the European Research Council (Learning&Achievement 338065). BB was supported by the European Research Council (starting grant, NEUROCODEC, #309865)

Introduction

Many group decisions are made by counting the number of votes in favour of each available option – with the option favoured by most votes being the winner. Under this rule, each vote carries the same weight, regardless of the voter's expertise. However, in everyday life, we often weight the opinions of others' according to their estimated level of expertise. Weighting opinions appropriately is especially important when individuals with different levels of expertise are brought together to reach a group decision (Grofman, Owen, & Feld, 1983). This situation arises in many learning and educational contexts, such as school classrooms where students with different abilities are brought together to work on a common project. Here, we tested how individuals with different levels of expertise combine (continuous) estimates into a joint estimate in the context of a numerical task and tested whether the benefit of interaction differs for novices and experts.

Research on judgement and decision-making has shown that (simple) averaging (or the median) of continuous estimates obtained from different individuals can improve accuracy in a range of domains – often over and above the single-best individual estimate (Armstrong, 2001; Clemen, 1989; Galton, 1907; Jacobson, Dobbs-Marsh, Liberman, & Minson, 2011; Minson, Liberman, & Ross, 2011). As a rule-of-thumb, averaging unbiased but noisy estimates causes uncorrelated errors to cancel out (Bruce, 1935; Eysenck, 1939; Galton, 1907; Gordon, 1924, 1935; Preston, 1938; Smith, 1931). Averaging, however, only works well when the individual estimates are equally likely to fall on either side of the truth – a principle known as the 'bracketing principle'. When the individual estimates are biased towards the same side of the truth (overestimation or underestimation by both dyad members), the average estimate is, at best, as accurate as the single-best individual estimate (Jacobson et al., 2011; Larrick & Soll, 2006; Minson et al., 2011; Soll & Larrick, 2009).

Averaging can be beneficial not only for groups but also for single individuals. For example, the average of two estimates from the same participant tend to be more accurate than either estimate (Vul & Pashler, 2008). This within-individual benefit may arise when the two estimates are ‘incorrect’ in different ways – for example, because the estimates are based on different sources of knowledge (Herzog & Hertwig, 2009). This ‘dialectal bootstrapping’ can be induced by asking people for a second opinion after considering the opposite stance (Lord, Lepper, & Preston, 1984) or reasoning why their initial estimate might be incorrect (Hoch, 1985). Individuals may also spontaneously engage in dialectal bootstrapping after having had experience with combining the estimates of others with their own. For example, Liberman et al. (2012) found that, after dyad members had interacted about joint estimates, subsequent individual estimates reached a high level of accuracy – a result which does not depend on feedback (Minson et al., 2011) or task expertise (Jacobson et al., 2011). The question remains, however, whether interacting individuals would use weighted averaging to combine continuous estimates in the face of differences in expertise.

A recent set of studies, using visual psychophysics, have shown that dyad members only obtain a true *interaction benefit* – that is, when joint decisions are more accurate than those of the best dyad members – when they have similar levels of expertise (Bahrami, Didino, Frith, Butterworth, & Rees, 2013; Bahrami et al., 2012a, 2012b; Bahrami et al., 2010; Bang et al., 2014; Mahmoodi, Bang, Ahmadabadi, & Bahrami, 2013). This pattern of responses is thought to arise because dyad members assign equal weights to their opinions (akin to simple averaging), even when they are explicitly informed about their differences in expertise or offered monetary incentives to maximise their joint accuracy (Mahmoodi et al., 2015). This “equality bias” might be of a social nature; for example, the “novice” may insist on playing a role, or the “expert” may feel obliged to treat the novice as an equal (Harvey & Fischer, 1997; Mahmoodi et al., 2015). It might also arise because people often are bad at identifying the relative expertise of different group members (Henry, 1995; Miner, 1984; Trotman, Yetton, & Zimmer, 1983); for example, expertise is often judged from cues such as confidence and tendency to talk (Littlepage, Robison, & Reddington, 1997), which correlate poorly with expertise in some domains (Klayman, Soll, González-Vallejo, & Barlas, 1999).

Here we sought to bring together, on one hand, the research on estimation strategies and, on the other hand, research on the effect of differences in expertise on joint accuracy. To induce differences in expertise, participants with backgrounds in either maths or humanities (combination of participants: Maths-Maths, Maths-Humanities, and Humanities-Humanities) estimated the location of different numbers on a number line privately, together and privately again. Within each dyad, we classified participants as expert or novice based on their performance on the number line task (Sella, Sader, Lolliot, & Cohen Kadosh, 2016; Siegler & Opfer, 2003).

First, we asked how the benefit of joint performance varied with individual expertise. We would expect the joint responses to be more accurate than the pre-interaction estimates made by novices, whereas we would expect such an improvement to be smaller for experts (Minson et al., 2011). Nevertheless, the dyadic interaction might bring a benefit (i.e., joint decisions might be more accurate than those of the best dyad member) when the two members of the dyads have similar levels of expertise (Bahrami et al., 2013; Mahmoodi et al., 2013). Second, we asked whether novices were more swayed by the experts – or whether the use of a simple averaging strategy would prevail. Lastly, we asked how any post-interaction benefit varied with individual expertise. We would expect that novice's post-interaction performance is higher than their pre-interaction performance, whereas such an improvement would be limited or possibly absent for experts (Liberman et al., 2012). For example, novices might recognise their poor performance in comparison with experts and then adjust their mapping strategy when performing the number line task in the post-interaction condition.

Method

Participants

Sixty university students (26 males; M_{age}=20.68, *SD*=2.45) from the faculties of maths and humanities took part in the present study. The sample size was doubled compared to a previous study exploring individuals' interaction in the numerical domain (Bahrami et al., 2013; Experiment 1). The maths (*n*=31) and humanities (*n*=29) students were assigned to a dyad according to their availability, resulting in 8 maths-maths, 7 humanities-humanities and 15 maths-humanities dyads. Participants did not know each other in advance and were compensated for their time (approximately 1 hour 20 minutes) with £5 and entry into a lottery to win £100. The study was approved by the Medical Sciences Inter Divisional Research Ethics Committee at University of Oxford.

Tasks

The Number line task (Sella, Sader, Lolliot, & Cohen Kadosh, 2016; Siegler & Opfer, 2003). This task required participants to locate a target number on a visual horizontal line, where only the extremes of the line were labelled. The number line task was originally implemented by Siegler and Opfer (2003) to investigate numerical estimation in children and adults. It has been repeatedly found that the pattern of estimates shifts from a biased (log-like) to an accurate (linear) mapping when children increase their numerical knowledge and experience with the proposed numerical intervals (Siegler & Booth, 2004). For instance, second graders display an accurate mapping when placing numbers in the interval 0-100 but still display a biased mapping with the interval 0-1000. The biased pattern was originally thought to resemble the logarithmic compression of the mental number line (Dehaene, 2003). However, several alternative theoretical accounts have been proposed to explain the shift from biased to linear mapping (Barth & Paladino, 2011; Cohen & Sarnecka, 2014; Hurst, Leigh Monahan, Heller, & Cordes, 2014; Moeller, Pixner, Kaufmann, & Nuerk, 2009). In particular, it has been suggested that the performance in the number line task can be interpreted as a proportional judgment (Barth & Paladino, 2011). Accordingly, adults seem to implement a proportional estimation strategy, which usually leads to high level of accuracy (Cohen & Blanc-Goldhammer, 2011; Sullivan, Juhasz, Slattery, & Barth, 2011), even though individual differences can emerge. For instance, mathematicians outperformed non-mathematicians when mapping positive numbers on the line, albeit this relation is fully explained by visuospatial skills (Sella et al., 2016). Similarly, those with better mental rotation abilities are able to perform more accurately on the number line task (Thompson, Nuerk, Moeller, & Cohen Kadosh, 2013). Overall, the task requires knowledge of the symbolic numerical system, the representation of symbolic numerical quantities, strategies to map numbers onto space and visuospatial skills (Sella et al., 2016; Sullivan et al., 2011).

Each trial began with the presentation of an unmarked number line extending from -1000 to +1000, with the same target number displayed above both extreme ends (Figure 1). On each LCD monitor, the white number line was presented on a black background across 1000 pixels; each key press moved the slider by one pixel, which corresponded to two numerical magnitudes on the number line. The monitors were identical, widescreen (51cm x 32cm) and had a spatial resolution of 1920 x 1200 pixels. We created the task in Matlab, using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

The task was composed of four sequential conditions: pre-interaction, to-be-shared, joint and post-interaction (Figure 1). Forty-eight different target numbers were presented for each condition, except for the to-be-shared and joint conditions in which the same target numbers were displayed. To avoid a repetition effect, different target numbers were presented in the pre-interaction, to-be-shared, and post-interaction conditions. In each condition, half of the target numbers were positive and half were negative, though the same absolute values were used for both halves. None of the numbers were divisible by five and none fell within 100 digits of the endpoints (-1000 and +1000). Numbers that met these criteria were selected pseudorandomly and presented in a pseudorandom order. The magnitude of target numbers was similar across conditions (Pre-interaction: *M*=470, *SD*=257; To-be-shared: *M*=408, *SD*=228; Post-interaction: *M*=390, *SD*=205). Every participant was presented with the same numbers in the same order. Two practice trials were presented at the beginning of each condition.

In the pre-interaction condition (Figure 1a), participants privately estimated the location of the target numbers on the line by pressing the left and right arrow keys to move their slider to the desired position. All sliders were initially presented at the extreme left or right of the number line on negative and positive target trials respectively.

---------------------------- Please insert Figure 1 here --------------------------

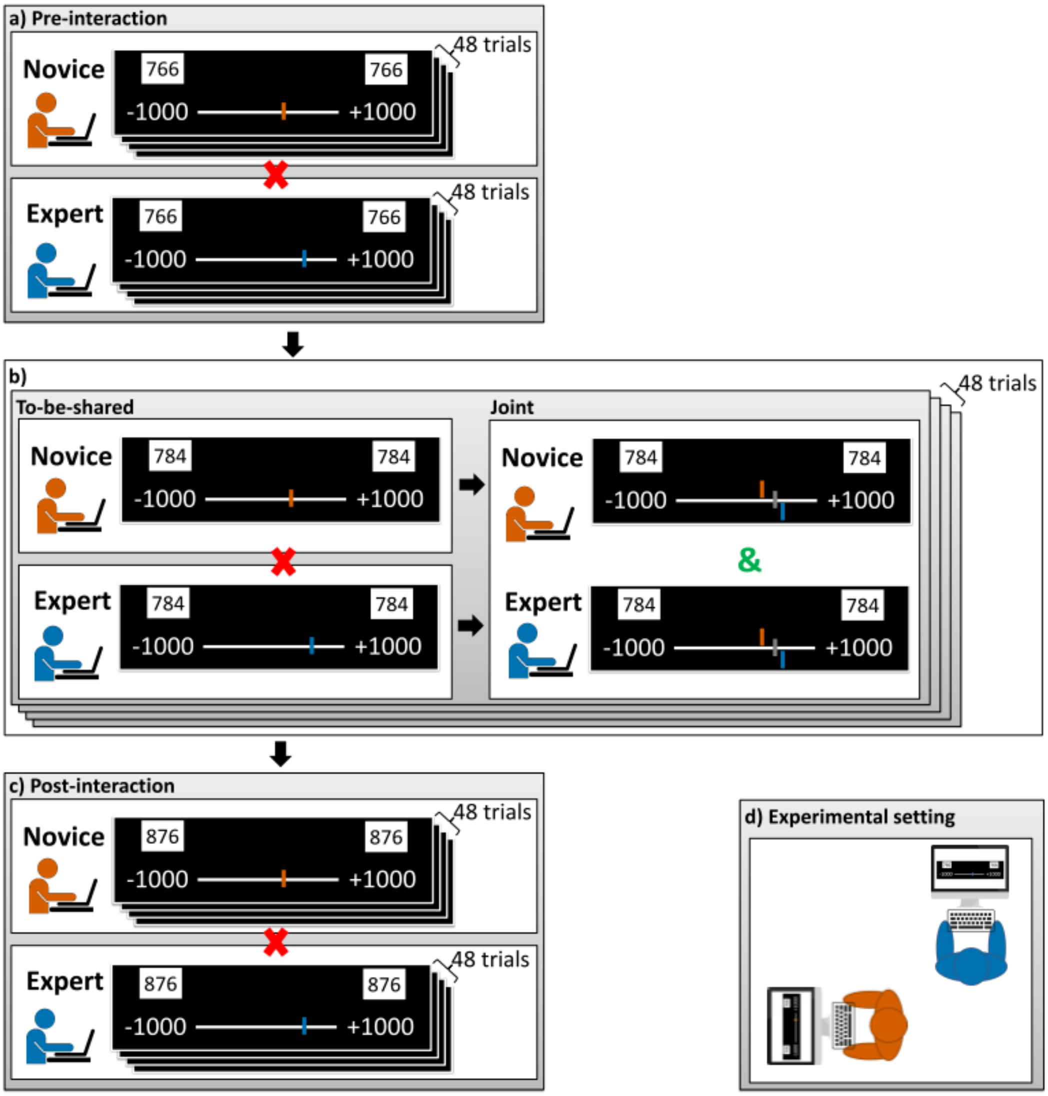

Figure 1. Diagram of the number line task. a) An example of a number line task trial in the pre-interaction condition: The novice and expert completed 48 trials without interacting. b) An example of the to-be-shared and joint conditions: in the to-be-shared trial, the novice and expert privately estimated the position of a target number (in white boxes) on the line; thereafter, in the joint trial, participants could see their estimates from the previous trial (i.e., a to-be-shared one), distinguished by red and blue colours on the same number line. Then, the two members of the dyad alternately adjusted the slider to reach a shared joint estimate. c) An example of a trial in the post-interaction condition: novice and expert completed 48 trials without interacting. d) The configuration of monitors and participants in the testing room.

Upon completion, participants pressed the ‘Next’ key (i.e., the “b” key for the red participant and the “up arrow” key for the blue participant; the keys were labelled “Next” with stickers), received no feedback and waited for their partner to finish before progressing onto the next trial. In order that individuals did not recognise whether they were the slower or faster member of the dyad, participants were informed that they were waiting for the computer to load the next number line. A random delay (varying between 1 and 3 seconds) was inserted between the point at which the slowest dyad member pressed 'Next' and onset of the next trial.

In the to-be-shared condition (Figure 1b), the nature of the task was identical to the pre-interaction condition except participants were aware that the next trial was a joint one. In the joint condition, both dyad members were simultaneously shown the locations of their own and their partner's estimates from the previous trial (i.e., a to-be-shared one), distinguished by different colours on the same number line (i.e., blue and yellow, even though in the figures we used red and blue to increase discriminability). One of the dyad members (A from hereon) was handed control of a joint slider, which was also presented on the same number line. It was alternated across trials whether dyad member A or B was handed first control of the joint slider. Participant A was encouraged to discuss the desired location of the joint response, move the slider to this location and then press 'Next', after which control of the joint slider was handed to participant B. Participant B either adjusted the joint slider's location (in the case of disagreement) or not (in the case of agreement) and then pressed 'Next' to hand control back to participant A. This turn-taking process continued until both the participants stopped readjusting the slider before pressing 'Next'. The entire process was displayed to both dyad members on each of their monitors simultaneously.

Finally, participants completed the post-interaction condition (Figure 1c) in which the nature of the task was identical to the pre-interaction condition, except that participants were aware that this was the last phase of the task. For each estimate, we calculated the absolute deviation (see formula below) as a measure of error, as was done in previous studies (Sella et al., 2016).

$$absolute\ deviation = |estimate - target\ number|$$

Competence scale. Participants estimated theirs and their partner's competence in the computational estimation task and in the number line task separately by writing down a whole number between 1 and 100, where 1 indicates that estimates were extremely inaccurate/very far from the true answer and 100 indicates that estimates were extremely accurate/very close to the true answer. One dyad did not complete the competence scale then the following analysis is limited to 29 dyads.

Procedure

Participants reported their field of study, together with demographic information (age, gender, university, degree type, year of study, nationality and ethnicity), on the questionnaire provided. Participants were informed of whether their dyadic partner was studying maths or one of the humanities. Throughout the study, participants sat in the same room at right angles to one another, each with their own keyboard and monitor (Figure 1d). After completing the demographic questionnaire, participants completed the first ten trials of a computational estimation task and then the conditions (pre-interaction, to-be-shared, joint, and post-interaction) of the number line task. At the end of the number line task, participants responded to the remaining ten trials of the computational estimation task. Finally, participants estimated theirs and their partner's competence in the computational estimation task and in the number line task using the competence scale. The results from the computational estimation task are reported in the supplementary materials. Specifically, the pattern of results (see below) remained stable when we categorised participants in expert and novices using the performance in the computational estimation task (see supplementary materials). The data can be found at https://osf.io/wnzhm/?view_only=f92fa71c4a88452fb32388b92c6b94f6.

Results

Statistical analyses were conducted using the free software JASP (JASP Team, 2017) using default priors and R (R Core Team, 2014) along with the package ggplot2 for graphs (Wickham, 2009). We ran Bayesian analyses and reported Bayes factors (BF_{10}) expressing the probability of the data given H1 relative to H0 (i.e., values larger than 1 are in favour of H1 whereas values smaller than 1 are in favour of H0). In case of model comparison, we computed the BF as the ratio between the BF_{10} of the first model and the BF_{10} of the second model (i.e., values larger than 1 are in favour of the first model whereas values smaller than 1 are in favour of the second model). We described the evidence associated with BFs as "anecdotal" (1/3 < BF < 3), "moderate" (BF < 1/3 or BF > 3), "strong" (BF < 1/10 or BF > 10), "very strong" (BF < 1/30 or BF > 30), and "extreme" (BF < 1/100 or BF > 100) (Jeffreys, 1961). Results from NHST approach and associated *p*-values can be found in the supplementary materials.

Participants

There was anecdotal evidence for a difference in age and the odds of being British/other nationality, Caucasian/other ethnicity or an undergraduate/postgraduate between maths and humanities students or between (see below) experts and novices (all BFs_{10} between 0.31 and 2.19). There was strong evidence for higher odds of being male for maths students than for humanities students, $BF_{10} = 20$. However, there was anecdotal evidence for a difference in the odds of being male between experts and novices, $BF_{10} = 0.36$.

Number line task

For each participant in each task condition, we calculated the mean absolute deviation and removed those trials below and above three standard deviations (percentages of trials removed: 1% in the pre-interaction condition, 0.76% in the to-be-shared condition, 0.35% in the Joint condition, 0.97% in post-interaction condition).

We classified participants as experts and novices based on their performance in to-be-shared condition of the number line task because this condition represents the estimates that participants were about to share with their partner, thereby determining the actual level of expertise within the dyad. Hence, the more accurate performer within each dyad was labelled the expert (*n*=30) and the less accurate member was labelled the novice (*n*=30).

In the joint condition, experts and novices evaluated their estimates from the to-be-shared trial and then negotiated a joint estimate. Consequently, both experts and novices obtained the same absolute deviations in the joint condition. Nevertheless, to be thorough, we analysed the absolute deviation (log base-10 transformed) in a Bayesian mixed ANOVA with Condition [pre-interaction, to-be-shared, joint, post-interaction] as a within-subjects factor and Expertise [Expert, Novice] and Dyad type [Maths-Maths, Maths-Humanities, Humanities-Humanities] as the between-subjects factors. The inclusion of Dyad type into the model led to an inclusion Bayes factor of 0.095 (this BF is obtained from averaging BFs from all the models including a specific effect, compared to all models that do not include the effect), therefore the Dyad type was removed from the model. There was extreme evidence in favour of the model including the interaction Condition x Expertise compared to the model with the two main effects (BF>100). We compared experts' and novices' absolute deviations in the pre-interaction, to-be-shared, and post-interaction conditions as well as observing within each group whether there were relevant differences between pre-interaction and to-be-shared, to-be-shared and joint, joint and post-interaction, and pre-interaction and post-interaction conditions using Bayesian *t*-tests (Figure 2). The experts displayed less absolute deviation in the pre-interaction (Experts: *M*=1.65, *SD*=0.13; Novices: *M*=1.81, *SD*=0.12; $BF_{10}>100$, extreme evidence) and post-interaction conditions compared to novices (Experts: *M*=1.57, *SD*=0.12; Novices: *M*=1.74, *SD*=0.16; $BF_{10}>100$, extreme evidence). Both groups displayed a reduction in absolute error from the pre-interaction to the to-be-shared conditions (Experts: $BF_{10}=72$, very strong evidence; Novices: $BF_{10}>100$, extreme evidence). Only novices displayed a reduced absolute deviation in the joint condition compared to the to-be-shared condition (Joint condition, *M*=1.55, *SD*=0.12; Novices: $BF_{10}>100$, extreme evidence) whereas experts' absolute deviation remained stable (Experts: $BF_{10}=0.21$, moderate evidence). Novices also showed an increase in absolute deviation in the post-interaction condition compared to the joint condition (Novices: $BF_{10}>100$, extreme evidence) whereas experts' performance appeared to remain stable (Experts: $BF_{10}=0.25$; moderate evidence). Finally, only experts displayed a strong reduction in absolute deviation in the post-interaction condition compared to the pre-interaction condition (Experts: $BF_{10}=24$, strong evidence) whereas this difference appeared to be modest for Novices (Novices: $BF_{10}=4.77$; moderate evidence). A detailed comparison of experts' and novices' spatial mapping can be found in the supplementary materials.

---------------------------- Please insert Figure 2 here --------------------------

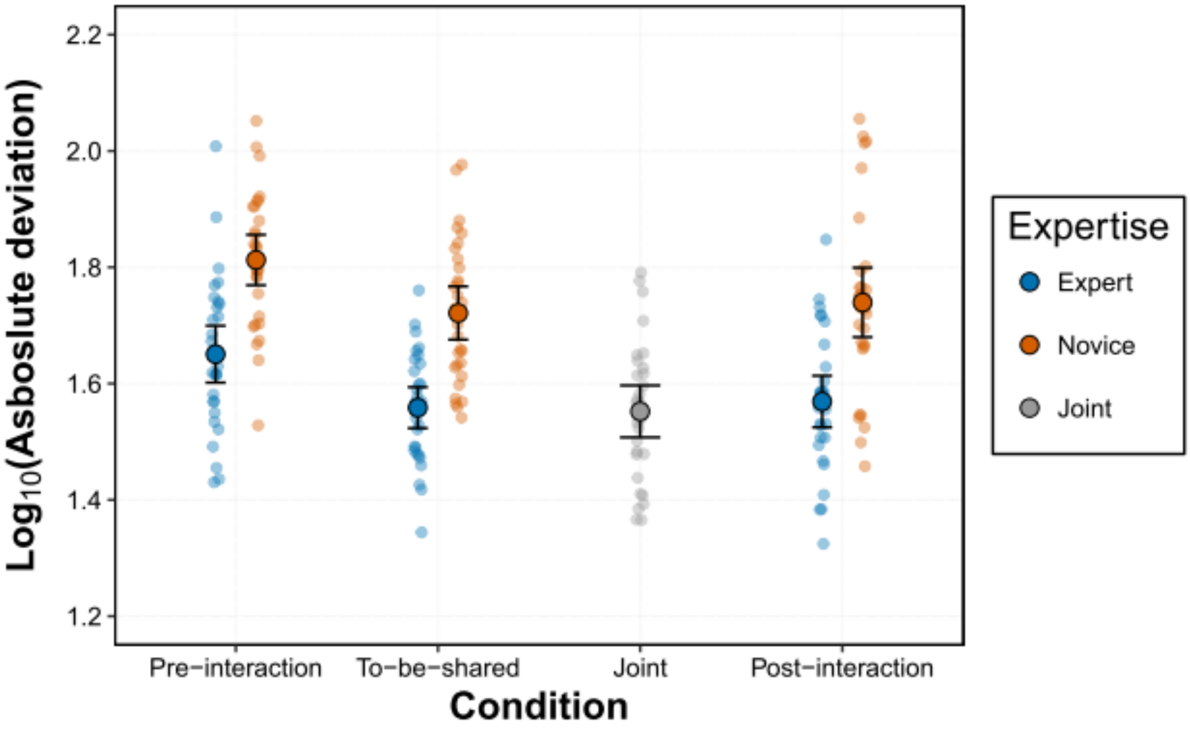

Figure 2. Mean absolute deviation ($\log_{10}$ transformed) in the number line task for each condition separately for experts (blue dots) and novices (red dots; error bars represent 95%CIs).

Joint strategy

In the following analyses, we aimed to individuate whether experts or novices led the positioning of the joint estimates on the line. Additionally, we investigated the distribution of joint estimates in order to assess the presence of averaging.

In the joint condition, each member of the dyad saw the other's member estimate from the to-be-shared trial and then both members negotiated a joint estimate. We removed from the analysis eleven trials, in which the expert and novice provided exactly the same estimate in the to-be-shared condition and therefore no negotiation was required to obtain their joint response. For each joint trial, we defined the decision maker (i.e. the main contributor) as the participant who moved the slider more (as measured in pixels). In 14% of trials, each member of the dyad moved the slider the same number of pixels along the line, so their contribution was equal (i.e., "equal contribution" in the same trial). The experts moved the mouse cursor more pixels compared to the novices in 43% of trials, whereas in the remaining 43% of trials the novices moved the mouse cursor more pixels. Therefore, the two members of the dyad contributed equally to the joint estimates across trials.

Figure 3 shows the percentage of estimates in the joint condition as a function of the proportional distance between the novice's (x=0) and the expert's (x=1) estimates in the to-be-shared condition. For instance, consider a trial in which the target number is 100 and the novice places the cursor on 50 whereas the expert places the cursor on 120. Then, in the joint condition, the two members of the dyad agree in placing the cursor on 80. The distance between the two estimates is 70 (i.e., 120-50), the distance between the joint estimate and the novice's estimate is 30 (i.e., 80-50) and the distance between the joint estimate and the expert's estimate is 40 (i.e., 120-80). Similarly, the distance between the correct position and the novice's estimate is 50 (i.e., 100-50) whereas the distance between correct position and the expert's estimate was 20 (i.e., 120-100). Then, we set the novice's estimate to be 0 and the expert's estimate to be 1, therefore the distance between the two estimates, which is 70, is now set to 1. Thereafter, the distance between the novice's estimate and the joint estimate is rescaled to 0.43 (i.e., 30/70) whereas the distance between the expert's estimate and the join estimate is the reciprocal, 0.57. The distance between the correct position and the novice's estimated is rescaled to 0.71 (i.e., 50/70). Consequently, when the value on the *x*-axis is zero or one, the estimate in the joint condition equalled the novice's or expert's estimate in the to-be-shared condition, respectively. It was found that 98% of the joint estimates fell within the interval created by the two members' estimates from the to-be-shared condition. Specifically, most of the joint estimates were placed halfway between members' estimates or on one of the member's estimate. The joint estimates were rarely (2% of trials) placed outside the obtained interval, either on the novice's or expert's side, even though the target position was outside the interval in 60% of the trials. Notably, experts and novices were both responsible for averaging in the joint condition.

---------------------------- Please insert Figure 3 here --------------------------

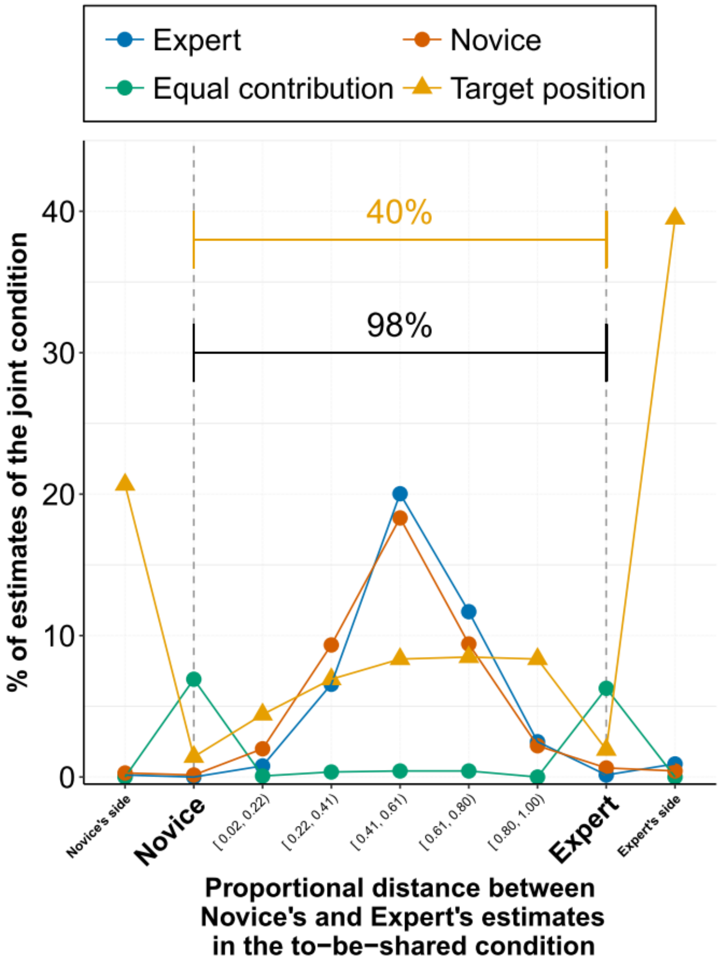

Figure 3. The percentage of trials in the joint condition that were placed within and outside the numerical interval created by the novice's (x=0) and expert's (x=1) estimates in the to-be-shared condition, subdivided by the trials in which the expert or the novice mainly contributed to the joint estimate. The classification of novices and experts was based on their performance in the to-be-shared condition of the number line task. In 14% of trials, the two members of the dyad moved the slider the same number of pixels along the line, so their contribution was equal (green dots and line). In the remaining 86% of trials, experts (blue dots and line) contributed most to half of the joint estimates and novices contributed most to the other half (red dots and line).The pattern of results indicates that both expert and novice displayed a largely overlapping mapping strategy with strong use of averaging. The orange triangles represent the target position with respect to the interval created by the expert's and novice's estimate in the to-be-shared condition. In 40% of trials of the joint condition, the position of the target fell within the interval created by the novice's and expert's estimates in the to-be-shared trial. In 98% of trials of the joint condition, experts

and novices decided to place the joint estimate within the interval they created in the to-be-shared trial.

Interaction benefit

Here we analysed whether the dyad interaction led to a benefit and whether such a benefit could be explained by similarity in performance between the two members of the dyad and their average performance as dyad.

For each dyad, we calculated the interaction index as the ratio between the dyad's absolute deviation in the joint condition and the dyad's best estimator's (BE) absolute deviation calculated on all the trials of the to-be-shared condition. With a ratio larger than one, the joint performance of the dyad was worse than the performance of the BE in the dyad, indicating that the interaction led to a loss in performance. Conversely, with a ratio lower than one, the dyad's performance exceeded that of the BE in the dyad, indicating that the interaction led to a benefit in performance. The mean obtained interaction index was 1 (*SD*=0.19), indicating that the performance of the dyad in the joint condition tended to match the performance of the dyad's BE in the to-be-shared condition and did not vary as a function of the dyad type (BF_{10}=0.28, moderate evidence). In line with previous literature (Bahrami et al., 2010), we investigated whether the interaction benefit index varied as a function of the similarity in performance of the two dyad members in the to-be-shared condition. The similarity index was calculated as the ratio of the worst to the best estimators' (z-scored) absolute deviation, with higher values indicating higher differences in individual performance. We also investigated whether the mean absolute deviation (z-scored) of the two members of the dyad in the to-be-shared condition explained any variance in the interaction index or may change depending on similarity. It is possible that the effect on similarity on the interaction benefit may change depending on the average ability of the dyad, although we acknowledge the exploratory nature of this analysis. We ran a regression analysis with the interaction index as the dependent variable and similarity in performance and average performance as the predictors (Table 1; Model 1). We found extreme evidence for the model including the interaction term compared to the model including only the two main effects (BF>100; Model 2).

---------------------------- Please insert Table 1 here --------------------------

Model	Measures	B	95% CIs		Model comparison	ΔR^2	$BF_{m1/m2}$
1	Average of performance	0.03	[-0.03	0.08]	1vsNull	.56	>100
	Similarity in performance	0.125	[0.07	0.18]			
2	Average of performance	0.04	[0.0002	0.085]	2vs1	.18	>100
	Similarity in performance	0.15	[0.11	0.20]			
	Average x Similarity	-0.07	[-0.11	-0.04]			

Table 1. Regression analyses with the interaction index as the outcome variable. For each dyad, the interaction index was calculated as the ratio between the dyad's absolute deviation and the dyad's best estimator's (BE) absolute deviation. For both regression models: Multicollinearity was absent (i.e., all Variance Inflation Factors lower than 4); No outliers were identified (i.e., Bonferroni's tests on Studentized residuals were associated with *p*s>.05) and no influential observations were found (i.e., all Cook's distances were below or equal 1); Normal distribution of residuals was respected (i.e., Shapiro tests were associated with *p*s>.05).

From the visual inspection of the Figure 4, it emerged that for dyads with a sufficiently accurate mean performance in the to-be-shared condition, an interaction benefit emerged if the two dyad members displayed similar performance in the to-be-shared condition (for further exploration of this interaction see the supplementary materials).

---------------------------- Please insert Figure 4 here --------------------------

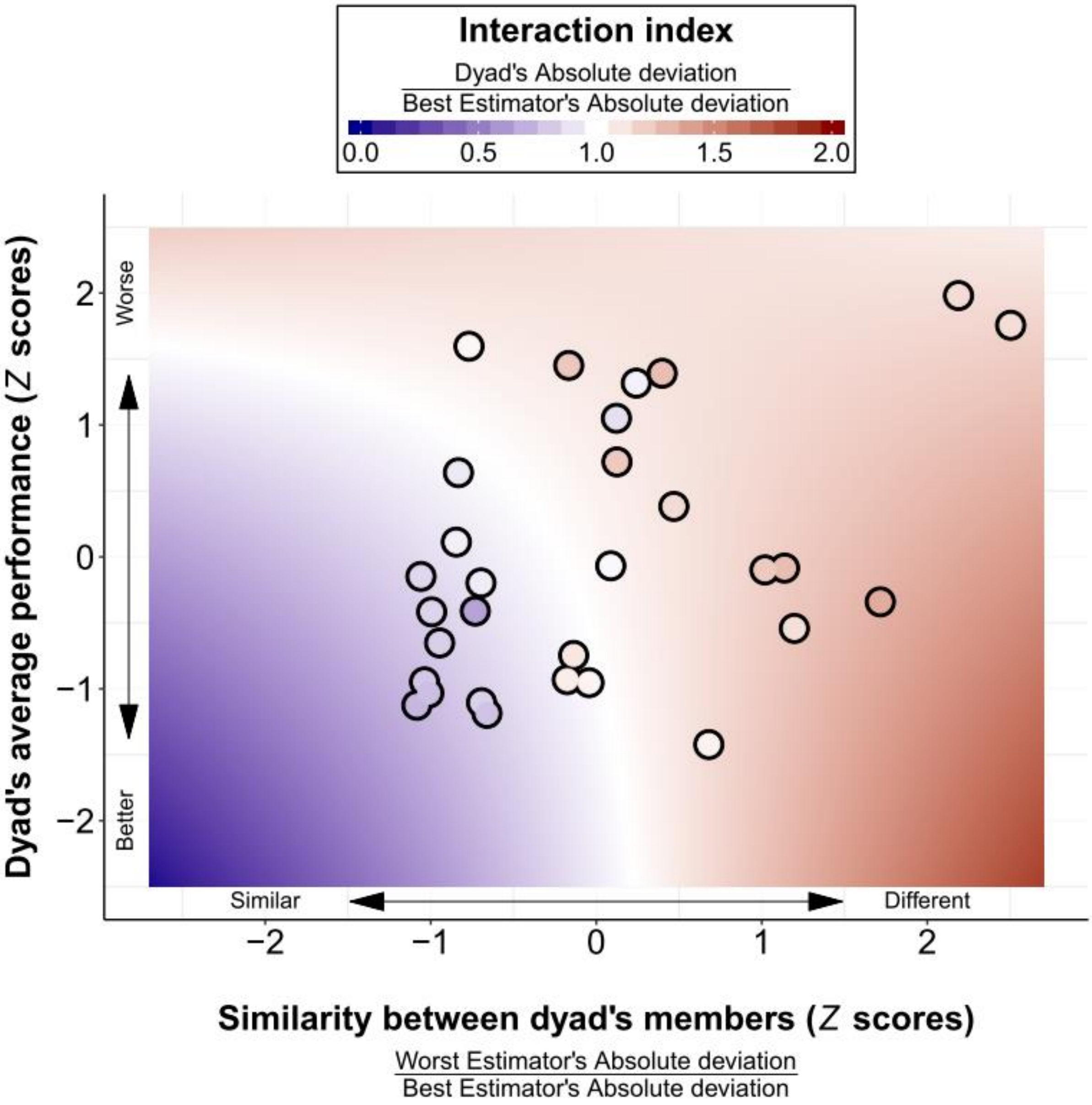

Figure 4. Predicted values (the hue outside the dots obtained with linear interpolation) from the regression analysis on interaction index as a function of the statistical interaction between similarity between the dyad's members and the dyad's average performance. The circles represent the actual dyad's scores for similarity and average performance (*z*-scored), with the colour inside the circles representing the dyad's actual interaction index. As the average performance of the dyad decreased, the relationship between similarity and the interaction index progressively disappeared.

Real and perceived competence

Both members rated their and their partner ability in mapping numbers on the competence scale. Therefore, for each participant, we could derive measures of real competence and perceived competence with respect to their partner. For real competence, we computed the logarithm[1] of the ratio between the partner's absolute deviation and the self-absolute deviation in the to-be-shared condition (i.e., ratio of real competence), whereby positive values represent a better performance of the participant compared to her partner and negative values represent a worse performance compared to her partner. For perceived competence, we computed the logarithm of the ratio between the self-perceived competence and the partner's perceived competence (i.e., ratio of perceived competence), whereby positive values represent a participant considering her more accurate than her partner and negative values represent a participant considering her less accurate than her partner. A ratio of perceived competence of zero means that a participant considered herself having the same competence of her partner. In Figure 5, we plotted the relation between the real and perceived competences. Points landing in the upper right quadrant and in the bottom left quadrant are correct identifications whereas points in the other quadrants are wrong identifications. Only 17% of the participants (8 novices and 2 experts) wrongly classified themselves as the novice or the expert in the dyad. Therefore, participants appeared to be aware about their level of competence in the task. It is interesting to notice that the distribution of ratios of perceived competence is narrow and to some extent skewed toward positive values. Individuals considered themselves to be marginally better or worse than their partners with a tendency to grant themselves a better ability in placing numbers on the line.

[1] We log transformed the ratios to overcome asymmetry. For example, a member of the dyad rated her competence to be 50 and her partner's to be 100, yielding a ratio of 1/2 (i.e., 50/100). The other member of the dyad rated her competence to be 80 and her partner's to be 40, yielding a ratio of 2 (i.e., 80/40). In this specific case, the perceived levels of competence are symmetrical: one participant considered her competence to be half compared to the partner and the other participant considered her competence to be double compared to the partner's. This symmetry is not properly assessed by ratios (i.e., comparing 0.5 and 2) but it is preserved with the log transformation, whereby $\ln(0.5)=-0.69$ and $\ln(2)=0.69$. The logged ratios have the same magnitude but different sign.

---------------------------- Please insert Figure 5 here --------------------------

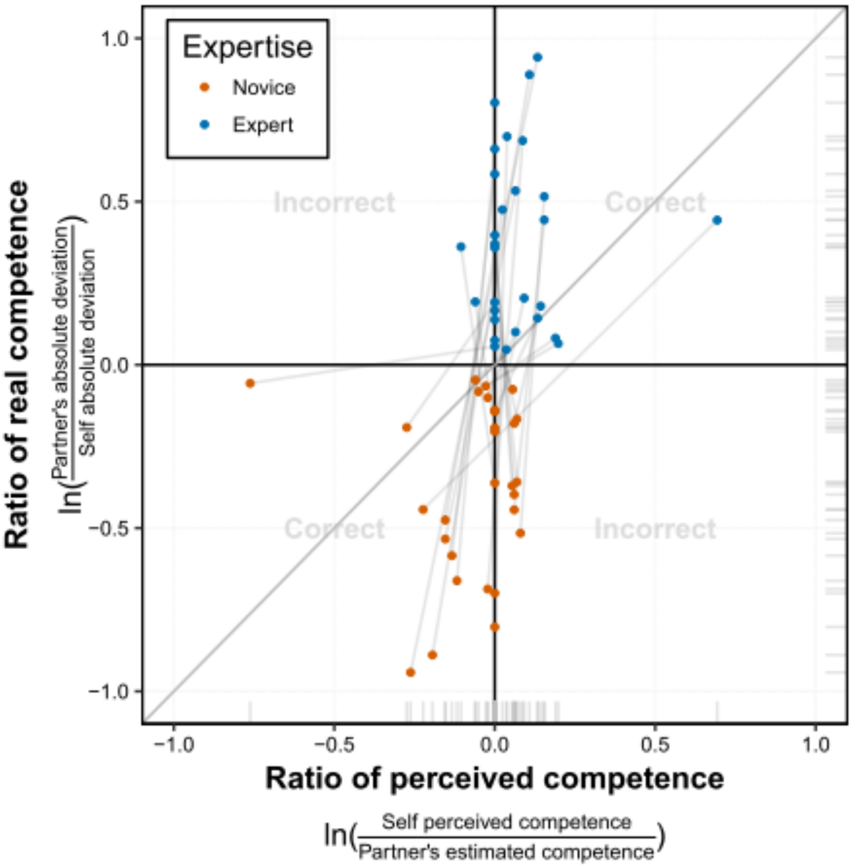

Figure 5. Relation between ratios of real and perceived competence for novices (red dots) and experts (blue dots). Grey lines connecting dots represent individuals from the same dyad. The grey bands at the bottom margin of the plot denote the distribution of ratio of perceived competence whereas grey bands on the right margin of the plot denote the distribution of ratio of real competence. The oblique grey line defines the perfect correspondence between real and perceived competence.

Furthermore, there seems to be a discrepancy between the real and perceived competence as represented by the points on the plot being far away from the oblique grey line, which represents a perfect correspondence between the real and perceived competence. We calculated for each participant the difference between the ratio of perceived competence and the ratio of real competence: a positive value indicates that a participant overestimated her actual ability compared to her partner whereas a negative value indicates an underestimation of competence. Interestingly, there was extreme evidence for a difference between novices and experts: Novices overestimated their competence whereas experts underestimated their competence (Novices: *M*=0.31, *SD*=0.31; Experts: *M*=-0.3, *SD*=0.3; BF_{10}>100). Overall, novices realised to be the novice in the dyad but they tended to overestimate their competence as they admitted to be less competent but only slightly worse than their partner. Experts, instead, realised to be more competent but were humble and considered their performance to be only slightly better than their novice partner.

---------------------------- Please insert Figure 6 here --------------------------

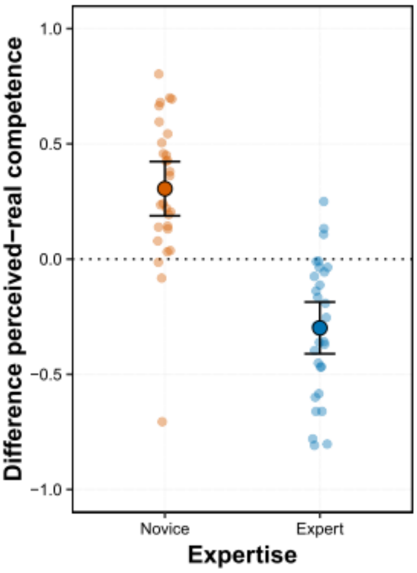

Figure 6. Difference between ratios of perceived and real competence. Positive values denote that participants overestimated their real ability compared to their partner whereas negative values denote an underestimation of competence.

Discussion

In the present study, young adults with an academic background in mathematics and the humanities were assigned to balanced dyads to complete the number line task privately, privately but knowing that they were about to share their estimates (i.e., the to-be-shared condition), jointly, and privately again. Afterwards, participants rated their and their partner's ability in mapping numbers on a numerical scale.

First, within each dyad, participants were categorized as an expert or novice based on their performance in the to-be-shared condition of the number line task. Not surprisingly, experts outperformed novices also in the pre-interaction condition. Both experts and novices mapped numbers more accurately in the to-be-shared condition compared to the pre-interaction condition. This improvement in the to-be-shared condition could be attributed to a Hawthorne-like effect (Landsberger, 1958) given that participants knew that their estimates would be shared with the other member of the dyad. When members of the dyad negotiated their shared response in the joint condition, the displayed joint accuracy was higher compared to the accuracy of the to-be-shared trials for novices but not for experts (Minson et al., 2011). Moreover, novices, but not experts, showed a mapping accuracy in the post-interaction condition that was similar to the pre-interaction one. Therefore, in the long run, novices obtained a minimal benefit from interacting with an expert in the current setting. Overall, the type of dyad (i.e., Maths-Maths, Humanities-Maths, Humanities-Humanities) did not influence performance across the different conditions of the number line task.

In the joint condition, each member of the dyad saw their respective estimates from the to-be-shared condition and was asked to move a third slider to agree upon a joint estimate. Ninety-eight percent of joint estimates fell within the interval that was created by novices' and experts' estimates in the to-be-shared condition, thereby indicating that in the joint condition, participants averaged their prior responses (Bang et al., 2017; Mahmoodi et al., 2015). According to the bracketing principle, averaging nearly all the time would have been the most effective strategy if the interval created by individual estimates had nearly always included the target number (Jacobson et al., 2011; Larrick & Soll, 2006; Minson et al., 2011). However, this was clearly not the case in the current experiment, with the correct location of the target number falling outside the dyad's estimates on 60% of occasions. Participants failed to conceive a scenario in which both their estimates were either under or over the target number position (i.e., an under- or over-estimation pattern). This result highlights the importance of advising interacting people that both novices and experts can be biased in the same direction; in which case, they should completely reconsider their opinions.

The two members of each dyad assigned an equal weight to their estimates (i.e., averaged their estimates) despite their different levels of competence, thereby showing a clear equality bias (Mahmoodi et al., 2015). Both experts and novices were responsible for the heavy reliance on averaging. In fact, neither experts nor novices attempted to pull their joint estimate towards their individual estimate in the to-be-shared condition; instead, both dyad members sought the average. The presence of a strong equality bias is in line with previous studies, which have shown that the assignment of equal weight to judgments emerges even when the difference in performance is large, when trial-by-trial feedback about the true answer is provided, when cumulative feedback about the performance of each observer is provided and when participants have monetary incentives to deviate from equal weighting (Bang et al., 2017; Mahmoodi et al., 2015). The strong reliance on averaging may be explained by the novice's desire to participate in the joint condition and the expert's implicit obligation to integrate the novice's contribution (Harvey & Fischer, 1997; Mahmoodi et al., 2015) or by the fact that participants failed to identify the expert in the dyad (Miner, 1984; Trotman et al., 1983), even though the latter does not seem to be the case in the present study. Therefore, in the case of peer interaction, as often happens in learning contexts, the expert should be clearly identified and rendered responsible for leading the novice toward a better accomplishment of the assigned task.

People typically discount others' opinion (egocentric discounting), a behaviour which, paradoxically, is strongest among people of low competence who really should take others' advice (i.e., Dunning-Kruger effect) (Dunning, 2011; Kruger & Dunning, 1999). One difference between paradigms investigating egocentric discounting and the current one is that they did not involve interaction, and, as such, there was no social obligation to take the other's opinion into account. Another difference is that studies on advice-taking typically focus on situations where it would be in participants' best interest to take the others' advice into account. The current study stands out in at least two ways. First, we study advice-taking in a more natural setting where social dynamics are more likely to come into play. Second, we study both sides of the interaction: the perspectives of the less competent individual who has something to gain and the more competent individual who has something to lose. These features of our study together with our findings allow us to refine the idea that people typically discount others' opinion. While the observed 'equal'-weighting strategy implies that the less competent dyad member did not give a sufficiently high weight to their more competent partner (classic egocentric discounting), it also implies that the more competent dyad member did not give a sufficiently low weight to their less competent partner (a phenomenon which perhaps should be called 'allocentric' discounting).

In our study, performance in the joint condition tended to equal performance of the best estimator in the dyad (Bahrami et al., 2012b); hence the beneficial effect of interaction was generally limited. Nevertheless, the interaction benefit was not uniform but better explained by the statistical interaction between the similarity in performance of the two dyad members and the average accuracy of dyad members in the to-be-shared condition. In line with previous studies (Bahrami et al., 2013), interaction led to a benefit when the two dyad members exhibited a similar level of accuracy in the to-be-shared condition. In an exploratory analysis, we found that this relationship was evident when the averaged performance of the two dyad members in the to-be-shared condition was above the mean and progressively disappeared with lower average performance. Interaction was beneficial when the dyad members had similar levels of expertise and when, on their own, they were sufficiently accurate. Therefore, an optimal arrangement of working dyads should take into consideration both their similarity and their level of ability. In fact, two individuals with the same but low level of expertise appeared to represent the least effective means of achieving an interaction benefit.

After performing the conditions of the number line task, participants estimated their and their partner’s ability in the number line task by completing a competence scale. Overall, participants displayed an accurate categorisation of themselves and their partner as the expert or the novice in the dyad. Nevertheless, novices overestimated their actual competence and considered themselves as only slightly worse than their partner. Experts, instead, were humble and underestimated their performance in comparison with their novice partner.

In summary, the present study demonstrated that, after interacting with an expert in a numerical estimation task, novices showed minimal benefit from the interaction. This result occurred at least when feedback and the explicit identification of the expert dyad member were absent, as was the case in the current study. Despite declaring different levels of expertise, interacting individuals assigned an equal weight to their estimates, thereby showing a clear equality bias (Mahmoodi et al., 2015). Finally, we demonstrated that for the interaction benefit to emerge, dyad members must have similar levels of expertise and sufficiently accurate average performance. Taken together, our findings provide important information for dyads interacting within learning and organisational contexts. We suggest that experts should lead the interaction and convince the novices to follow their experienced guidance. Moreover, the dyad should be assembled so that the two interacting individuals have similar levels of expertise and that their average performance is sufficiently accurate. In particular, we suggest that individuals with similar but low average performance should not be paired during interaction given the resultant lack of an interaction benefit.

References

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30): Springer Science & Business Media.

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2013). Collective enumeration. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 338.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012a). Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 3.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012b). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 367*(1594), 1350-1365.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*(5995), 1081-1085.

Bang, D., Aitchison, L., Moran, R., Castanon, S. H., Rafiee, B., Mahmoodi, A., . . . Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour, 1*(6), s41562-41017-40117.

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., . . . Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and cognition, 26*, 13-23.

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental science, 14*(1), 125-135.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision, 10*, 433-436.

Bruce, R. S. (1935). Group judgments in the fields of lifted weights and visual discrimination. *The Journal of Psychology, 1*(1), 117-121.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic bulletin & review, 18*(2), 331-338.

Cohen, D. J., & Sarnecka, B. W. (2014). Children's number-line estimation shows development of measurement skills (not number representations). *Developmental psychology, 50*(6), 1640.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in cognitive sciences, 7*(4), 145-147.

Dunning, D. (2011). 5 The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. *Advances in experimental social psychology, 44*, 247.

Eysenck, H. (1939). The validity of judgments as a function of the number of judges. *Journal of Experimental Psychology, 25*(6), 650.

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature, 75*, 450-451.

Gordon, K. (1924). Group Judgments in the Field of Lifted Weights. *Journal of Experimental Psychology, 7*(5), 398.

Gordon, K. (1935). Further observations on group judgments of lifted weights. *The Journal of Psychology, 1*(1), 105-115.

Grofman, B., Owen, G., & Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision, 15*(3), 261-278.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*(2), 117-133.

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes, 62*(2), 190-197.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychol Sci, 20*(2), 231-237.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 719.

Hurst, M., Leigh Monahan, K., Heller, E., & Cordes, S. (2014). 123s and ABCs: Developmental shifts in logarithmic-to-linear responding reflect fluency with sequence values. *Developmental science, 17*(6), 892-904.

Jacobson, J., Dobbs-Marsh, J., Liberman, V., & Minson, J. A. (2011). Predicting Civil Jury Verdicts: How Attorneys Use (and Misuse) a Second Opinion. *Journal of Empirical Legal Studies, 8*(s1), 99-119.

Jeffreys, H. (1961). *Theory of probability*: Oxford university press.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*(3), 216-247.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology, 77*(6), 1121.

Landsberger, H. A. (1958). Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science, 52*(1), 111-127.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology, 48*(2), 507-512.

Littlepage, G., Robison, W., & Reddington, K. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes, 69*(2), 133-147.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology, 47*(6), 1231.

Mahmoodi, A., Bang, D., Ahmadabadi, M. N., & Bahrami, B. (2013). Learning to make collective decisions: the impact of confidence escalation. *PLoS One, 8*(12), e81195.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., . . . Frith, C. D. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences, 112*(12), 3835-3840.

Miner, F. C. (1984). Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance, 33*(1), 112-124.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin*, 0146167211410436.

Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H.-C. (2009). Children's early mental number line: Logarithmic or decomposed linear? *Journal of experimental child psychology, 103*(4), 503-515.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision, 10*(4), 437-442.

Preston, M. G. (1938). Note on the reliability and the validity of the group judgment. *Journal of Experimental Psychology, 22*(5), 462.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sella, F., Sader, E., Lolliot, S., & Cohen Kadosh, R. (2016). Basic and advanced numerical performances relate to mathematical expertise but are fully mediated by visuospatial skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development, 75*(2), 428-444.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation evidence for multiple representations of numerical quantity. *Psychol Sci, 14*(3), 237-250.

Smith, M. (1931). Group judgments in the field of personality traits. *Journal of Experimental Psychology, 14*(5), 562.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780.

Sullivan, J. L., Juhasz, B. J., Slattery, T. J., & Barth, H. C. (2011). Adults' number-line estimation strategies: Evidence from eye movements. *Psychonomic bulletin & review, 18*(3), 557-563.

Thompson, J. M., Nuerk, H. C., Moeller, K., & Cohen Kadosh, R. (2013). The link between mental rotation ability and basic numerical representations. *Acta Psychol (Amst), 144*(2), 324-331. doi:10.1016/j.actpsy.2013.05.009

Trotman, K. T., Yetton, P. W., & Zimmer, I. R. (1983). Individual and group judgments of internal control systems. *Journal of Accounting Research*, 286-292.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychol Sci, 19*(7), 645-647.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Supplementary Materials

Comparison of spatial mapping: Experts and Novices.

We further explored behaviour in the number line task in order to identify differences in the mapping of numbers onto the line between experts and novices. In Figure S1, we reported the mean estimates as a function of target numbers separately for novices and experts across the task conditions. Both experts and novices displayed a clear linear mapping of target numbers along the number line.

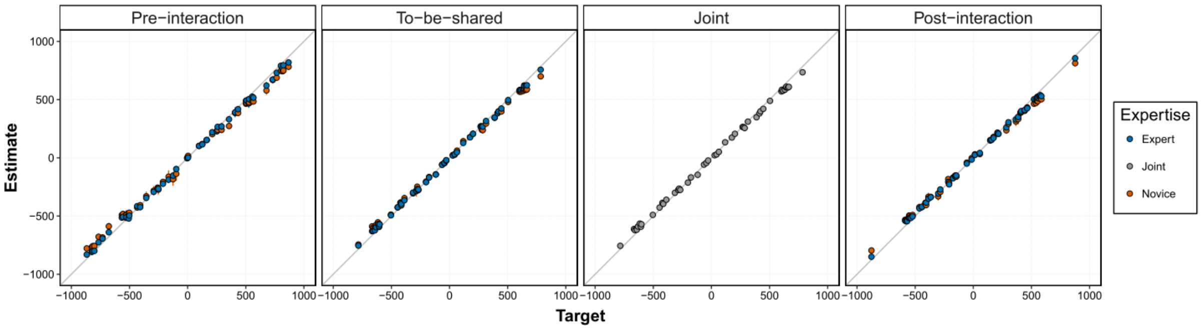

Figure S1. Mean estimates as function of target numbers in the number line task separately for experts (blue dots), novices (red dots) and joint trials (grey dots). Bars represent 95% CIs.

This pattern was confirmed by the individual R^2-values from a linear regression in which estimates were predicted from target numbers for each participant. Means of the individual R^2 approached one for both groups (Figure S2). There was anecdotal evidence for a different performance in the pre-interaction condition (Experts: *M*=0.989, *SD*=0.01; Novices: *M*=0.98, *SD*=0.024; BF_{10}=1.27), whereas experts displayed higher linearity compared to novices in the to-be-shared (Experts: *M*=0.993, *SD*=0.003; Novices: *M*=0.985, *SD*=0.013; BF_{10}=30; strong evidence) and the post-interaction conditions (Experts: *M*=0.992, *SD*=0.004; Novices: *M*=0.982, *SD*=0.016; BF_{10}=38; very strong evidence).

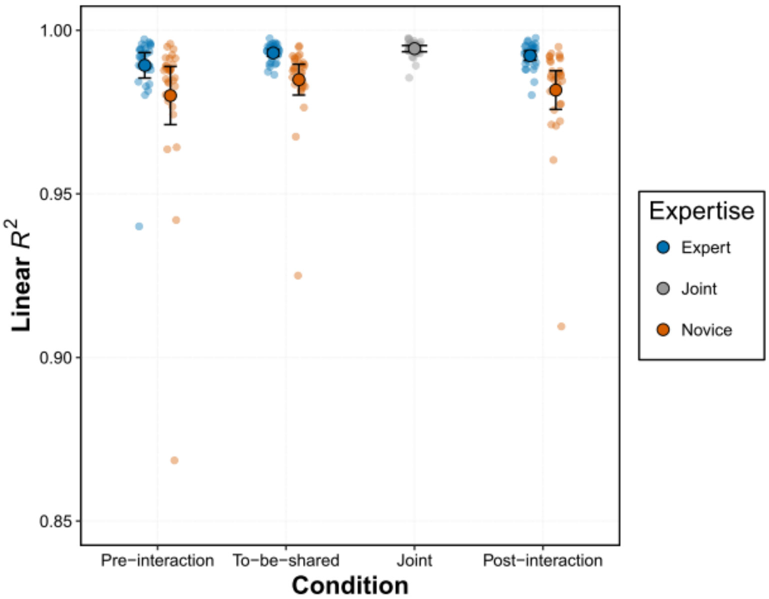

Figure S2. Mean of R^2 separately for experts (blue dots), novices (red dots) along with joint trials (grey dots) in the number line task. Transparent points represent individual values. Bars represent 95% CIs.

We also assessed the ordinality of the mapping by computing the Spearman rank correlation between estimates and the target numbers for each participant. Means of the individual Spearman rank correlation approached one for both groups (Figure S3). There was anecdotal evidence of a similar performance in the pre-interaction condition (Experts: *M*=0.993, *SD*=0.006; Novices: *M*=0.986, *SD*=0.019; BF_{10}=0.86), whereas experts displayed higher ordinality compared to novices in the to-be-shared (Experts: *M*=0.993, *SD*=0.003; Novices: *M*=0.987, *SD*=0.008; BF_{10}=47; very strong evidence). There was anecdotal support in favour of a higher ordinality for experts compared to novices in the post-interaction conditions (Experts: *M*=0.992, *SD*=0.004; Novices: *M*=0.984, *SD*=0.016; BF_{10}=2.97).

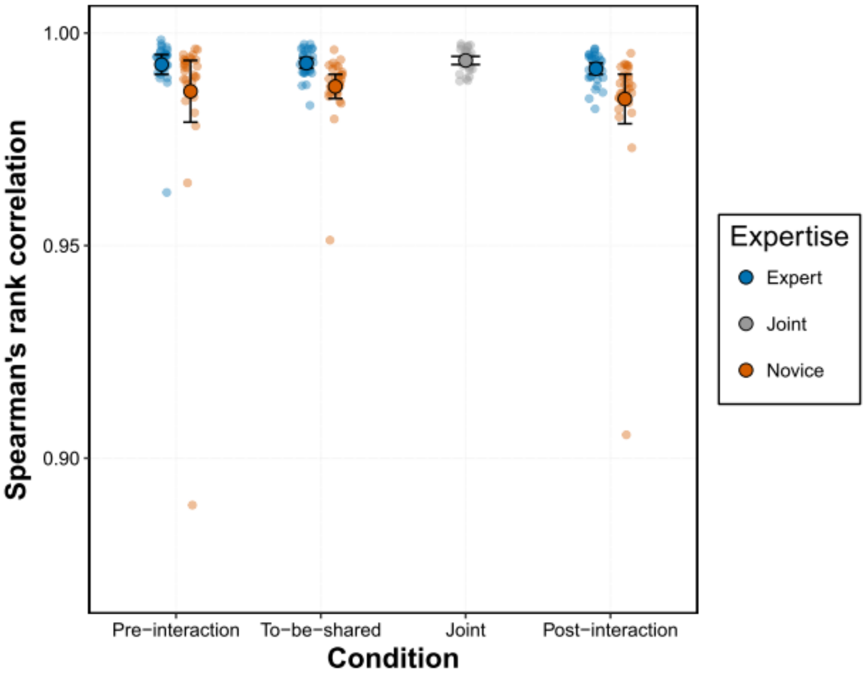

Figure S3. Mean of Spearman's rank correlation separately for experts (blue dots), novices (red dots) in the pre-interaction, to-be-shared and post-interaction condition along with joint trials (grey dots) from the number line task. Transparent points represent individual values. Bars represent and 95% CIs.

Finally, we reported the mean deviation of estimates for experts and novices across the conditions of the number line task (Figure S3). Both groups seem to show a slight pattern of underestimation in the pre-interaction condition whereas in the other conditions there was not a clear pattern of over- or under-estimation and there was moderate/anecdotal evidence for a similar performance in both groups (Pre-interaction: Experts, *M*=-6.87, *SD*=14.12, Novices, *M*=-8.996, *SD*=20.4, BF_{10}=0.29; To-be-shared: Experts, *M*=-0.97, *SD*=8.996, Novices, *M*=-4.14, *SD*=18.95, BF_{10}=0.35; Post-interaction: Experts, *M*=-1.57, *SD*=14.87, Novices, *M*=-5.13, *SD*=20.85, BF_{10}=0.33).

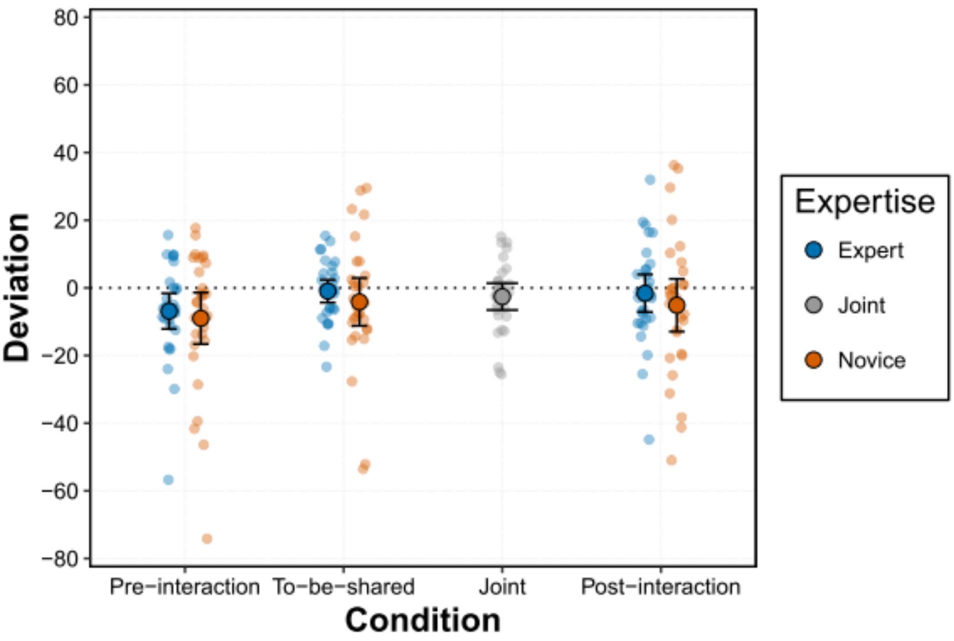

Figure S4. Mean deviations for experts and novices in the number line task. Transparent points represent individual values. Bars represent 95% CIs.

Analysis of absolute deviation in the number line task (NHST approach)

In all analyses of variance (ANOVAs), *p* values were adjusted using the Greenhouse-Geisser correction ($p_{[gg]}$) when Mauchly's test indicated a violation of the assumption of sphericity. In the following sets of multiple comparisons, *p* values were adjusted using the Bonferroni correction.

We analysed the mean absolute deviation in the number line task using a mixed ANOVA with Condition [pre-interaction, to-be-shared, joint, post-interaction] as a within-subjects factor and Expertise [Expert, Novice] and Dyad type [Maths-Maths, Maths-humanities, Humanities-Humanities] as the between-subjects factors. The main effect of Expertise, *F*(1, 54)=22.55, MSE=0.039, *p*<.001, η_p^2=.29, and the main effect of Condition, *F*(3, 162)=34.69, MSE=0.008, $p_{[gg]}$<.001, η_p^2=.39, were both significant. The Expertise x Condition interaction also reached significance *F*(3, 162)=12.42, MSE=0.008, $p_{[gg]}$<.001, η_p^2=.19. We explored the interaction by comparing Experts' and Novices' mean absolute deviation in the pre-interaction, to-be-shared, and post-interaction conditions as well as observing within each group whether there was any significant difference between pre-interaction and to-be-shared, to-be-shared and joint, joint and post-interaction, and pre-interaction and post-interaction conditions (11 comparisons with *p* values corrected using the Bonferroni formula).

The experts displayed less absolute deviation in the pre-interaction (Experts: *M*=1.65, *SD*=0.13; Novices: *M*=1.81, *SD*=0.12, *t*(58)=5.07, *p*<.001, *d*=1.31), to-be-shared (Experts: *M*=1.56, *SD*=0.09; Novices: *M*=1.72, *SD*=0.12, *t*(58)=5.76, *p*<.001, *d*=1.49), and post-interaction conditions (Experts: *M*=1.57, *SD*=0.12; Novices: *M*=1.74, *SD*=0.16, *t*(58)=4.69, *p*<.001, *d*=1.21) compared to novices. Both groups displayed a reduction in absolute error from the pre-interaction to the to-be-shared conditions (Experts: *t*(29)=3.98, *p*=.005, *d*=0.73; Novices: *t*(29)=4.33, *p*=.002, *d*=0.79). Only novices displayed a reduced absolute deviation in the joint condition compared to the to-be-shared condition (Experts: *t*(29)=0.43, *p*=1, *d*=0.08; Novices: *t*(29)=12.65, *p*<.001, *d*=2.31) and an increase in absolute deviation in the post-interaction condition compared to the joint condition (Experts: *t*(29)=0.74, *p*=1, *d*=0.14; Novices: *t*(29)=6.63, *p*<.001, *d*=1.21). Finally, only experts displayed a reduced absolute deviation in the post-interaction condition compared to the pre-interaction condition (Experts: *t*(29)=3.53, *p*=.016, *d*=0.64; Novices: *t*(29)=2.78, *p*=.103, *d*=0.51).

Analysis of interaction benefit (NHST approach).

Model	Measures	B	95% CIs		Model comparison	*ΔR*²
1	Average of performance	0.03	[-0.03	0.08]	1vsNull	.56
	Similarity in performance	0.125***	[0.07	0.18]		
2	Average of performance	0.04*	[0.0002	0.085]	2vs1	.18
	Similarity in performance	0.15***	[0.11	0.20]		
	Average x Similarity	-0.07***	[-0.11	-0.04]		

Table S1. Regression analyses with the interaction benefit as the outcome variable. **p*<.05, ***p*<.01, ****p*<.001.

Computational estimation task

The computational estimation task (Levine, 1982). Participants were asked to complete ten multiplication and division problems before the Number line task (76x89, 135x48, 64.6x0.16, 24.3x18.5, 0.47x0.26, 6375/17, 648.9/22.4, 764/44.5, 66/0.86, 0.73/0.94) and the remaining ten after the Number line task (63x92, 145x37, 72.5x0.12, 12.5x11.4, 0.37x0.43, 4645/18, 737.1/27.2, 546/33.5, 55/0.73, 0.76/0.89); these are hereon referred to as the first and second sessions. The trials (multiplications/divisions) were presented simultaneously for 1 minute and participants were asked to write down their best estimate for each arithmetic problem. Participants were explicitly told to estimate without calculating the exact answer. For each response, we calculated the proportion of its absolute deviation from the exact answer (the proportion of absolute deviation=|estimate-correct answer|/ correct answer). When estimates differed from the correct answer by more than the value of the correct answer, such responses were deemed outliers and so were replaced with '1', as were missing values. Therefore, '1' was the maximum error score. Outliers were replaced rather than removed in order to avoid unfairly disadvantaging participants who misunderstood the task (e.g., by failing to notice the change in arithmetic sign from multiplication to division). Missing values were also replaced with '1' rather than removed in order to avoid unfairly advantaging participants who decided not to complete the task, with the aim of spending more time on fewer calculations.

Analysis of absolute deviation in the Number Line task with the classification in experts and novices based on the performance in the Computational Estimation task (NHST and Bayesian approach).

We classified participants as experts and novices based on their performance in the first session of the computational estimation task. Hence, the more accurate performer in the computational estimation task within each dyad was labelled the expert (*n*=30) and the less accurate member was labelled the novice (*n*=30). The performance in the computational estimation task can be considered as an external and independent criterion to assess participants' expertise.

We compared experts' and novices' deviation in the pre-interaction, to-be-shared, and post-interaction conditions as well as observing within each group whether there were significant differences between pre-interaction and to-be-shared, to-be-shared and joint, joint and post-interaction, and pre-interaction and post-interaction conditions (Figure S5). In the joint condition, experts and novices evaluated their estimates from the to-be-shared trial and then negotiated a joint estimate. Consequently, both experts and novices obtained the same absolute deviations in the joint condition. Nevertheless, for transparency we report here the mixed ANOVA and later the planned *t*-test comparisons using the Bonferroni correction. We analysed the deviation in the number line task using a mixed ANOVA with Condition [pre-interaction, to-be-shared, joint, post-interaction] as a within-subjects factor and Expertise [Expert, Novice] and Dyad type [Maths-Maths, Maths-Humanities, Humanities-Humanities] as the between-subjects factors. The main effect of Expertise, $F(1, 54)=11.01$, MSE=0.046, $p=.002$, $\eta_p^2=.17$, and the main effect of Condition, $F(3, 162)=32.3$, MSE=0.008, $p_{[gg]}<.001$, $\eta_p^2=.37$, were both significant. The Expertise x Condition interaction also reached significance, $F(3, 162)=7.25$, MSE=0.008, $p_{[gg]}<.001$, $\eta_p^2=.12$.

We additionally ran a Bayesian mixed ANOVA with Condition [pre-interaction, to-be-shared, joint, post-interaction] as a within-subjects factor and Expertise [Expert, Novice] and Dyad type [Maths-Maths, Maths-Humanities, Humanities-Humanities] as the between-subjects factors. The inclusion of Dyad type into the model led to an inclusion Bayes factor of 0.097, therefore the Dyad type was removed from the model. There was extreme evidence for the model including the interaction Condition x Expertise compared to the model with the two main effects (BF>100, extreme evidence).

T-test comparisons using the Bonferroni correction revealed that experts displayed less deviation in the pre-interaction (Experts: *M*=1.68, *SD*=0.16; Novices: *M*=1.78, *SD*=0.12, *t*(58)=2.88, *p*=.061, *d*=0.74, approached significance; BF_{10}=7.57, moderate evidence), to-be-shared (Experts: *M*=1.58, *SD*=0.12; Novices: *M*=1.7, *SD*=0.13, *t*(58)=3.69, *p*=.005, *d*=0.95; BF_{10}=56, very strong evidence) and post-interaction conditions (Experts: *M*=1.58, *SD*=0.12; Novices: *M*=1.73, *SD*=0.16, *t*(58)=4.05, *p*=.002, *d*=1.05; BF_{10}>100, extreme evidence) compared to novices. Both groups displayed a reduction in such deviation from the pre-interaction to the to-be-shared conditions (Experts: *t*(29)=4.22, *p*=.002, *d*=0.77, BF_{10}=129, extreme evidence; Novices: *t*(29)=4.09, *p*=.003, *d*=0.75, BF_{10}=93, very strong evidence). Only novices displayed a reduced deviation in the joint condition compared to the to-be-shared condition (Experts: *t*(29)=1.70, *p*=1, *d*=0.31, BF_{10}=0.70, anecdotal evidence; Novices: *t*(29)=8.02, *p*<.001, *d*=1.46, BF_{10}>100, extreme evidence) and an increase in deviation in the post-interaction condition compared to the joint condition (Experts: *t*(29)=1.13, *p*=1, *d*=0.21, BF_{10}=0.35, anecdotal evidence; Novices: *t*(29)=5.95, *p*<.001, *d*=1.09, BF_{10}>100, extreme evidence). Finally, only experts displayed a reduced deviation in the post-interaction condition compared to the pre-interaction condition (Experts: *t*(29)=4.39, *p*=.002, *d*=0.80, BF_{10}>100, extreme evidence; Novices: *t*(29)=2.08, *p*=.514, *d*=0.38, BF_{10}=1.26, anecdotal evidence).

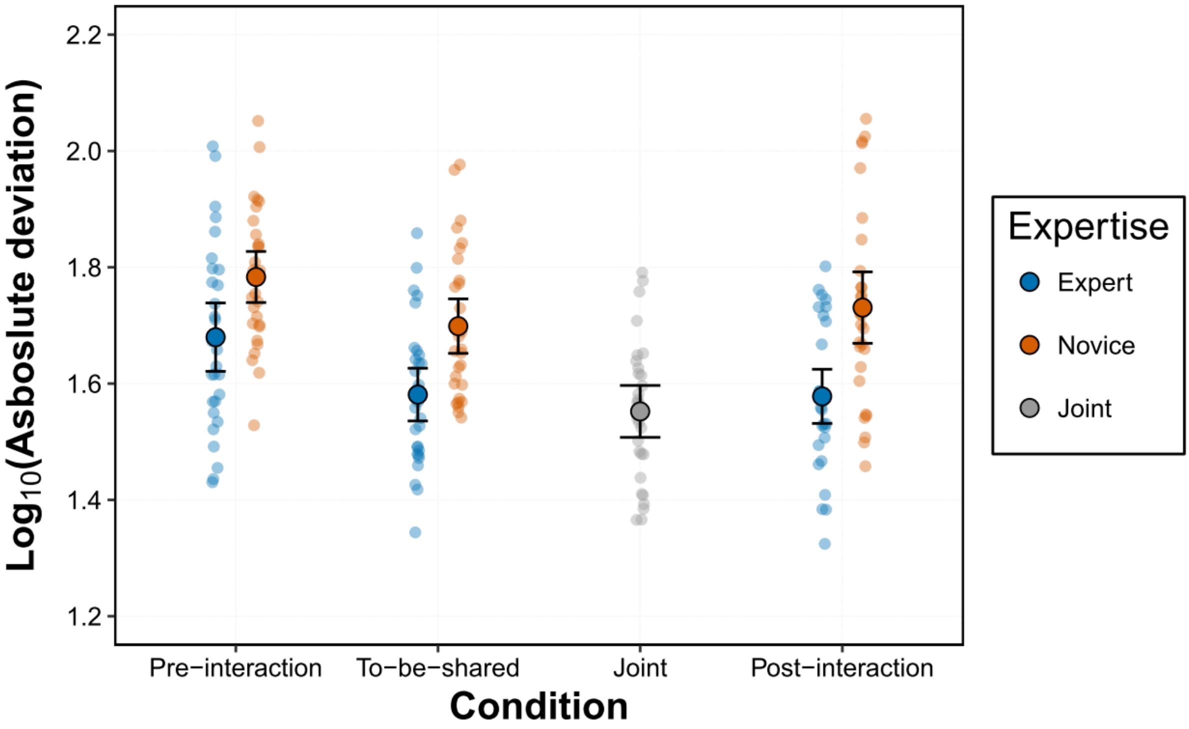

Figure S5. Mean absolute deviation (log$_{10}$ transformed) in the number line task for each condition separately for experts (blue dots) and novices (red dots; error bars represent 95%CIs) based on the performance of the Computational Estimation task.