Judgment and Decision Making (2023), **Vol. 18**:e2 1–22
doi:10.1017/jdm.2023.3

EMPIRICAL ARTICLE

Expertise determines frequency and accuracy of contributions in sequential collaboration

Maren Mayer[1,2*], Marcel Broß[3], and Daniel W. Heck[3]

[1]Leibniz-Institut für Wissensmedien (Knowledge Media Research Center), Tübingen, Germany; [2]Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany and [3]Department of Psychology, University of Marburg, Marburg, Germany

***Corresponding author.** E-mail: maren.mayer@iwm-tuebingen.de

Received: 30 January 2023; **Accepted:** 30 January 2023

Keywords: wisdom of crowds; group decision-making; mass collaboration; team work

Abstract
Many collaborative online projects such as Wikipedia and OpenStreetMap organize collaboration among their contributors sequentially. In sequential collaboration, one contributor creates an entry which is then consecutively encountered by other contributors who decide whether to adjust or maintain the presented entry. For numeric and geographical judgments, sequential collaboration yields improved judgments over the course of a sequential chain and results in accurate final estimates. We hypothesize that these benefits emerge since contributors adjust entries according to their expertise, implying that judgments of experts have a larger impact compared with those of novices. In three preregistered studies, we measured and manipulated expertise to investigate whether expertise leads to higher change probabilities and larger improvements in judgment accuracy. Moreover, we tested whether expertise results in an increase in accuracy over the course of a sequential chain. As expected, experts adjusted entries more frequently, made larger improvements, and contributed more to the final estimates of sequential chains. Overall, our findings suggest that the high accuracy of sequential collaboration is due to an implicit weighting of judgments by expertise.

1. Introduction

Online collaborative projects such as Wikipedia and OpenStreetMap have become increasingly important sources of information over the last two decades and are frequently used by many people. Prior research showed that Wikipedia yields highly accurate information both in general (Giles, 2005) and for specific topics (Kräenbring et al., 2014; Leithner et al., 2010). Also, OpenStreetMap provides geographic information with a similar accuracy as commercial map services and governmental data (Ciepłuch et al., 2010; Haklay, 2010; Zhang and Malczewski, 2018; Zielstra and Zipf, 2010). To gather information, both Wikipedia and OpenStreetMap build on a sequential process referred to as sequential collaboration (Mayer and Heck, 2022). In this process, one contributor creates an entry which is then sequentially adjusted or maintained by the following contributors.

Mayer and Heck (2022) showed that sequential collaboration represents a successful way of eliciting group judgments. In three online studies, participants either answered general-knowledge questions or located European cities on geographic maps. Participants were randomly assigned to sequential chains of four or six contributors. Each chain started with a contributor providing an independent judgment. Next, other contributors encountered the latest version of the judgment and could then decide whether to adjust or maintain it. For instance, the first individual may start by locating Rome on a map of Italy

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

without additional information. The second contributor may then maintain the location, whereas the third contributor may move the location more to the south. Participants were unaware of their position in the sequential chain, the change history of judgments, and how often a judgment had already been adjusted.

In the three studies by Mayer and Heck (2022), change probability and change magnitude decreased over the course of a sequential chain, whereas judgment accuracy improved. Observing an incremental increase in judgment accuracy over the course of a sequential chain represents a rather weak benchmark of performance since it lacks a comparison standard for the accuracy of the final judgments. As a remedy, one can also compare the accuracy of sequential collaboration against a stronger benchmark. In fact, the studies by Mayer and Heck (2022) provided preliminary evidence that the final judgments of sequential chains were similarly accurate, and in some cases even more accurate, than unweighted averaging, that is, computing the mean of independent individual judgments for the same number of participants. Similar results were reported by Miller and Steyvers (2011) for a more complex ordering task. This is an important finding given that unweighted averaging is known to yield highly accurate estimates in various contexts and tasks, a phenomenon known as wisdom of crowds (Hueffer et al., 2013; Larrick and Soll, 2006; Steyvers et al., 2009; Surowiecki, 2004).

Even though these initial results are promising, the mechanisms contributing to the increase in accuracy of sequential judgments are still unclear. In the present paper, we investigate whether the expertise of contributors affects both the probability of adjusting a presented judgment and the accuracy of revised judgments. We hypothesize that individuals with higher expertise are better at distinguishing between presented judgments they can improve and those they cannot improve. This would lead to a systematic opt-out mechanism: experts provide new, more accurate judgments when possible, but otherwise maintain a presented judgment (Mayer and Heck, 2022). Sequential collaboration would thus facilitate an implicit weighting of judgments by expertise, in turn leading to increasingly accurate judgments over the course of a sequential chain.

In the following, we first define expertise and discuss its relevance for judgment accuracy in various contexts. Based on the literature on the role of expertise for individual judgments, we propose a theoretical framework of how individuals' expertise drives a differential opt-out mechanism that improves the accuracy of sequential judgments. We conducted three experimental studies using a city-location task and a random-dots estimation task. In each study, we either measured individuals' knowledge or manipulated their skill for the task at hand. Thereby, we examined whether expertise influences how frequently presented judgments in sequential collaboration are adjusted and how much they are improved. As expected, we found that contributors with higher expertise adjust presented judgments more frequently and also provide larger improvements if adjustments are made. Furthermore, individuals with higher expertise have a larger impact on sequential chains than individuals with lower expertise, and this effect is more pronounced the later experts enter into the chain.

1.1. Expertise in judgment and decision-making

Expertise is a concept for which many definitions have been proposed (Herling, 2000), with a focus on different facets depending on the research question of interest (Baumann and Bonner, 2013). Typically, expertise is described as domain-specific (Herling, 2000) and comprises task-related knowledge, skills, and abilities (Schulze and Krumm, 2017; Stevens and Campion, 1994). This conceptualization is found in many research areas in judgment and decision-making (e.g., Kruger and Dunning, 1999; Martire et al., 2018; Schunn and Anderson, 1999). In our studies on sequential collaboration, we adopt a knowledge definition of expertise in Experiment 1 and skill definition of expertise in Experiment 2. By ensuring that our conceptualization of expertise is highly task-related, it is plausible to assume a positive effect of expertise on task performance (Kruger and Dunning, 1999).

Compared with novices, experts work on tasks in qualitatively different ways (Dubrovsky et al., 1991; Franz and Larson, 2002; Schunn and Anderson, 1999) and usually show better performance (Budescu and Chen, 2014; Kruger and Dunning, 1999; Merkle et al., 2020; Wang et al., 2021). In group

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

decision-making, the more individuals are aware of the expertise of other group members, the more accurate group decisions become (Baumann and Bonner, 2013). However, in such settings, it is crucial to explicitly communicate the expert status of group members before the discussion starts (Bonner et al., 2002). Moreover, when eliciting independent judgments from a group of individuals, weighting these judgments by expertise improves the accuracy of the aggregated estimates (Budescu and Chen, 2014; Lin and Cheng, 2009). For this purpose, expertise can be estimated statistically based on the observed performance (Mayer and Heck, 2023; Merkle et al., 2020; Merkle and Steyvers, 2011) or measured empirically by asking participants to rate their own, task-relevant knowledge (Ungar et al., 2012).

However, tasks need to have a certain level of 'demonstrability' for expertise to have an impact on a group decisions (Bonner et al., 2022; Laughlin and Ellis, 1986). For a task to be demonstrable, team members completing the task need to rely on the same system of communication and require sufficient information to solve the task. Moreover, team members who cannot solve the task still need to recognize and accept a correct solution if it is proposed by others, whereas members who can solve the task need to have sufficient motivation, ability, and time to demonstrate the accuracy of their solution to others. Highly demonstrable tasks ('intellective tasks') can profit from group members' task-related expertise. In contrast, less demonstrable, highly subjective tasks ('judgmental tasks') may not profit from expertise in a similar way since forming, communicating, and recognizing a correct answer is less clear (Bonner et al., 2022). According to this theoretical framework, sequential collaboration requires a sufficient level of task demonstrability, and thus we focus on intellective tasks in the following.

1.2. Implicit weighting of judgments by expertise

We hypothesize that sequential collaboration provides accurate outcomes because the process facilitates an implicit weighting of judgments by expertise. A weighting of judgments emerges due to the opportunity for contributors to opt out of providing a judgment. By opting out and maintaining the presented judgment, contributors assign more weight to the presented judgment (Mayer and Heck, 2022). In contrast, when opting in and adjusting a presented judgment, contributors give more weight to their own judgment compared with the presented judgment. Thus, if contributors show a differential opt-in and opt-out behavior depending on their expertise, judgments are implicitly weighted. This should, in turn, lead to increasingly accurate judgments over the course of a sequential chain since weighting by expertise improves aggregation of individual judgments (e.g., Budescu and Chen, 2014; Mayer and Heck, 2023; Merkle et al., 2020).

Such a process requires contributors to rely on task-related, metacognitive knowledge about their expertise. Metacognition describes contributors' 'cognition about cognitive phenomena' (Flavell, 1979). In the context of sequential collaboration, metacognitive knowledge (Lai, 2011) about one's own expertise allows contributors to evaluate the accuracy of presented judgments and one's own capacity to provide improvements (Kruger and Dunning, 1999; Laughlin and Ellis, 1986). Given that contributors decide whether to opt out of providing a judgment based on their metacognitive knowledge of their expertise (Bennett et al., 2018), sequential collaboration does not require a particular mechanism for identifying experts. It is thus neither necessary to assign expert roles (Baumann and Bonner, 2013), to directly assess individuals' expertise (Ungar et al., 2012), or to estimate expertise statistically (Mayer and Heck, 2023; Merkle et al., 2020). Instead, contributors determine the weighting of judgments within sequential chains implicitly based on their metacognitive assessment of their expertise and the evaluation of the presented judgment. Achieving high accuracy only requires that some of the contributors have sufficient expertise to detect and improve inaccurate judgments by others.

1.2.1. Probability of making adjustments

Sequential collaboration requires a two-stage response process in which contributors first decide whether to adjust a presented judgment (opt in) or whether to maintain it (opt out); only in the former case, they provide a new, revised judgment (Mayer and Heck, 2022). In the following, we derive predictions for the first stage in terms of the probability of adjusting a presented judgment.

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

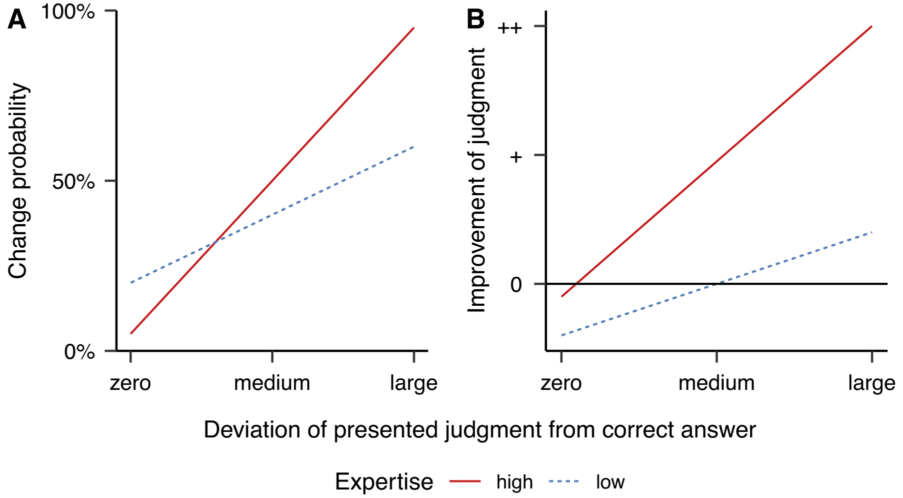

Figure 1. *Expected patterns for change probability and improvement.*

Note: In Panel B, a positive (negative) value of improvement indicates that a revised judgment is more (less) accurate than the presented judgment.

As discussed above, we expect that contributors with high expertise are better at distinguishing between presented judgments they can improve and those they cannot improve (Bennett et al., 2018). Figure 1A illustrates our expectations of how the following two factors influence contributors' decision whether to adjust or maintain a presented judgment: first, contributors' expertise, and second, the deviation of the presented judgment from the correct answer (referred to as presented deviation in the following). Contributors high in expertise should be able to detect highly accurate judgments as being correct, and in turn refrain from adjusting such judgments. Moreover, they should be able to detect even small deviations of the presented judgment from the correct judgment and show a high probability of adjusting presented judgments that differ considerably from the correct answer. In contrast, we assume that contributors with less expertise cannot reliably distinguish between presented judgments they can improve and those they cannot improve. Such contributors should show a substantial probability of (unnecessarily) adjusting already accurate judgments but a lower probability of adjusting inaccurate judgments compared with experts. Figure 1A shows that the expected pattern implies an interaction, that is, a steeper slope of the change probability (as a function of the presented deviation) for contributors with higher than for those with lower expertise.

As long as contributors have sufficient task-relevant expertise, a positive relationship between presented deviation and change probability should emerge (i.e., a main effect), meaning that larger presented deviations are more likely to be detected and adjusted. Since we assume that contributors with higher expertise can detect even small presented deviations and tend to adjust such judgments, we also expect an overall higher change probability for contributors with higher than for those with lower expertise. However, since we expect a crossed interaction (Figure 1), this effect will only emerge when the range of presented deviations across items is sufficiently large.

1.2.2. Improvement of presented judgments

When deciding to adjust a presented judgment, contributors in sequential collaboration have to provide a new, revised judgment in the second stage of the response process. We assume that, similar to change probability, the amount of improvement of presented judgments also depends on the two factors expertise and presented deviation. As illustrated in Figure 1B, we expect a main effect of presented deviation, meaning that with increasing deviation, contributors improve presented judgments to a larger degree. We also expect that improvements increase with increasing expertise since contributors high in expertise can provide more accurate judgments compared with those low in expertise (Merkle et al.,

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

2020; Ungar et al., 2012). As shown in Figure 1, we also assume an interaction of contributors' expertise and presented deviation. Individuals high in expertise should make no or only minor adjustments to presented judgments that are already accurate while providing medium improvements to moderately inaccurate judgments and large improvements to highly inaccurate judgments. In contrast, contributors with lower expertise may not be able to make similarly large improvements to highly and moderately inaccurate presented judgments. In fact, contributors low in expertise may even revise presented judgments that are already accurate, leading to negative improvements if the presented deviation is zero.

1.2.3. The role of expertise in chains of sequential judgments

The predictions above focus on a single step of sequential collaboration and apply to the level of individual contributions. In the following, we derive additional predictions for actual chains of sequential judgments of groups of at least two contributors. First, we consider whether contributors respond differently when encountering presented judgments of contributors who are high or low in expertise. As discussed above, contributors high in expertise should generally be better at distinguishing between accurate and inaccurate judgments. Assuming that the expertise of contributors is linked to the accuracy of their judgments (e.g., Merkle et al., 2020), we thus expect that contributors high in expertise adjust presented judgments more frequently if judgments were made by others low in expertise than by those high in expertise. In contrast, we expect that contributors low in expertise show a similar change probability irrespective of whether presented judgments were made by contributors with higher or lower expertise.

Concerning the accuracy of revised judgments, presented judgments should be improved most by contributors with high expertise who encounter judgments of contributors with low expertise. We expect smaller improvements if contributors revise presented judgments of others who have a similar level of expertise, that is, when contributors with high (low) expertise correct others with high (low) expertise. However, contributors with low expertise are expected to worsen presented judgments provided by others with high expertise. Our last predictions concern the overall accuracy of chains of sequential judgments. We expect that final estimates become more accurate the more contributors with high expertise enter a sequential chain. Moreover, accuracy should be higher if individuals with high expertise contribute later in the chain, since it becomes less likely that their judgments are changed (and possibly worsened) by other, less-skilled contributors.

To test our predictions, we conducted three experiments. Experiments 1 and 2 focus on a single sequential step of sequential collaboration (i.e., at the level of individual contributions) which allows us to experimentally control the deviation of presented judgments. In contrast, Experiment 3 studies the effects of expertise and presented deviation in actual chains of sequential judgments to examine the role of contributors with varying expertise entering the sequential chain at different points.

2. Experiment 1

2.1. Methods

In Experiment 1, we measured expertise in a city-location task and manipulated the presented deviation of judgments before letting participants decide whether to adjust or maintain location judgments with varying distances to the correct answer. To this end, we draw on the paradigm established by Mayer and Heck (2022) to investigate sequential collaboration. In the original study, participants positioned 57 European cities on maps. We modified the paradigm with some of these items serving as a baseline measure of individual expertise. Thereby, expertise was operationalized as knowledge acquired in the past (Schunn and Anderson, 1999). The remaining items were used to examine how participants adjust judgments in terms of change probability and improvement. The study design, sample size, hypotheses, and planned analyses were preregistered at https://aspredicted.org/cj9uu.pdf. Materials, analysis scripts, and data are available at https://osf.io/z2cxv/.

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

2.1.1. Participants

We recruited 290 participants who were compensated with 0.75€ for a median study duration of 9.63 minutes via a German panel provider. We excluded one participant who provided judgments that were on average more accurate than the mean accuracy of judgments found in a small test sample in which we instructed participants to look up the correct locations of each city before providing a judgment. Furthermore, we excluded eight participants who positioned more than 10% of the cities outside the highlighted areas which marked the countries of interest. After these exclusions, the final sample comprised 281 participants who were on average 46.49 years old (*SD* = 15.33) with 48.80% of participants being female. Concerning educational background, 15.70% had a college degree, 15% held a high school diploma, 31.10% had vocational education, and 38.20% had a lesser educational attainment.

2.1.2. Materials and procedure

Participants had to locate 57 European cities on seven different European maps, namely (1) Austria and Switzerland, (2) France, (3) Italy, (4) Spain and Portugal, (5) United Kingdom and Ireland, (6) Germany, and (7) Poland, Czech Republic, Hungary, and Slovakia. All maps had a resolution of 800 × 500 pixels and were scaled to 1:5,000,000. Table A1 in the Appendix provides a list of all cities and the phase they were presented in. The 17 cities used for measuring expertise were selected based on the accuracy of independent location judgments for these cities collected in Experiment 3 of Mayer and Heck (2022). Cities were selected to have a wide range of difficulty while ensuring that all seven European maps were represented.

In the first phase of Experiment 1, participants provided independent location judgments for the 17 cities which served as a measure of expertise. Each trial showed the instruction to place one city on the map as accurately as possible. Next, in the sequential phase, participants were instructed that each map already showed a location judgment of a previous participant and that they could decide whether to adjust or maintain the position. Again, only one of the remaining 40 cities was presented in each trial, but the map already contained a preselected location judgment. By clicking on the map, participants could adjust the presented judgment by providing a new position for the city, whereas clicking the button 'continue' allowed them to maintain the presented judgment. Participants were not provided with any additional information about the source of the judgment or the expertise of the previous contributor. Figure 2 displays the map of Italy with four preselected location judgments for Rome reflecting different distances from the correct location. All presented judgments were placed in the country or countries of interest colored in white. The seven maps and the corresponding cities were presented in block-randomized order (i.e., both the order of maps and of cities within maps were randomized).

Participants also provided demographic information and indicated their subjective knowledge concerning the locations of large European cities. Finally, they were debriefed and thanked for their participation.

Unknown to the participants, the locations presented in the sequential phase were not provided by other participants. Instead, we manipulated the presented deviation by selecting locations with a certain Euclidean distance to the correct answer (0, 40, 80, or 120 pixels). The presented deviations were selected based on judgments obtained in Mayer and Heck (2022) and pretested in a pilot study ensuring that participants were on average able to improve the presented distances. Moreover, in the Supplementary Materials (https://osf.io/z2cxv/), we show that the presented deviations correspond to plausible values from the empirical distribution of independent judgments of participants (which were collected for measuring expertise). The plots for all cities show that correct judgments as well as distances of 40, 80, or 120 pixels were inside the range of provided answers. For all 40 cities, one deviation was randomly selected such that each deviation was presented 10 times. For each map, the four levels of presented deviations were duplicated as rarely as possible. To manipulate presented deviation, it was necessary to deceive participants about the presented locations allegedly being judgments of other participants. The study was reviewed and approved by the ethics committee of the University of Mannheim, and participants were debriefed after participation.

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

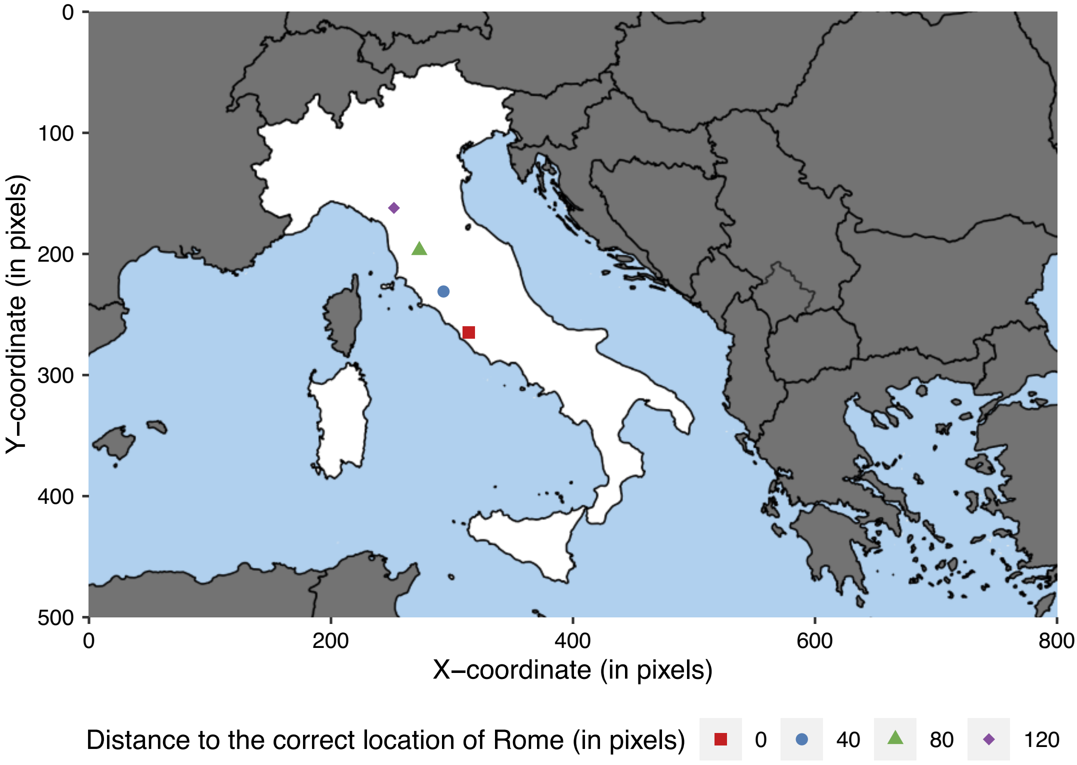

Figure 2. *Expected patterns for change probability and improvement.*

Note: Each participant only saw one of the four preselected location judgments.

To ensure that participants complied to the instructions and completed the study without technical issues, the online study was accessible only for participants using a computer (but not for mobile devices). We prevented looking up correct answers by implementing a time limit of 40 seconds for each response. Moreover, we already excluded participants during participation if they left the browser tab more than five times despite repeated warnings.

2.2. *Results and discussion*

We estimated participants' expertise based on the independent location judgments for the 17 cities that were shown without a previous judgment. As an operationalization of expertise, we computed the mean of the Euclidean distances between the location judgments and the correct positions for each participant. To ensure that larger values indicate higher expertise, we multiplied the average distances by -1. This measure of expertise was included as a continuous predictor in the analyses below. To assess the validity of this task-related expertise measure, we computed the correlation with the self-reported knowledge about the location of European cities. The large, positive correlation of $r = 0.43$ ($t(279) = 7.91, p < .001$) indicates a satisfactory convergent validity.

We tested the effects of participants' expertise, presented deviation, and their interaction on change probability using a generalized linear mixed model. The model used a logistic link function to predict the decision whether to adjust (= 1) or maintain (= 0) a presented judgment. We standardized our expertise measure for all analyses. Moreover, we applied a mean-centered linear contrast with values $-1.5, -0.5, 0.5$, and 1.5 for the four levels of deviations between presented and correct locations. Standardizing the expertise measure and applying a mean-centered contrast to the presented deviations allows us to interpret the additive terms in the model as main effects and the multiplicative term as

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

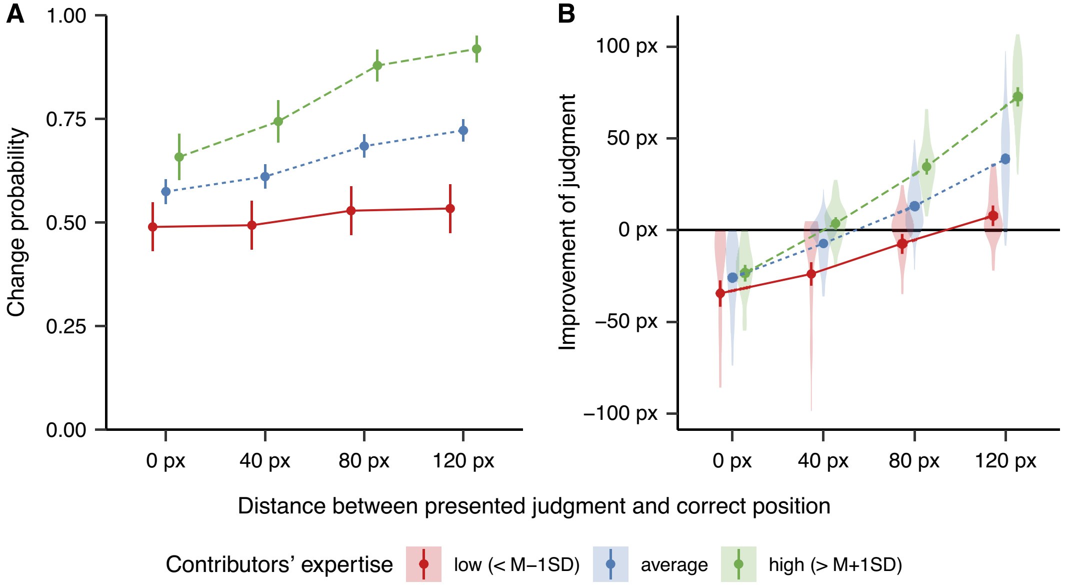

Figure 3. *Change probability and improvement of presented judgments in Experiment 1.*

Note: Points display empirical means with error bars showing the corresponding 99% between-subjects confidence intervals. Violin plots indicate the distribution of the dependent variable aggregated across items within each person. The plot for improvement only includes presented judgments that were adjusted by participants.

interaction. The model accounts for the nested data structure by including random intercepts for items and participants (Pinheiro and Bates, 2000).[1]

Figure 3A displays the average change probability for cities depending on participants' expertise and presented deviation. Table 1 shows the estimated regression coefficients. As expected, the linear contrast of presented deviation was positive and significant (β = 0.444, CI = [0.392, 0.495]). The model also indicated a significant positive relationship between expertise and change probability (β = 0.622, CI = [0.202, 1.042]). Furthermore, we found a significant interaction between expertise and presented deviation ($\beta = 0.218$, $CI = [0.165, 0.272]$).

However, contrary to our predictions, Figure 3A shows that individuals with higher expertise changed correct judgments more frequently than individuals with lower expertise. The high change probability for accurate presented judgments may be due to demand effects. In fact, participants did not know that 25% of the presented judgments had a perfect accuracy. Hence, they may not have expected that optimal behavior required maintaining a substantial proportion of the presented judgments.

Next, we examined the effects of presented deviation, expertise, and their interaction on the improvement of presented judgments. As dependent variable, we computed the improvement by subtracting the accuracy of the presented judgments from that of the revised judgments. For this purpose, accuracy was defined as the Euclidean distance between a judgment and the correct position. For the presented judgments, accuracy is thus equivalent to the presented deviation (i.e., a distance of 0, 40, 80, or 120 pixels to the correct position). Positive (negative) values of the improvement measure imply that a revised judgment is more (less) accurate than the presented judgment. Since participants could decide whether to adjust or maintain presented judgments, we only included trials in the analysis in which participants actually adjusted the presented judgments.[2]

[1] Including random slopes on items for presented deviation led to comparable results. For the sake of consistency with the models in Experiment 2 and 3, we only report the results of models which include random intercepts in Experiment 1.

[2] The statistical analysis yielded similar results when including non-adjusted judgments (coded with an improvement score of zero) in the analysis (main effect of expertise: β = 12.200, CI = [10.523, 13.876], $t(275.82)$ = 14.263, p < .001; main effect

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

Table 1. *Fixed-effects coefficients of the fitted (generalized) linear mixed models.*

	Independent variable	β	*SE*	95% CI		*p*
				LL	UL	
Dependent variable: Change probability						
Experiment 1	Presented deviation	0.444	0.026	0.392	0.495	>.001
	Expertise	0.622	0.214	0.202	1.042	.004
	Presented deviation × expertise	0.218	0.027	0.165	0.272	>.001
Experiment 2	Presented deviation (V-shaped contrast)	0.208	0.024	0.160	0.256	>.001
	Presented deviation (linear contrast)	0.312	0.069	0.176	0.448	>.001
	Expertise	0.570	0.205	0.169	0.972	.005
	Presented deviation (V-shape) × expertise	0.056	0.030	−0.003	0.114	.062
	Presented deviation (linear) × expertise	−0.329	0.089	−0.503	−0.155	>.001
Experiment 3	Presented deviation (V-shaped contrast)	0.141	0.013	0.116	0.167	>.001
	Presented deviation (linear contrast)	0.367	0.038	0.292	0.441	>.001
	Expertise	0.052	0.178	−0.297	0.401	.771
	Presented deviation (V-shape) × expertise	0.075	0.018	0.040	0.111	>.001
	Presented deviation (linear) × expertise	0.052	0.053	−0.051	0.156	.322
Dependent variable: Improvement of presented judgments						
Experiment 1	Presented deviation	32.289	0.455	31.398	33.181	>.001
	Expertise	15.545	1.047	13.492	17.598	>.001
	Presented deviation × expertise	3.819	0.453	2.930	4.707	>.001
Experiment 2	Presented deviation (V-shaped contrast)	6.782	0.234	6.323	7.240	>.001
	Presented deviation (linear contrast)	−0.593	0.579	−1.728	0.541	.315
	Expertise	8.827	2.229	4.458	13.196	>.001
	Presented deviation (V-shape) × expertise	0.647	0.233	0.191	1.104	.005
	Presented deviation (linear) × expertise	−0.069	0.546	−1.140	1.002	.899
Experiment 3	Presented deviation (V-shaped contrast)	6.653	0.185	6.290	7.016	>.001
	Presented deviation (linear contrast)	0.791	0.465	−0.122	1.703	.089
	Expertise	16.518	2.968	10.701	22.336	>.001
	Presented deviation (V-shape) × expertise	−0.024	0.257	−0.527	0.479	.925
	Presented deviation (linear) × expertise	−1.516	0.627	−2.745	−0.287	.016

Note: CI = confidence interval. All models included crossed random effects for participants and items. The models for change probability (0 = no adjustment, 1 = adjustment) assumed a logistic link function.

We used improvement as dependent variable in a linear mixed model with (standardized) expertise and presented deviation (linear contrast) as independent variables and added random intercepts for participants and items. Figure 3B displays the average improvement in judgment accuracy, whereas Table 1 shows the estimated regression coefficients. As expected, improvement increased for larger presented deviations ($\beta = 32.289$, $CI = [31.398, 33.181]$) and higher expertise ($\beta = 15.545$, $CI = [13.492, 17.598]$). In line with the expected pattern shown in Figure 1, the model also showed a significant interaction such that more knowledgeable participants showed a steeper increase in improvement than less knowledgeable participants ($\beta = 3.819$, $CI = [2.930, 4.707]$).

of deviation: $\beta = 21.932$, $CI = [21.291, 22.574]$, $t(10, 861.76) = 66.979$, $p < .001$; interaction of expertise and deviation: $\beta = 5.726$, $CI = [5.085, 6.367]$, $t(10, 860.76) = 17.500$, $p < .001$).

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

3. Experiment 2

Experiment 1 allows only weak causal conclusions since expertise was merely measured rather than manipulated. As a remedy, we implemented a new study design in which expertise was operationalized as a skill or strategy. We manipulated the level of expertise in a random-dots estimation task (Honda et al., 2022) in which participants had to estimate the number of randomly positioned, colored dots. Participants in the experimental group learned a strategy to provide accurate estimates for the number of presented points. Importantly for sequential collaboration, the same strategy can also be used to evaluate the accuracy of presented judgments. In contrast, participants in the control condition completed a control task and should thus have no advantage in providing and evaluating judgments. In a pilot study, we examined whether the manipulation of expertise was successful and whether participants in the control condition came up with any solution strategy themselves, which was not the case. The preliminary data were also used to calibrate the time limit per item and to define outliers. Hypotheses, study design, sample size, and planned analyses were preregistered at https://aspredicted.org/8c6wh.pdf. Materials, data and analysis scripts are available at https://osf.io/z2cxv/.

3.1. Methods

3.1.1. Participants

We recruited 124 college students from the University of Marburg and a study exchange platform. Participants received course credit or the opportunity to take part in a gift-card lottery in exchange for participation. The median time to complete the study was 17.30 minutes. We excluded five participants from the analysis. One participant did not complete the study conscientiously, one vastly underestimated and another vastly overestimated the number of dots for most items, one almost always provided the perfectly exact number of dots, and one did not answer the attention-check questions about the instructions correctly.[3] The remaining 119 participants (69.70% female) had a mean age of 25.50 (*SD* = 9.94).

3.1.2. Procedure

Participants were randomly assigned either to the expertise-manipulation condition (referred to as ‘experts’ in the following) or the control condition (‘novices’). Experts were introduced to raster scanning, a strategy for accurately estimating the number of objects on a presented image by mentally overlaying a 3 × 3 raster on top of the presented image. With the raster in mind, one can pick one of the nine areas with an approximately average number of dots and count the number of dots within this box. Next, one simply multiplies the result by nine to obtain an estimate for the total number of dots in the image. To facilitate multiplication in one’s head, we advised participants to multiply the number of dots by 10 and then subtract the number of counted dots once. Participants in the control condition only read an essay about the importance of accurate judgments. Afterward, both groups answered four attention-check questions concerning the instructions.

As practice trials, all participants had to independently estimate the number of dots for five images. Only in the experimental condition, these five images were overlaid with a visible 3 × 3 raster to train raster scanning. Next, participants saw another set of five images, now always shown without a raster, and were again asked to provide independent judgments for the number of presented dots. The judgments in this phase served as a manipulation check. The following sequential phase was similar to the one in Experiment 1. In each trial, participants saw one of the 30 remaining images (20 test images and 10 easy images for motivational purposes), each with an (alleged) judgment of a previous participant regarding the number of shown dots. They decided whether to adjust or maintain

[3] Moreover, six participants in the experimental condition indicated that they did not apply the learned strategy. A robustness analysis in which we also excluded these participants led to similar results as the main analysis: For change probability, all tested effects were significant, whereas for improvement, only the V-shaped contrast for presented deviation, expertise, and their interaction were significant.

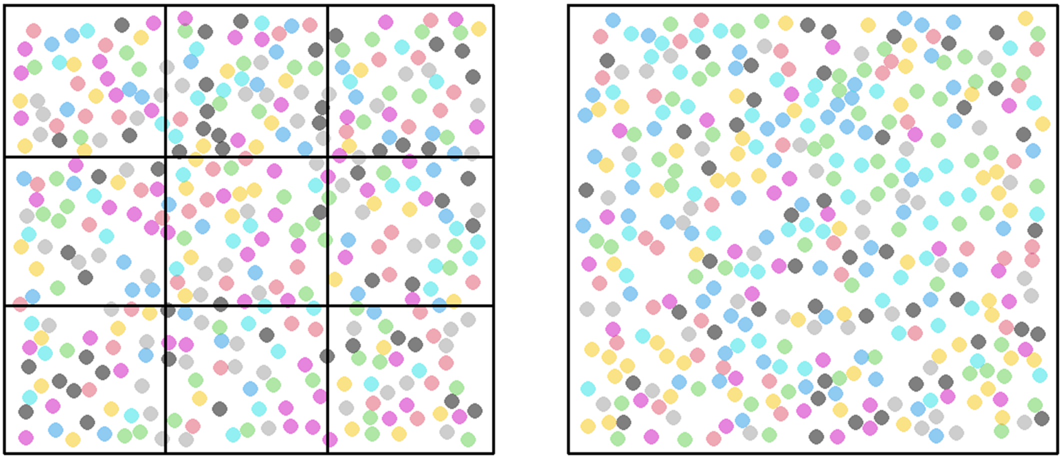

Figure 4. *Example images in the random-dots estimation task presented in Experiments 2 and 3.*
Note: Both images show 379 dots. The left image was used in the training phase for the control condition. The right image displays the 3 × 3 raster overlaid during training in the expertise-manipulation condition. Images presented for the manipulation check and in the sequential phase resembled the left image.

the presented judgment by clicking on respective buttons. Only if they decided to adjust the presented judgment, an open-text box appeared in which the new judgment could be entered. After indicating to maintain the judgment or providing a new judgment, participants continued to the next image. The images were shown in random order with a time limit of 60 seconds (including a warning after 40 seconds). Similar as in Experiment 1, presented judgments were not provided by previous participants but rather preselected to manipulate the deviation of presented judgments from the correct answer. Participants received no additional information about the (alleged) previous contributor. After providing demographic information, participants in the experimental condition were asked whether they had actually used raster scanning, whereas participants in the control condition were asked whether they had used any particular strategy to estimate the number of dots. Finally, we asked whether they completed the study conscientiously and debriefed participants.

3.1.3. Materials

We generated 30 images (600 × 600 pixels; Figure 4) with white background depicting between 100 and 599 randomly positioned, nonoverlapping, colored dots using the R package ggplot2 (Wickham, 2016). Five of these images were used to train participants, and five were used for the manipulation check. The remaining 20 images were shown jointly with an (alleged) judgment of the number of dots. These preselected values were either correct (deviation = 0%) or deviated by ±35% or ±70% from the correct answer. In contrast to Experiment 1, presented deviations were not randomly assigned to items but were fixed for all participants. Moreover, for motivational purposes, we also showed 10 additional images depicting only 10–59 dots which were displayed with a judgment that was either correct or deviated by ±20% or ±35% from the correct answer. For these items, a pilot study showed that participants in both conditions could easily detect whether the presented judgment was correct or not since the time limit allowed to simply count the small number of dots. We manipulated the deviation of presented judgments on five levels. Similar as in Experiment 1, it is thus in principle possible for participants to infer the manipulation if they knew the exact number of dots presented. However, since the deviation was not operationalized by a fixed additive error but rather by a multiplicative constant, it is unlikely that participants could detect the manipulation or acted differently than they would have when seeing actual judgments of previous participants. Moreover, the Supplementary Materials

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

(https://osf.io/z2cxv/) provide plots showing that the five levels of presented deviations fall within the empirical distribution of the independent judgments which were collected for the manipulation check.

3.2. Results and discussion

To test whether the manipulation was successful, we examined whether experts showed a higher accuracy than novices for the five items shown without a previous judgment. As a measure of accuracy, we computed the percentage error for each item, defined as the absolute difference between the judgment and the correct answer divided by the correct answer and multiplied by 100. Using this measure allowed us to analyze average accuracy across items even though the number of dots varied from 100 to almost 600. Including only the independent judgments for the five items in the manipulation-check phase, we fitted a linear mixed model with condition as independent variable (dummy-coded with 1 = expert condition, 0 = novice condition). We found a significant negative effect of condition on the percentage error ($\beta = -16.214$, $CI = [-23.645, -8.784]$, $t(117.16) = -4.277$, $p < .001$). The effect of the manipulation of expertise was large with novices showing a mean error of 35.81% and experts showing a mean error of only 19.59%.

We first tested the expected patterns for change probability shown in Figure 1A. While expertise was coded with a dummy contrast (1 = expert condition, 0 = novice condition), we used two orthogonal, centered contrasts for presented deviation. Since the presented deviation includes both over- and underestimation of the correct answer, we used a centered, V-shaped contrast (values: $4, -1, -6, -1, 4$) to test whether change probability is lowest for correct presented judgments but increases the more presented judgments deviate from the correct judgment. The regression coefficient of this contrast is positive for a V-shape, negative for an inverse V-shape, and zero in the absence of such an effect. Participants, however, may not equally often adjust over- and underestimated presented judgments. Hence, we also included a centered, linear contrast (values: $2, 1, 0, -1, -2$) which tests whether the slope of the V-shaped contrast differs between these two cases. A value of zero indicates a symmetric V-shape, whereas a positive (negative) coefficient indicates a steeper (less steep) slope for underestimated than for overestimated presented judgments.

Figure 5A illustrates the average change probability including 99% confidence intervals. Change probabilities followed the expected V-shape as a function of the presented deviation. Moreover, experts generally changed items more frequently than novices. This impression was confirmed by a significant, positive V-shaped contrast for presented deviation in the linear mixed model ($\beta = 0.208$, $CI = [0.160, 0.256]$) and a significant positive effect of condition ($\beta = 0.570$, $CI = [0.169, 0.972]$) (Table 1). Moreover, we found a positive linear contrast indicating a smaller effect of presented deviation (i.e., a smaller slope of the V-shape) for underestimated than for overestimated judgments ($\beta = 0.312$, $CI = [0.176, 0.448]$). As expected, the interaction between condition and the V-shaped contrast of the presented deviation was positive, meaning that experts better distinguished between accurate and inaccurate judgments ($\beta = 0.056$, $CI = [-0.003, 0.114]$). However, in contrast to our predictions, participants in the expert condition adjusted correct presented judgments more frequently than participants in the novice condition (Figure 5). Besides demand effects, this could be due to the raster-scanning strategy providing only an approximate estimate of the actual number of presented dots. While the approximation leads to improved judgments, it is still prone to errors. Hence, for already accurate presented judgments, participants may have adjusted the judgment even though it was already correct. Lastly, we found a significant interaction between condition and the linear contrast of presented deviation indicating that the V-shape was more symmetric (with respect to over- and underestimated judgments) for experts than for novices ($\beta = -0.329$, $CI = [-0.503, -0.155]$).

Next, we tested whether expertise, presented deviation, and their interaction affect the improvement of presented judgment. Similar to Experiment 1, improvements were computed as the difference between the percentage errors of the presented and the revised judgment. Again, positive (negative) values indicate that presented judgments are improved (worsened). We used a linear mixed model to predict the improvement of presented judgments using the same contrasts for condition and

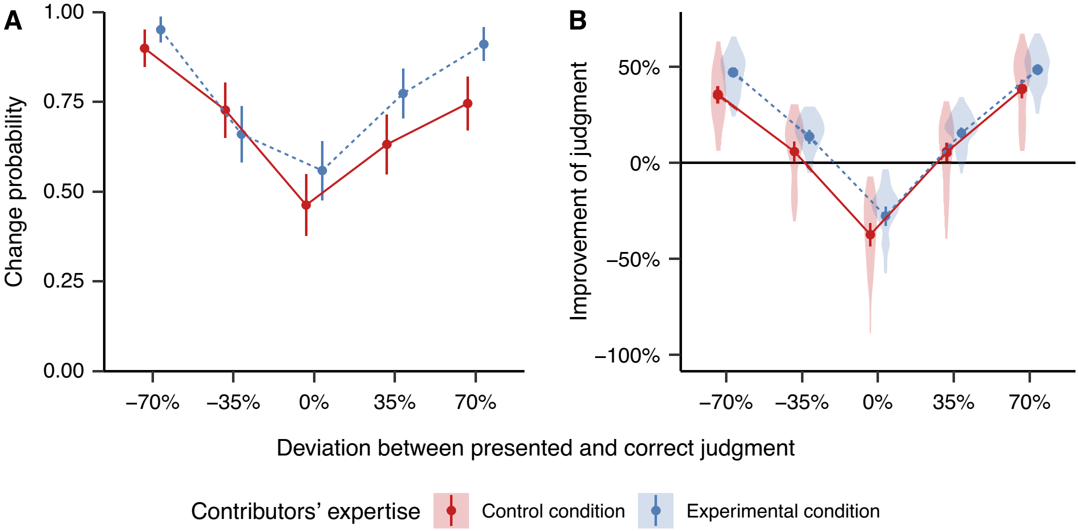

Figure 5. *Change probability and improvement of presented judgments for Experiment 2.*
Note: Points display empirical means with error bars showing the corresponding 99% between-subjects confidence intervals. Violin plots show the distribution of the dependent variable for participants aggregated across items. The plot for improvement only includes trials in which presented judgments were adjusted by participants.

presented deviation as in the model for change probability. Similar as in Experiment 1, we only included trials in which participants adjusted the presented judgment.[4] Figure 5B displays the mean improvement of presented judgments including 99% confidence intervals and violin plots; Table 1 shows the estimated regression coefficients. As expected, presented deviation had a significant V-shaped effect on improvement such that presented judgments were improved more the larger the deviation from the correct judgment was ($\beta = 6.782$, $CI = [6.323, 7.240]$). Compared with the novice condition, participants in the expert condition improved presented judgments more if there was room for improvement and also worsened correct judgments less ($\beta = 8.827$, $CI = [4.458, 13.196]$). Furthermore, the model showed a positive interaction between condition and the V-shaped contrast for presented deviation ($\beta = 0.647$, $CI = [0.191, 1.104]$). These results are closely in line with the expected patterns derived from our theoretical framework (Figure 1).

4. Experiment 3

Experiments 1 and 2 show that change probability and improvement of presented judgments depend on expertise, presented deviation, and their interaction. However, both studies implemented only a single incremental step in sequential collaboration using preselected values for the presented judgments. The same effects should hold if individuals encounter actual judgments of previous individuals rather than preselected judgments. Moreover, as outlined in the Introduction, the benefits of expertise on the accuracy of chains of sequential judgments should be especially pronounced for the final estimates.

To test these assumptions, we again relied on the random-dots estimation task using the raster-scanning strategy as a manipulation of expertise. However, we now implemented a sequential-collaboration paradigm in which participants actually encountered judgments made by previous participants (Mayer and Heck, 2022). The design allowed us to manipulate the number and position

[4] Similar results were obtained when analyzing all trials while assigning an improvement score of zero to maintained judgments (condition: $\beta = 8.659$, $CI = [4.804, 12.514]$, $t(117.01) = 4.403$, $p < .001$; V-shaped contrast for presented deviation: $\beta = 4.833$, $CI = [4.485, 5.181]$, $t(39.41) = 27.225$, $p < .001$; interaction of condition and V-shaped contrast: $\beta = 1.271$, $CI = [0.878, 1.663]$, $t(2, 230.19) = 6.344$, $p < .001$; all other effects were not significant).

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

of experts and novices in a sequential chain. The hypotheses, study design, sample size, and planned analyses were preregistered at https://aspredicted.org/HZT_QW3. Materials, data, and analysis scripts can be found at https://osf.io/z2cxv/.

4.1. Methods

4.1.1. Materials and procedure

We used the same experimental paradigm as in Experiment 2 while making some minor changes. In the expertise condition, we already excluded participants during participation if they did not answer at least three questions about the raster-scanning strategy correctly. Thereby, we avoided the necessity to exclude participants who subsequently contributed to the same sequential chains. We also generated five new items for the sequential-collaboration phase.

Participants were randomly assigned either to the expertise-manipulation or the control condition. We then built sequences of two participants which differed with respect to the status and order of contributors (i.e., novice–novice, expert–novice, novice–expert, and expert–expert). Similarly as in Experiment 2, the first participant in each chain saw preselected judgments which were either correct or deviated by ±35% or ±70% from the correct number of dots. Again, the distribution of independent judgments obtained in the manipulation-check phase revealed that the manipulation for preselected judgments resembled actual judgments made by participants (Supplementary Materials, https://osf.io/z2cxv/). If the first participant in a chain made an adjustment, the second participant saw the revised judgment; otherwise, the second participant merely saw the originally presented judgment. Overall, this procedure results in a mixed design with expertise and composition of dyads as between-subjects factors and presented deviation as within-subjects factor.

4.1.2. Participants

Using a German panel provider, we recruited 464 participants who were compensated by the panel provider for a median participation time of 18.30 minutes. We excluded one participant because they answered ‘1’ to all items, which in turn made it necessary to remove another participant assigned to the same sequential chain. Moreover, we excluded five participants for technical reasons due to duplicate assignments to sequential chains. The final sample included 457 participants (46.80% female) with mean age of 46.16 (SD = 14.36) and heterogeneous educational background (college degree: 34.80%; high-school diploma: 26%; vocational education: 24.10%; lesser educational attainment: 15.10%).

4.2. Results and discussion

We computed the same dependent measures as in Experiment 2. As a manipulation check, we fitted a linear mixed model to test whether the independent judgments for the five items during the manipulation-check phase were more accurate for experts than for novices. As expected, the expertise manipulation leads to a decrease in the percentage error ($\beta = -28.898$, $CI = [-36.319, -21.477]$, $t(117.16) = -4.277$, $p < .001$), indicating that judgments of experts were twice as accurate as those of novices (mean error = 27.46% vs. mean error = 56.36%, respectively).

4.2.1. Effects of one step in sequential collaboration

Replicating the analyses of Experiment 2, we first tested the effects of presented deviation, expertise, and their interaction on change probability and improvement of presented judgments in a single sequential step. We analyzed change probabilities in the sequential phase by including only those participants who saw preselected judgments (but not those who saw the judgments of other participants). Similarly as in Experiment 2, we fitted a generalized linear mixed model to predict whether a presented judgment was changed using the same contrasts for presented deviation and condition.

Figure 6A displays the average change probabilities in Experiment 3. As expected, the V-shaped effect of presented deviation emerged, with a steeper slope for underestimated than for overestimated

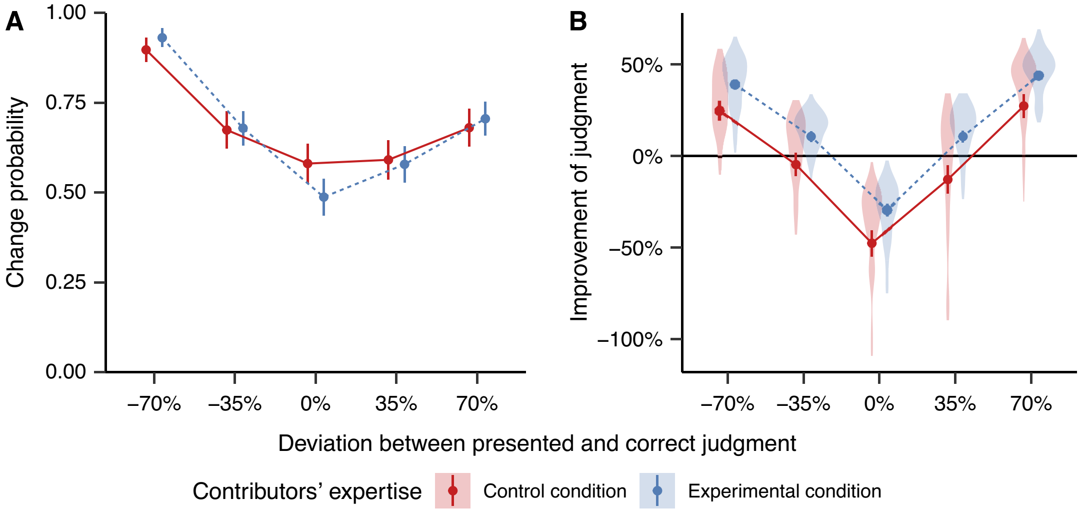

Figure 6. *Change probability, and percentage improvement of presented judgments for Experiment 3.*
Note: Points display empirical means with error bars showing the corresponding 99% between-subjects confidence intervals. Violin plots show the distribution of the dependent variable for participants aggregated across items. The plot for improvement only includes trials in which presented judgments were adjusted by participants.

judgments. In contrast to Experiment 2, the plot does not indicate a main effect of condition. These impressions were supported by the model-based analysis (Table 1) showing a significant V-shaped contrast of presented deviation ($\beta = 0.141$, $CI = [0.116, 0.167]$), but no effect of experimental condition ($\beta = 0.052$, $CI = [-0.297, 0.401]$). The linear contrast of deviation was also significant, indicating a steeper slope for the left than the right limb of the V-shaped effect ($\beta = 0.367$, $CI = [0.292, 0.441]$). Most importantly for sequential collaboration, Figure 6A illustrates the interaction of expertise and presented deviation on change probability. In line with our expectations, experts adjusted presented judgments less often than novices if judgments were already correct, but more often if judgments deviated by $\pm 70\%$ from the correct answer. In the mixed model, the corresponding interaction term of condition and the V-shaped contrast was significant ($\beta = 0.075$, $CI = [0.040, 0.111]$). As described in the Introduction, given that we predicted and found a crossed interaction (Figure 1), the absence of a main effect of expertise may simply be due to a limited range of presented deviations.

Next, we tested the effect of expertise and presented deviation on the improvement of presented judgments. For this analysis, we only included participants at the first chain position.[5] Figure 6B displays the improvement of presented judgments which followed a V-shaped pattern, with already correct presented judgments being slightly worsened. We fitted a linear mixed model for the percentage improvement again using the same contrasts for condition and presented deviation. In line with our hypotheses, the model showed a V-shaped effect of presented deviation ($\beta = 6.653$, $CI = [6.290, 7.016]$), a main effect of condition, indicating more improvement of judgments provided by participants in the expertise than in the novice condition ($\beta = 16.518$, $CI = [10.701, 22.336]$), but no interaction of condition and presented deviation ($\beta = -0.024$, $CI = [-0.527, 0.479]$). Moreover, the interaction between the linear slope for presented deviation and expertise was significant, indicating a steeper slope for the left than the right limb of the V-shape for experts compared with novices ($\beta = -1.516$, $CI = [-2.745, -0.287]$).

[5] Similar results were obtained when including judgments that were maintained with an improvement score of zero (V-shaped contrast of deviation: $\beta = 4.871$, $CI = [4.596, 5.145]$, $t(5, 599.39) = 34.822$, $p < .001$; linear contrast of deviation: $\beta = 1.143$, $CI = [0.417, 1.870]$, $t(5, 599.24) = 3.084$, $p = .002$; condition: $\beta = 12.784$, $CI = [7.998, 17.571]$, $t(236.25) = 5.235$, $p < .001$; all other terms were not significant).

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

4.2.2. Robustness analyses using all participants

As a robustness check, we also tested our predictions for a single sequential step while including participants at both chain positions. The deviation of presented judgments thus becomes a continuous variable since participants at the second chain position may see revised judgments of participants at the first position. In the linear mixed models, we thus included the standardized deviation and the corresponding, quadratic trend as predictors.

For change probability, the results were similar as when including only participants at the first chain position. The model showed a significant quadratic effect of presented deviation ($\beta = 0.308$, $CI = [0.235, 0.381]$, $z = 8.286$, $p < .001$) and, most importantly, a significant interaction with condition ($\beta = -0.112$, $CI = [-0.204, -0.019]$, $z = -2.371$, $p = .018$). Again, the main effect of condition on change probability was not significant ($\beta = 0.145$, $CI = [-0.111, 0.400]$, $z = 1.108$, $p = .268$).

Concerning the improvement of the presented judgments, results were again similar to analyzing only participants at the first chain position. We found a positive effect of the quadratic trend of deviation ($\beta = 12.333$, $CI = [11.530, 13.135]$, $t(7, 289.03) = 30.106, p < .001$) and a positive effect of condition ($\beta = 19.530$, $CI = [14.489, 24.570]$, $t(490.81) = 7.594$, $p < .001$), but also a significant negative interaction ($\beta = -3.261$, $CI = [-4.208, -2.315]$, $t(7, 286.53) = -6.753, p < .001$).

4.2.3. Effects on chains of sequential judgments

We tested the expected impact of experts at the chain level based on data of participants at the second chain position. We first fitted a generalized linear mixed model to predict whether change probability for the second contributor in a sequential chain differed between the four compositions of sequential chains (i.e., novice–novice, expert–novice, novice–expert, or expert–expert). For this purpose, we implemented two contrasts: The first compared novice–novice chains against expert–novice chains, whereas the second compared novice–expert chains against expert–expert chains. In line with our expectations, change probability was larger for experts correcting novices than for expert correcting other experts ($\beta = 0.326$, $CI = [0.063, 0.588]$, $z = 2.432$, $p = .015$). In contrast, novices changed the entries of experts and novices similarly frequently ($\beta = 0.136$, $CI = [-0.098, 0.370]$, $z = 1.140$, $p = .254$). These patterns are illustrated in Figure 7A.

To test how novices and experts improve each other's judgments, we only considered judgments that were adjusted by participants at the second chain position[6] and implemented a linear mixed model with percentage improvement as dependent variable and composition of sequential chain as predictor. We additionally used Helmert contrasts to test our expectations concerning the improvement of each other's judgments by contrasting the novice–expert chain with all other chains, the expert–novice chain with the novice–novice and expert–expert chains, and, lastly, testing the novice–novice and expert–expert chains against each other. Figure 7B displays the empirical means for percentage improvement for all compositions of sequential chains. In line with this pattern, we found a significant Helmert contrast for the novice–expert sequential chain ($\beta = 3.760$, $CI = [1.264, 6.256]$, $t(215.08) = 2.952$, $p = .004$). Furthermore, we found a significant contrast for the expert–novice chain ($\beta = -3.852$, $CI = [-7.227, -0.477]$, $t(221.47) = -2.237, p = .026$). In fact, Figure 7 shows that novices worsen judgments of experts. Lastly, we did not find a significant difference in improvement between expert–expert and novice–novice groups ($\beta = -5.965$, $CI = [-12.137, 0.208]$, $t(222.70) = -1.894, p = .060$). Overall, these findings are in line with our expectations that experts improve judgments of novices most, novices worsen judgments of experts, and only little improvement can be found when novices correct novices and experts correct experts.

[6] Similar results are obtained when maintained judgments are included with an improvement score of zero ($\beta = 3.182$, $CI = [1.195, 5.169]$, $t(214.61) = 3.139, p = .002$ for comparing relative improvement of judgments of novice–expert chains to all other types of sequential chains, $\beta = -2.985$, $CI = [-5.633, -0.336]$, $t(214.50) = -2.209, p = .028$ for comparing expert–novice chains to novice–novice and expert–expert chains, and $\beta = -3.404$, $CI = [-8.161, 1.352]$, $t(214.60) = -1.403, p = .162$ for comparing expert–expert and novice–novice chains).

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

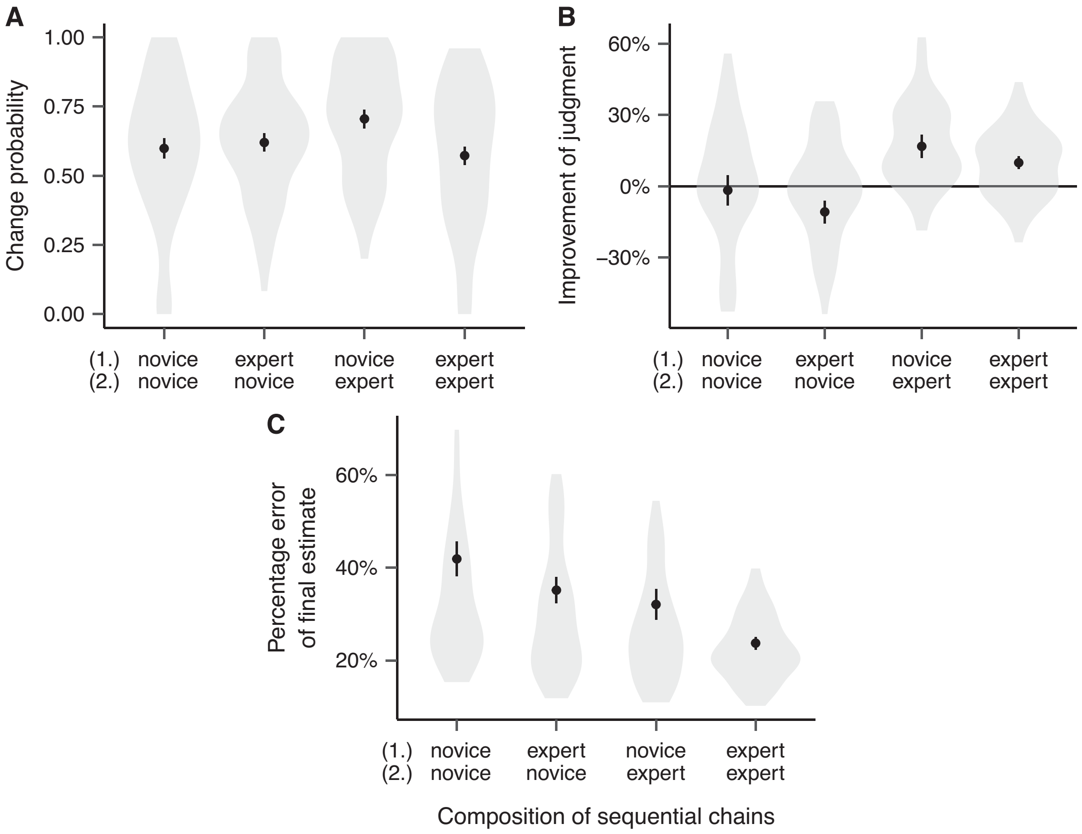

Figure 7. *Change probability, improvement, and accuracy of judgments for the four compositions of sequential chains in Experiment 3.*

Note: Points display empirical means with error bars showing the corresponding 99% between-subjects confidence intervals. Violin plots illustrate the distribution of changes and judgments aggregated for each participant across items.

Finally, we tested which composition of sequential chains lead to the most accurate estimates at the end of a sequential chain. We fitted a linear mixed model with percentage error of the final judgment in a sequential chain as dependent variable and chain composition as predictor. Depending on whether the two participants in each chain adjusted the presented judgment, the final judgment could either be the presented judgment, the judgment entered by the first participant, or the judgment entered by the second participant. We used a linear contrast to test whether percentage error decreases, or equivalently, whether accuracy increases across chain compositions.

As expected, we found a significant linear trend between chain composition and accuracy of the final estimates ($\beta = 5.779$, $CI = [2.199, 9.359]$, $t(216.79) = 3.164, p = .002$). Figure 7C illustrates this pattern with the percentage error being largest for sequential chains with two novices and smallest for sequential chains with two experts. Regarding mixed sequential chains which included both an expert and a novice, the percentage error was smaller when chains ended rather than started with an expert. Overall, the more and the later experts enter sequential chains, the better the final estimates.

5. General discussion

In three experiments, we studied when and how contributors with varying expertise adjust presented judgments in sequential collaboration. The results for individual contributions (i.e., a single sequential step) show that the probability of changing a judgment increases as the deviation to the correct answer increases and as participants' expertise increases. Most importantly, compared with novices, contrib-

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

utors with high expertise were better at distinguishing between accurate and inaccurate judgments as indicated by a steeper slope of the change probability. Core aspects of the predicted data pattern in Figure 1 were thus supported. However, the data did not consistently show that contributors with high expertise adjust perfectly accurate judgments less frequently than contributors with low expertise. Concerning the accuracy of revised judgments, the improvement of presented judgments increased for larger presented deviations and higher expertise in two of three experiments.

Expertise is thus an important predictor of change probability and the amount of improvement of judgments in sequential collaboration. This supports our theoretical assumption that contributors adjust and maintain judgments based on their expertise which in turn facilitates an implicit weighting of judgments. Even though this weighting happens at the individual level within each sequential step, the increased accuracy due to overweighting judgments of contributors with higher expertise can be observed at the chain level (Mayer and Heck, 2022). The data provided evidence for an important prerequisite for such a weighting, namely, contributors with high expertise better differentiate between accurate and inaccurate judgments than contributors with low expertise. However, we found only mixed support for our prediction that contributors with high expertise have a lower change probability for perfectly accurate judgments than contributors with low expertise.

In Experiment 3, we also studied chains of two sequential judgments. As expected, experts adjusted judgments of novices more frequently than those of other experts, and experts improved judgments of novices most, whereas novices tend to worsen judgments of experts. Moreover, the final estimates of sequential chains became more accurate the more and the later experts entered a sequential chain. This shows that the number of experts and the position in which they enter a sequential chain affects the accuracy of group estimates. Accurate judgments of experts at the beginning of a sequential chain may be obstructed by novices later, in turn resulting in reduced accuracy. In contrast, possibly inaccurate judgments by novices at the beginning can be corrected by experts later.

Our findings add to the literature on the wisdom of crowds, supporting the notion that weighting judgments by expertise increases accuracy (Budescu and Chen, 2014; Mayer and Heck, 2023; Merkle et al., 2020). In contrast to other experimental designs and statistical techniques, sequential collaboration does not require researchers to identify experts before or after the judgment task. Instead, sequential collaboration results in an implicit weighting of judgments by expertise. This is achieved by the contributors' metacognitive assessment of whether they can improve a presented judgment. Our results thus shed light on the mechanisms of why the aggregation of individual judgments in sequential collaboration results in high accuracy. Note, however, that the evidence for the high accuracy of sequential collaboration is still sparse (Mayer and Heck, 2022; Miller and Steyvers, 2011). Thus, further studies are necessary to test the robustness and performance of sequential collaboration in different tasks and populations.

5.1. Limitations and future research directions

In all our experiments, we deceived participants about the source of the presented judgments. Both the presented city locations and the number of dots were not judgments of previous participants as stated in the instructions. Instead, we manipulated presented deviation experimentally by generating hypothetical judgments that closely resembled actual judgments. Even though the manipulation used only few levels of deviation, participants would require substantive knowledge about the correct answers for a considerable amount of items in order to become aware of the manipulation. Moreover, due to the design of the sequential-collaboration paradigm (Mayer and Heck, 2022), it is plausible that presented judgments were actually made by other participants previously. This is also supported by the empirical distribution of independent judgments which were collected for measuring expertise (Experiment 1) and as a manipulation check (Experiments 2 and 3). For these items, the preselected deviations fall within the distribution of actual judgments, which provides evidence for their plausibility. In addition, a design presenting participants with authentic judgments by others was implemented in Experiment 3 in which participants formed sequential chains and encountered actual judgments of previous participants.

Irrespective of the source of the presented judgments, all three studies provided converging evidence for the predicted data patterns. Overall, it thus seems unlikely that participants noticed the manipulation and acted differently toward the presented judgments than they would have when seeing actual judgments of previous participants.

We designed the tasks in our experiments to be highly demonstrable, meaning that contributors have the opportunity to demonstrate their expertise (Bonner et al., 2022). However, demonstrability can still be low if participants are not sufficiently motivated to complete a task (Laughlin and Ellis, 1986). If this was the case in our study, contributors with high expertise may have opted out more frequently than would be beneficial for achieving a high accuracy. Also, contributors may have provided generally imprecise judgments and guesses to proceed more quickly. However, these appear to be minor concerns for the validity of our results. Moreover, in ‘natural’ applied settings (e.g., online collaborative projects), the motivation of volunteers to provide demonstrable solutions should be very high.

Our studies show that expertise predicts change probability and improvement in chains of sequential judgments. However, it remains unclear whether the high accuracy of final estimates is due to the sequential judgment process itself or due to the possibility to opt out of answering. Bennett et al. (2018) showed that merely providing an opportunity to opt out increases the accuracy of independent individual judgments. Essentially, individuals use their metacognitive knowledge to select those tasks that fit their individual expertise best. Regarding sequential collaboration, future research should thus disentangle the effects of the judgment-elicitation process (i.e., contributors building a chain of sequential judgments) and of the opportunity to opt out of providing a judgment.

Our three studies are also limited in their generalizability to online collaborative projects such as Wikipedia and OpenStreetMap since they differ in various features. First, we examined the effect of expertise only in one sequential step of sequential collaboration and for short sequential chains of only two contributors even though sequential chains are typically much longer and complex in online collaborative projects. Simplifying such a process into single steps is a typical approach in experimental research. Nevertheless, we expect that the effects of expertise and deviation on change probability and improvement of judgments should similarly hold for longer sequential chains, given that participants are not aware of the number of contributors or previous judgments. Moreover, tasks in our experiments considerably vary from tasks in online collaborative projects. Tasks in these projects are typically more judgmental and less demonstrable than providing numeric or geographical judgments with decisions on which, where, and how to include information while also providing more infrastructure for the contributions such as discussion forums and change logs. In contrast to scientific experiments, contributors are not fully anonymous and typically volunteer for editing in these projects. All these factors may influence whether contributors adjust or maintain Wikipedia articles or OpenStreetMap objects and how they contribute to these projects.

6. Conclusion

Sequential collaboration is a key mechanism found in many large-scale, online collaborative projects. Our studies show that expertise is an important predictor of whether individuals adjust or maintain presented entries, how much they improve an entry, and how accurate the final estimates are. Thereby, we provide first evidence for the implicit weighting of expertise in sequential collaboration, which can explain the high accuracy of online collaborative projects.

Data availability statement. Data and R scripts for the analyses are available at the Open Science Framework: https://osf.io/z2cxv/.

Author contributions. Conceptualization: all authors; Formal Analysis: M.M.; D.W.H.; Investigation: M.M.; M.B.; Methodology: M.M.; D.W.H.; Writing—original draft: M.M.; Writing—review and editing: all authors.

Funding statement. This publication is part of the research group ‘Shared Data Sources’ in the program ‘Collective Decision-Making’ of the WIN-Kolleg of the Heidelberg Academy of Sciences and Humanities, financed by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg. This work was also supported by the Research Training Group ‘Statistical Modeling in Psychology’ funded by the German Research Foundation (DFG grant GRK 2277).

Competing interest. The authors declare no competing interests exist.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Baumann, M. R., & Bonner, B. L. (2013). Member awareness of expertise, information sharing, information weighting, and group decision making. *Small Group Research*, *44*, 532–562. https://doi.org/10.1177/1046496413494415

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, *1*, 90–99. https://doi.org/10.1007/s42113-018-0006-4

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes*, *88*, 719–736. https://doi.org/10.1016/S0749-5978(02)00010-9

Bonner, B. L., Shannahan, D., Bain, K., Coll, K., & Meikle, N. L. (2022). The theory and measurement of expertise-based problem solving in organizational teams: Revisiting demonstrability. *Organization Science*, *33*(4), 1452–1469. https://doi.org/10.1287/orsc.2021.1481

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*, 267–280. https://doi.org/10.1287/mnsc.2014.1909

Ciepłuch, B., Jacob, R., Mooney, P., & Winstanley, A. C. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In *Proceedings of the ninth international symposium on spatial accuracy assessment in natural resources and environmental sciences, 20–23 July 2010* (pp. 337–340). Leicester, UK: University of Leicester.

Dubrovsky, V. J., Kiesler, S., & Sethna, B. N. (1991). The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human–Computer Interaction*, *6*, 119–146. https://doi.org/10.1207/s15327051hci0602_2

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Franz, T. M., & Larson, J. R. (2002). The impact of experts on information sharing during group discussion. *Small Group Research*, *33*, 383–411. https://doi.org/10.1177/104649640203300401

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901. https://doi.org/10.1038/438900a

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*, 682–703. https://doi.org/10.1068/b35097

Herling, R. W. (2000). Operational definitions of expertise and competence. *Advances in Developing Human Resources*, *2*(1), 8–21. https://doi.org/10.1177/152342230000200103

Honda, H., Kagawa, R., & Shirasuna, M. (2022). On the round number bias and wisdom of crowds in different response formats for numerical estimation. *Scientific Reports*, *12*, 8167. https://doi.org/10.1038/s41598-022-11900-7

Hueffer, K., Fonseca, M. A., Leiserowitz, A., & Taylor, K. M. (2013). The wisdom of crowds: Predicting a weather and climate-related event. *Judgment and Decision Making*, *8*, 91–105. http://journal.sjdm.org/12/12924a/jdm12924a.html

Kräenbring, J., Monzon Penza, T., Gutmann, J., Muehlich, S., Zolk, O., Wojnowski, L., Maas, R., Engelhardt, S., & Sarikas, A. (2014). Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology. *PLoS One*, *9*(9), e106930. https://doi.org/10.1371/journal.pone.0106930

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*, 1121–1134.

Lai, E. R. (2011). Metacognition: A literature review. Pearson Research Report. http://images.pearsonassessments.com/images/tmrs/Metacognition_Literature_Review_Final.pdf

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*, 111–127. https://doi.org/10.1287/mnsc.1050.0459

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. https://doi.org/10.1016/0022-1031(86)90022-3

Leithner, A., Maurer-Ertl, W., Glehr, M., Friesenbichler, J., Leithner, K., & Windhager, R. (2010). Wikipedia and osteosarcoma: A trustworthy patients’ information? *Journal of the American Medical Informatics Association*, *17*, 373–374. https://doi.org/10.1136/jamia.2010.004507

Lin, S.-W., & Cheng, C.-H. (2009). The reliability of aggregated probability judgments obtained through Cooke’s classical model. *Journal of Modelling in Management*, *4*, 149–161. https://doi.org/10.1108/17465660910973961

Martire, K. A., Growns, B., & Navarro, D. J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*, *25*, 2346–2355. https://doi.org/10.3758/s13423-018-1448-3

Mayer, M., & Heck, D. W. (2022). Sequential collaboration: The accuracy of dependent, incremental judgments. *Decision*, Advance online publication. https://doi.org/10.1037/dec0000193

Mayer, M., & Heck, D. W. (2023). Cultural consensus theory for two-dimensional location judgments. *Journal of Mathematical Psychology*, *113*, 102742. https://doi.org/10.1016/j.jmp.2022.102742

Merkle, E. C., Saw, G., & Davis-Stober, C. (2020). Beating the average forecast: Regularization based on forecaster attributes. *Journal of Mathematical Psychology*, *98*, 102419. https://doi.org/10.1016/j.jmp.2020.102419

Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgments of magnitude. In J. Salerno, S. J. Yang, D. Nau, & S.-K. Chai (Eds.), *Social computing, behavioral–cultural modeling and prediction* (pp. 236–243). Cham : Springer. https://doi.org/10.1007/978-3-642-19656-0_34

Miller, B. J., & Steyvers, M. (2011). The wisdom of crowds with communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33). Leicester, UK: University of Leicester. https://escholarship.org/uc/item/4jt6q62c

Pinheiro, J. C., & Bates, D. M. (Eds.). (2000). Linear mixed-effects models: Basic concepts and examples. In *Mixed-effects models in S and S-PLUS* (pp. 3–56). New York: Springer. https://doi.org/10.1007/978-1-4419-0318-1_1

Schulze, J., & Krumm, S. (2017). The virtual team player: A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organizational Psychology Review*, *7*(1), 66–95. https://doi.org/10.1177/2041386616675522

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, *23*, 337–370. https://doi.org/10.1207/s15516709cog2303_3

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management*, *20*(2), 503–530. https://doi.org/10.1177/014920639402000210

Steyvers, M., Lee, M., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of order information. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1785–1793). New York: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/4c27cea8526af8cfee3be5e183ac9605-Paper.pdf

Surowiecki, J. (2004). *The wisdom of crowds* (1st ed.). New York: Anchor Books.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., & Baron, J. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. In *AAAI Fall Symposium: Machine Aggregation of Human Judgment*. AAAI publications.

Wang, J., Liu, Y., & Chen, Y. (2021). Forecast aggregation via peer prediction. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 131–142. https://ojs.aaai.org/index.php/HCOMP/article/view/18946

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York : Springer. https://ggplot2.tidyverse.org

Zhang, H., & Malczewski, J. (2018). Accuracy evaluation of the Canadian OpenStreetMap road networks. *International Journal of Geospatial and Environmental Research*, *5*(2). https://dc.uwm.edu/ijger/vol5/iss2/1

Zielstra, D., & Zipf, A. (2010). Quantitative studies on the data quality of OpenStreetMap in Germany. In M. Painho, M.Y. Santos, & H. Pundt (Eds.), *AGILE 2010: The 13th AGILE international conference on geographic information science*.

https://doi.org/10.1017/jdm.2023.3 Published online by Cambridge University Press

Appendix. Cities selected for different maps

Table A1. *Table of items for Experiment 1 using map material.*

Study phase	Map	Cities
Expertise measurement	Austria and Switzerland	Zurich and Basel
	France	Lyon and Nice
	Italy	Venice
	Spain and Portugal	Seville and Lisbon
	United Kingdom and Ireland	Glasgow
	Poland, Czech, Hungary and Slovenia	Budapest
	Germany	Berlin, Nuremberg, Bonn, Münster, Mannheim, Augsburg, Braunschweig, and Munich
Sequential collaboration	Austria and Switzerland	Geneva, Bern, Vienna, Graz, Linz, and Salzburg
	France	Paris, Marseille, and Toulouse
	Italy	Rome, Milan, Naples, and Florence
	Spain and Portugal	Madrid, Barcelona, and Porto
	United Kingdom and Ireland	London, Birmingham, Liverpool, and Dublin
	Poland, Czech, Hungary and Slovenia	Warsaw, Prague, and Bratislava
	Germany	Hamburg, Cologne, Frankfurt, Stuttgart, Düsseldorf, Leipzig, Dortmund, Essen, Bremen, Dresden, Hannover, Duisburg, Wuppertal, Bielefeld, Karlsruhe, Wiesbaden, and Kiel

Cite this article: Mayer, M., Broß, M., & Heck, D. W. (2023). Expertise determines frequency and accuracy of contributions in sequential collaboration. *Judgment and Decision Making*, e2. https://doi.org/10.1017/S1930297523000037