Expertise-Weighing of Judgments in Wisdom of Crowds: Investigating Independent Judgments and Sequential Collaboration

Maren Mayer

Inaugural Dissertation

Submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim

Thesis Defense:
23.09.2022

1st Supervisor:
Prof. Dr. Arndt Bröder

2nd Supervisor:
Prof. Dr. Daniel W. Heck

Dean of the School of Social Sciences:
Prof. Dr. Michael Diehl

Thesis Reviewers:
Prof. Dr. Arndt Bröder
Prof. Dr. Edgar Erdfelder
Prof. Dr. Thorsten Meiser

Für Christian

Contents

Summary

In the past, collaboration to form an aggregate from individual judgments or preferences was mostly investigated for either wisdom of crowds or group decision making. While wisdom of crowds was typically examined by statistically aggregating independent individual judgments, group members form a consensus decision by information sharing and discussing in group decision making. Even though very different, both of these methods were demonstrated to profit from considering expert judgment in the aggregation process. However, the Internet vastly changed the way individuals can collaborate, share information, and form judgments. Large-scale online collaborative projects such as Wikipedia and OpenStreetMap rely on sequential collaboration, a process in which contributors consecutively adjust or maintain the latest versions of entries.

In this thesis comprising three articles, I add to research showing that weighing independent individual judgments by expertise improves resulting estimates. Moreover, I demonstrate that sequential collaboration is a successful way of aggregating individual judgments which relies at least partially on an implicit expertise-weighing of judgments by contributors. In the first paper, I extend Cultural Consensus Theory to two-dimensional continuous data which allows to derive estimates from independent individual location judgments while simultaneously considering individuals' competence. With this model extension, I show that aggregating location judgments with Cultural Consensus Theory yields more accurate estimates than unweighted averaging. In the second paper, I examine judgment aggregation with sequential collaboration showing that sequential collaboration is a successful way of integrating individual judgments which results in similar accurate estimates as unweighted averaging. Lastly, I investigate the role of expertise in sequential collaboration in the third paper. There I show that sequential collaboration allows contributors to implicitly weigh judgments by expertise which results in more accurate estimates the more and later experts enter sequential chains.

With my thesis, I aim to further deepen and extend the understanding of how expertise influences judgments and estimates in wisdom of crowds and establish a theoretical framework of sequential collaboration. Thereby, I hope to contribute to the understanding of successful judgment aggregation and provide a theoretical underpinning for the success and high information quality in large-scale online collaborative projects.

Articles

This cumulative thesis comprises three articles of which two are submitted for publication, and one is currently under review. The main text provides an overview of information collection and information sharing in online collaborative projects before detailing how collaboration was examined so far and how it is influenced by judgment dependency and expertise as a starting point for my research. Thereafter, the manuscripts are summarized and discussed in the order in which they are listed below. In these summaries, I refrain from providing specific details which can be found in the appended manuscripts.

Article I

Mayer, M., & Heck, D. W. (2022). Cultural Consensus Theory for Two-Dimensional Data: Expertise-Weighted Aggregation of Geographic Location Judgments. Manuscript under review. http://doi.org/10.31234/osf.io/unhvc

Article II

Mayer, M., & Heck, D. W. (2022). *Sequential Collaboration: The Accuracy of Dependent, Incremental Judgments. Decision*. Advanced online publication. https://doi.org/10.1037/dec0000193

Article III

Mayer, M., Broß, M., & Heck, D. W. (2022). *Expertise Determines Frequency and Accuracy of Contributions in Sequential Collaboration.* Manuscript under review. http://doi.org/10.31234/osf.io/s7vtg

1 Introduction

Over the past three decades, the Internet spread across the globe and the share of worlds' population regularly accessing the Internet rose from 1% in 1990 to almost 50% in 2017 (Roser et al., 2015). The Internet and its increased use around the globe offered new possibilities for people to collaborate as sharing information and making it publicly accessible became easier than ever. This enabled individuals to collaborate asynchronously as they did not need to stay in the same place and work at the same time anymore. In the following, large-scale online-collaborative projects such as Wikipedia and OpenStreetMap emerged which aim at collecting information and making it freely available to everyone. While contributors in Wikipedia create an online encyclopedia in various languages (Wikipedia Contributors, 2022b), contributors of OpenStreetMap seek to create a comprehensive world map (OpenStreetMap Contributors, 2022).

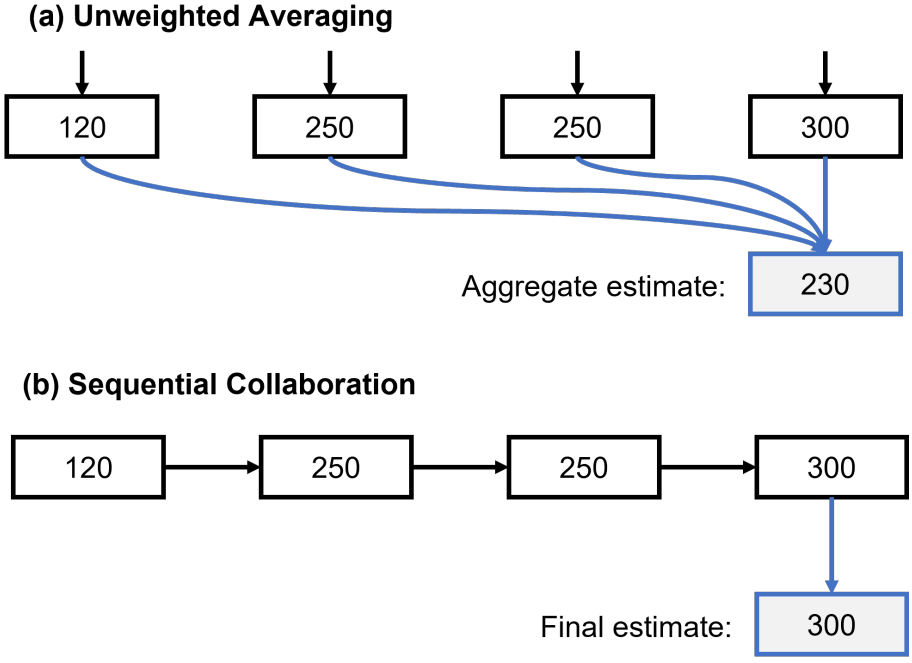
Both projects are frequently accessed, continuously developed and extended. As of May 2022, each month two billion unique devices accessed Wikipedia and nine million users created 15 million edits for all language versions (Wikipedia Contributors, 2022a). Similarly, OpenStreetMap had already 8.5 million registered users in January 2021. Each month 40,000 to 50,000 users were actively working on this project and added millions of new objects to the map each day (OpenStreetMap Contributors, 2021).

Despite the success of online collaborative projects, there are recurring concerns about information quality. Due to little monitoring by moderators (Giles, 2005), vandalism (The Signpost, 2021), or political interest (Steinwehr & Bushuev, 2021; The Signpost, 2021) which can result in frequent back-and-forth changing referred to as edit war (Kittur et al., 2007; Quinn & Bull, 2019), information in these projects may be misleading or even incorrect. Nonetheless, research has demonstrated highly accurate information for Wikipedia compared to the Encyclopedia Britannica (Giles, 2005), university textbooks (Leithner et al., 2010), or governmental information (Kräenbring et al., 2014) while OpenStreetMap yields highly accurate information compared with governmental sources (Haklay, 2010) and commercial map services (Ciepłuch et al., 2010; Zielstra & Zipf, 2010). These findings illustrate that even though these projects do not benefit from all contributions made, information quality is still overall very high.

To collect information, large-scale online collaborative projects such as Wikipedia and OpenStreetMap rely on wisdom of crowds which describes the phenomenon that aggregating individual judgments of a group of individuals concerning the same task results

in highly accurate estimates (Galton, 1907; Larrick & Soll, 2006; Surowiecki, 2005). Even though wisdom of crowds is often referred to as an explanation for high information quality in these projects (Arazy et al., 2006; Baeza-Yates & Saez-Trumper, 2015; Kittur & Kraut, 2008; Niederer & van Dijck, 2010), crowd wisdom was mostly assessed focusing on unweighted averaging of independent individual judgments (de Oliveira & Nisbett, 2018; Hueffer et al., 2013; Larrick & Soll, 2006). In contrast, contributors in Wikipedia and OpenStreetMap do not provide independent judgments that are statistically aggregated but rather judgment aggregation is organized in a dependent, sequential manner (Mayer & Heck, 2022b). In sequential collaboration, one contributor starts with creating an entry which is consecutively adjusted or maintained by subsequent contributors encountering it. This distinction is also depicted in Figure 1 contrasting unweighted averaging of independent individual judgments and sequential collaboration in forming an estimate from numerical judgments.

FIGURE 1: Process of generating estimates with either unweighted averaging or sequential collaboration for the question "How tall is the Eiffel Tower?" answered in meters.



In this thesis, I discuss how collaboration was investigated so far focusing on independent individual judgments on the one hand and group decision making on the other hand. Thereby, I highlight how judgments and estimates obtained in both settings can incorporate and benefit from expertise of single group or crowd members. In my research, I first look into how weighing judgments by expertise can improve the aggregation of independent individual judgments compared to unweighted averaging which

is frequently investigated in the context of wisdom of crowds. To this end, I develop an extension of Cultural Consensus Theory for two-dimensional continuous data in the first paper, a method that allows to derive latent cultural truth estimates from independent individual location judgments while simultaneously considering informants' competence and item difficulty. Applying this method to location judgments of various European cities, I demonstrate that aggregating independent individual location judgments is improved with weighing by expertise compared to unweighted averaging. Furthermore, I investigate sequential collaboration and the role of expertise in this judgment aggregation process which further extends the scope of wisdom of crowds to the aggregation of dependent, incremental judgments as present in online collaborative projects. In the second paper, I therefore systematically examine sequential collaboration in three experimental studies using numeric judgments and geographic location judgments. Thereby, I show that change probability and change magnitude decrease over the course of a sequential chain while judgment accuracy increases. This ultimately results in similarly accurate estimates as can be obtained with unweighted averaging of judgments. Lastly, I shed light into the role of expertise in sequential collaboration with another three experimental studies finding that more knowledgeable individuals contribute more to judgments in sequential collaboration than less knowledgeable individuals. Thus, sequential collaboration allows contributors to weigh their judgments implicitly by expertise without explicitly estimating individuals' expertise and weighing judgments accordingly.

2 Collaboration Between Wisdom of Crowds and Group Decision Making

In the following, I introduce wisdom of crowds and group decision making as methods to elicit group estimates which greatly differ in the extent to which judgments are dependent among judges. Thereby, I especially focus on the role expertise plays for the accuracy of provided judgments and derived estimates.

2.1 Wisdom of Crowds

Prior research has demonstrated that aggregating individual judgments, for instance concerning the dressed weight of an ox, yields highly accurate estimates which result in less error than the average error of provided judgments and are often more accurate than even the crowds best judgment (Galton, 1907; Larrick & Soll, 2006; Surowiecki, 2005). This phenomenon is referred to as wisdom of crowds (Surowiecki, 2005). Wisdom of crowds was not only found to yield highly accurate estimates for numerical judgments in forecasting (Clemen, 1989) but also for various other contexts and tasks such as small crowds (Wagner & Vinaimont, 2010), combinatorial problems like the traveling salesman problem (Yi et al., 2012), rank order tasks (Steyvers et al., 2009), or even estimating the day of ice melting on an Alaskan river (Hueffer et al., 2013).

Highly accurate estimates obtained with wisdom of crowds can be attributed to the cancellation of individual errors (Hogarth, 1978; Larrick & Soll, 2006). Due to this mechanism of error cancellation, judgment aggregation with wisdom of crowds performs best if judgments fall on both sides of the correct answer and thereby bracketing it as symmetrically as possible (Larrick & Soll, 2006). Moreover, if crowds are diverse (de Oliveira & Nisbett, 2018; Larrick et al., 2012), individuals provide judgments independently (Larrick et al., 2012), and judgments are negatively correlated (Davis-Stober et al., 2014), estimates obtained with wisdom of crowds are likely to be highly accurate since these features ensure that a wide range of relevant information is incorporated in the provided judgments further facilitating error cancellation. However, even if these optimal conditions are not met, obtained estimates are still accurate as wisdom of crowds is highly robust against biases (Davis-Stober et al., 2014).

Research on wisdom of crowds mostly focused on independent judgments even

though judgment independence is not a necessary precondition for wisdom of crowds. Thus, it is still unclear whether presenting judgments of others to the judge is beneficial or detrimental for subsequently provided judgments and resulting estimates. Since aggregating individual judgments as a measure of crowd wisdom profits from error cancellation, dependent judgments may cause a restriction in judgment range resulting in less error cancellation and ultimately in less accurate estimates (Becker et al., 2017; Larrick et al., 2012). In line with this notion, judgments may also become more homogeneous due to anchoring (Mussweiler et al., 2004; Tversky & Kahneman, 1974) if they are not independent. These ideas are supported by Lorenz et al. (2011) who found that presenting judges with others' judgments results in diminished diversity of subsequently provided judgments and reduced bracketing leading to a decrease in estimate accuracy. However, several other studies demonstrated that presenting judgments of others can be beneficial for wisdom of crowds under certain conditions. Becker et al. (2017) presented the aggregate of others' judgments to participants which either incorporated judgments equally weighted or overweighted one judgment compared to the others. Presenting an equally weighted aggregate to participants improved subsequent individual judgments even though these judgments became less diverse. However, presenting an unequally weighted aggregate resulted in participants' subsequent judgments converging towards this aggregate irrespective of whether it was accurate or inaccurate. Moreover, both Minson et al. (2018) and Navajas et al. (2018) found that allowing individuals to discuss their answers to a question after they provided initial independent judgments leads to more accurate group judgments and individual judgments after the discussion. Finally, presenting individuals with the current best judgment in the crowd leads to improved judgments through imitation (King et al., 2012). Overall, dependence among judgments might be beneficial under some condition such as providing initial judgments, encountering already highly accurate judgments, or being presented with an unweighted aggregate of others' judgments.

2.2 Group Decision Making

In contrast to wisdom of crowds, research on group decision making does not rely on eliciting individual judgments to statistically aggregate into an estimate. This area of research rather focuses on groups developing a consensus decision from a set of individual preferences or positions by discussion (N. L. Kerr & Tindale, 2004). Even though groups can outperform individual judgments in such settings (Sniezek & Henry, 1989, 1990), their judgments are worse than the one of best group member (Gigone & Hastie, 1997).

Importantly, group decision making is impaired if not all group members are pro-

vided with the same information prior to discussing a consensus decision. In the hidden profile paradigm developed by Stasser and Titus (1985), participants are provided with information on several choice alternatives, for instance potential student representatives (Stasser & Titus, 1985) or football players (Franz & Larson, 2002). Before engaging in a group discussion, all group members are presented with information on the choice alternatives which is partially shared by all group members and partially only presented to a single group member. Thereby information is distributed such that shared information indicates a sub-optimal choice alternative while shared and unshared information combined indicate the optimal alternative. Research on the hidden profile paradigm revealed that decisions are biased towards shared information which results in incorrect group decisions (Lu et al., 2012; Mesmer-Magnus & DeChurch, 2009; Stasser & Titus, 1985). This is due to a lack of information sharing such that unshared information is less mentioned and less discussed (Lu et al., 2012; Mesmer-Magnus & DeChurch, 2009). Moreover, information sharing is not increased if group discussion and interaction is computer mediated rather than in person and groups are equally likely to solve the hidden profile in both settings (Dennis, 1996; Dennis et al., 1998; D. S. Kerr & Murthy, 2009; Lu et al., 2012).

Impaired information sharing found in the hidden profile paradigm was attributed to different group-level as well as individual-level processes. Wittenbaum et al. (1999) found that the discussion can serve as social validation for information supporting initial preferences while Gigone and Hastie (1993) demonstrated that the discussion may be used to negotiate preferences rather than share information which results in premature decisions of an incorrect alternative. Furthermore, social norms (Postmes et al., 2001) or individuals' tendency towards social desirability (Henningsen & Henningsen, 2004) also foster repetition of shared information. However, even if all relevant information is exchanged, individuals nonetheless tend to retain their initial decision preference resulting in erroneous decisions (Faulmüller et al., 2010; Greitemeyer & Schulz-Hardt, 2003). These findings demonstrate that forming a consensus decision through discussion is prone to several biases leading to inaccurate group decisions.

2.3 Improving Estimates and Decisions with Weighing by Expertise

Expertise is a multi-faceted concept which has no clear definition but consists of several facets instead (Baumann & Bonner, 2013). It comprises abilities like logical reasoning (Kruger & Dunning, 1999), semantic knowledge like received information (Stewart & Stasser, 1995) or grammar rules (Kruger & Dunning, 1999), procedural skills such as design rules for experiments (Schunn & Anderson, 1999), or experience like students

judging their own curriculum (Dubrovsky et al., 1991) or handwriting experts judging handwriting features (Martire et al., 2018). Formal education can result in expertise as it incorporates several of these aspects (Martire et al., 2018). All these facets of expertise imply some domain specificity. Furthermore, experts are expected to conduct tasks differently (Dubrovsky et al., 1991; Franz & Larson, 2002; Schunn & Anderson, 1999) which results in better task performance on expertise-related tasks (Budescu & Chen, 2014; Merkle & Steyvers, 2011; Merkle et al., 2020).

Expertise can be determined in various ways depending on the facet of expertise that is assessed. Formal education can indicate individuals' expertise, thereby including various aspects of expertise such as knowledge and experience (Martire et al., 2018). Moreover, judgment confidence can serve as a measure of expertise as both are positively related (Meyen et al., 2021; Palley & Satopää, 2022) if individuals do not hold widespread but erroneous beliefs (Koriat, 2008, 2011). Another way to derive individuals' expertise is to rely on their predictions about others' judgments. Experts can accurately predict how likely others are to provide the same judgment (surprisingly popular method, Lee et al., 2018; Prelec et al., 2017) and which judgments others are likely to provide in general (social projection, Grüning & Krueger, 2021), or which judgments their peers are likely to provide (peer prediction, Wang et al., 2021). Lastly, expertise can be measured by the performance on previous tasks (Lin & Cheng, 2009; Mannes et al., 2014) or the same task (Budescu & Chen, 2014; Merkle & Steyvers, 2011; Merkle et al., 2020).

Both wisdom of crowds and group decision making were found to profit from considering crowd or group members' individual expertise. While experts help to improve group decisions, it is necessary to communicate the expert role explicitly to other group members before discussions starts (Baumann & Bonner, 2013; Bonner et al., 2002). Estimates obtained with unweighted averaging of independent individual judgments also improve if judgments are weighted by individuals' expertise for the aggregation (Budescu & Chen, 2014; Lin & Cheng, 2009; Mannes et al., 2014; Merkle & Steyvers, 2011; Merkle et al., 2020).

Mayer, M., & Heck, D. W. (2022). Cultural Consensus Theory for Two-Dimensional Data: Expertise-Weighted Aggregation of Geographic Location Judgments. Manuscript under review. http://doi.org/10.31234/osf.io/unhvc

Building on the findings that weighing independent individual judgments by expertise improves resulting estimates, we extend Cultural Consensus Theory (CCT, Romney et al., 1986), a method allowing the derive estimates from judgments while simultaneously considering individuals' competence, to two-dimensional continuous data. This

model extension allows to compare the accuracy of location estimates obtained with either unweighted averaging and a weighing by individuals' competence both for simulated and empirical data.

CCT was originally developed in anthropological research to examine cultural beliefs for which both correct answers and informants' competence are unknown to the researcher (Romney et al., 1986). Therefore, CCT is also referred to as "test theory without an answer key" (Batchelder & Romney, 1988). Optimally aggregating individual judgments is more difficult if correct answers to questions are not known since it is unclear which informants provide the most accurate judgments. As a remedy, CCT allows to identify latent cultural truths while taking into account informants' competence and item difficulty simultaneously by assuming that highly competent informants show similar answer patterns on a certain set of questions as they are expected to provide judgments similar to the latent cultural truth (Romney et al., 1986). For this purpose, CCT requires multiple informants to provide judgments for numerous items from the same knowledge domain (Weller, 2007). Since its development to elicit one cultural truth from a set of questions with dichotomous answer options, CCT was extended both to multiple cultural truths (Anders & Batchelder, 2012) as well as to other response formats such as continuous, ordinal, or mixed data (Anders & Batchelder, 2015; Anders et al., 2014; Aßfalg, 2018; Batchelder & Anders, 2012). Recently, CCT models have been implemented using hierarchical Bayesian modeling (Anders & Batchelder, 2012; Anders et al., 2014).

Even though CCT was originally developed for scenarios in which correct answers to the questions asked are unknown, CCT can also be applied to scenarios in which correct answers become available later as in forecasting tasks or are already known from the beginning as for knowledge questions. These cases are especially interesting for method comparison since the estimates obtained with different methods can be compared against the correct answer. For instance, Merkle et al. (2020) recently demonstrated that a CCT-inspired model outperforms simple unweighted averaging and Waubert de Puiseau et al. (2017) showed that accuracy of eyewitness testimonies increases when aggregated with CCT.

Despite various extensions and applications in the past, CCT has not been extended to two- or higher-dimensional data even though such an extension would allow to apply CCT to a larger set of scenarios such as geographical location judgments. Hence, we developed a model extending CCT to two-dimensional continuous data (CCT-2D) and applied this model to geographical location judgments in order to compare the accuracy of location estimates obtained with either CCT-2D or unweighted averaging.

To develop a CCT model for two-dimensional continuous data, we extended the model for one-dimensional continuous data by Anders et al. (2014). As Figure 2A dis-

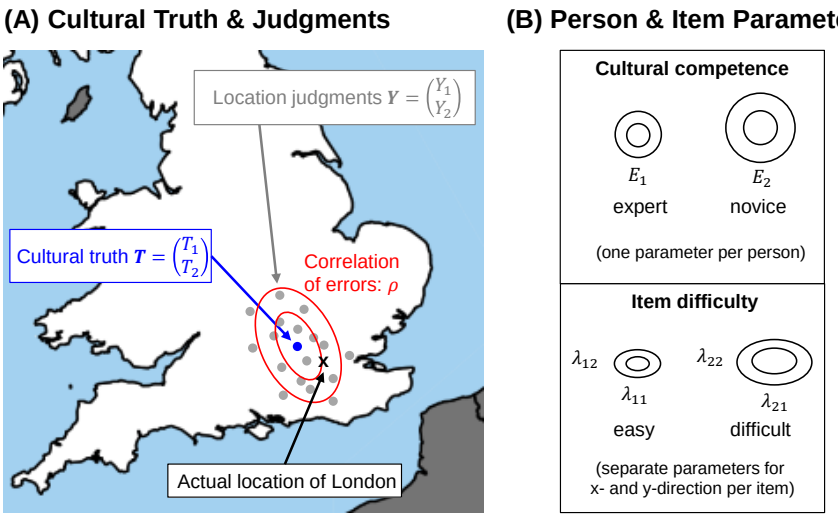
plays, observed judgments $\boldsymbol{Y}_{ik}$ for items k answered by informants i are modeled by one shared cultural truth and an unsystematic error

$$\boldsymbol{Y}_{ik} = \boldsymbol{T}_k + \boldsymbol{\varepsilon}_{ik}. \tag{2.1}$$

Similar to other CCT models, we assume conditional independence of errors $\boldsymbol{\varepsilon}_{ik}$ given the person competence E_i and the item difficulty $\boldsymbol{\lambda}_k$. Since judgments are continuous and two-dimensional, a bivariate normal distribution of errors is assumed

$$(\boldsymbol{\varepsilon}_{ik} \mid E_i, \boldsymbol{\lambda}_k) \overset{\text{iid}}{\sim} \text{MV-Normal}\,(\mathbf{0}, \boldsymbol{\Sigma}_{ik})\,. \tag{2.2}$$

FIGURE 2: Data structure and parameters of CCT-2D for location judgments of London.



The resulting covariance matrix of errors $\boldsymbol{\Sigma}_{ik}$ is modeled as a function of informants' competence and item difficulty. Thereby, error variances in both x- and y-direction are assumed to be smaller for informants with higher competence and easier items resulting in judgments closer to the cultural truth. Since dimensions may vary in difficulty as geographical features such as borders, coasts, lakes, or other landmarks may restrict positioning of locations, we assume one item difficulty parameter for each dimension, λ_{k1} and λ_{k2}. However, competence is not assumed to vary between dimensions and can be interpreted as a one-dimensional trait as it is not likely that informants' geographical knowledge varies between longitude and latitude of locations. Figure 2B displays the effect of informants' competence and item difficulty on the variance of the distribution of errors. As depicted, smaller values of E_i indicate higher competence while smaller values in $\boldsymbol{\lambda}_k$ indicate easier items both resulting in judgments closer to the cultural

truth. Lastly, certain geographical feature such as diagonal coastlines or borders may also lead to spatially correlated errors of location judgments. We therefore consider that errors for each dimension might correlate with ρ_k within each item as illustrated in Figure 2A. Nevertheless, errors do not correlate between items or informants and thereby not violate conditional independence of errors. These assumptions result in the following covariance matrix for two-dimensional judgment errors or CCT-2D:

$$\mathbf{\Sigma}_{ik} = \begin{pmatrix} (E_i\lambda_{k1})^2 & \rho_k E_i^2 \lambda_{k1}\lambda_{k2} \\ \rho_k E_i^2 \lambda_{k1}\lambda_{k2} & (E_i\lambda_{k2})^2 \end{pmatrix}. \tag{2.3}$$

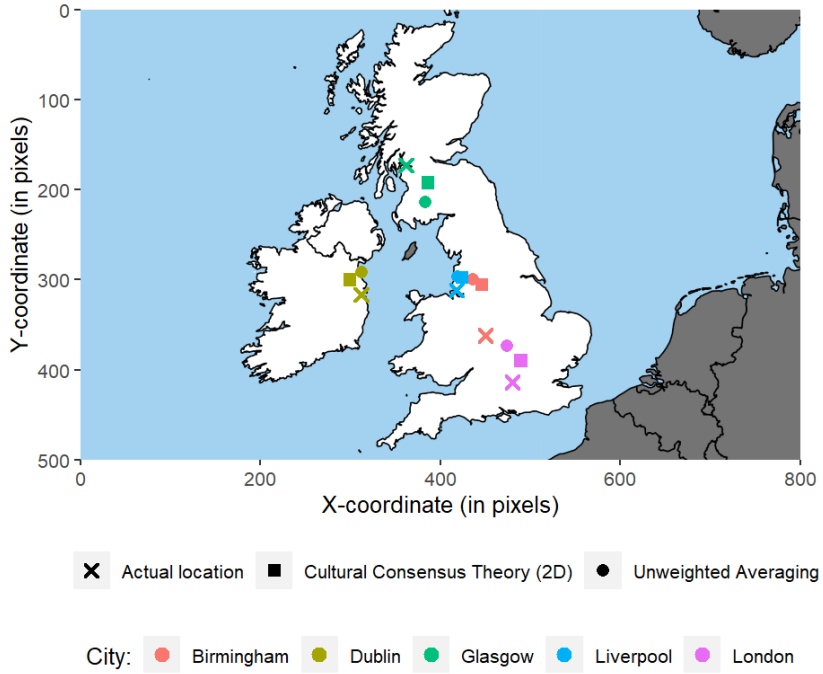
In contrast to Anders et al. (2014), we did not assume multiple cultural truths, an additive response bias, or a scaling bias. Since this model is developed to compare the accuracy of geographic location estimates obtained with either CCT-2D or unweighted averaging and correct answers to all questions need to be known in this case, we do not consider multiple cultural truths as CCT is most useful in such a case if it provides a single competence-weighted estimate for each item. Moreover, we consider an additive response bias as unlikely in the context of two-dimensional location judgments since this would imply that all judgments are shifted horizontally, vertically, or diagonally irrespective of coastlines, borders, or other geographical features constraining possible responses. Lastly, we also omitted a scaling bias from the model as a multiplicative bias heavily depends on the underlying scale of the judgments. However, for geographical location judgments such a bias would depend on the location of the of the origin of the underlying coordinate system which is unknown to the informants. To estimate all parameters included in CCT-2D, we adopted the hierarchical Bayesian modeling approach by Anders et al. (2014) assuming separate population distributions of competence parameters E_i across informants and item difficulty parameters $\boldsymbol{\lambda}_k$ across items and implemented CCT-2D in JAGS (Plummer, 2003).

To asses the models' parameter recovery and to compare the location estimates obtained with CCT-2D and unweighted averaging, we first performed a simulation study in which we varied the number of informants and items as well as the variance of informants' competence and item difficulty. Parameter recovery for cultural truths $\boldsymbol{T}_k$, individuals' competence E_i, item difficulty $\boldsymbol{\lambda}_k$, and correlation of errors ρ_k was satisfactory especially for larger numbers of participants and items and for larger variance of informants' competence and item difficulty. However, recovery decreased with only little variance in informants' competence and item difficulty combined with only few items or informants, respectively. Concerning the comparison of judgment aggregation with CCT-2D and unweighted averaging, we found similar accurate estimates if there was no variance in informants' competence. However, CCT-2D outperformed unweighted averaging with increasing variance in informants' competence as it takes informants'

competence into account when estimating locations while unweighted averaging does not. Nonetheless, unweighted averaging profits from error cancellation such that estimates became increasingly more accurate with larger numbers of informants.

In order to apply CCT-2D to empirical data, we reanalyzed the subset of independent individual location judgments of 228 participants for 57 European cities collected in a study on sequential collaboration (Mayer & Heck, 2022b) and compared the accuracy of resulting location estimates to those obtained with unweighted averaging. We found that CCT-2D on average led to an improvement of estimates by 12.35 pixels resembling the improvement for Glasgow displayed in Figure 3 which was 15.63 pixels.

FIGURE 3: Comparison of estimates obtained with CCT-2D and unweighted averaging for judgments on the map of the United Kingdom and Ireland.



In this study, we did not only extend CCT to two-dimensional continuous data, but also further contribute to the literature demonstrating that weighing individual judgments by expertise improves estimates compared to unweighted averaging (Merkle et al., 2020). Even though CCT-2D was originally developed to allow aggregating location judgments while simultaneously considering individuals' competence, it can also be applied to other two-dimensional data like ratings of arousal and valence of pictures or trustworthiness and likability of facial images. As CCT-2D was specifically developed to compare expertise-weighted location estimates with estimates obtained with unweighted averaging of judgments, we only assumed one latent cultural truth and did not include judgment biases into the model. However, there may be varying represen-

tations of the location of certain places by different groups (Friedman et al., 2002, 2005) making it useful to consider multiple latent cultural truths. Furthermore, we observed that judgments tend to shift towards the center of presented land mass hinting towards some response bias in judgments. Therefore, developing a model extension considering this bias may be a useful addition to CCT-2D.

3 Sequential Collaboration

Mayer, M., & Heck, D. W. (2022). *Sequential Collaboration: The Accuracy of Dependent, Incremental Judgments. Decision*. Advanced online publication. https://doi.org/10.1037/dec0000193

As outlined above, the spread of the Internet and the subsequent emergence of large-scale online collaborative projects such as Wikipedia and OpenStreetMap vastly changed how individuals can and do collaborate. While information generation in these projects can be regarded as wisdom of crowds, information generation and judgment aggregation differs from typical wisdom-of-crowds settings of averaging independent individual judgments. In sequential collaboration, one contributors starts a sequential chain by providing an initial independent entry. This judgment is subsequently encountered by other contributors who decide whether to adjust or maintain the presented entry. In case of an adjustment, the entry is updated such that only the latest version of the entry is displayed. The high accuracy of information in online collaborative projects suggests that contributors in sequential collaboration reach consensus on a highly accurate entry which is not adjusted and thereby possibly worsened anymore. Imagine contributors sequentially locating Rome on a map of Italy. The first contributor may position the initial judgment near Naples around 150 kilometers south of Rome while the second contributor may correct the judgment to be close to the actual location of Rome. Lastly, a third contributor may not adjust the already highly accurate judgment. Thus, compared to independent individual judgments that are averaged on the one hand and group decision making on the other hand, sequential collaboration has a medium degree of judgment dependency.

Even though sequential collaboration is successfully applied in online collaborative projects, it was vastly overlooked as a method of judgment aggregation (Miller & Steyvers, 2011). Despite its success, several theoretical arguments speak against contributors reaching consensus on a highly accurate judgment and stop adjusting judgments when they reached the correct answer in sequential collaboration. Anchoring on the currently presented judgment may bias the subsequent judgment (Mussweiler et al., 2004; Tversky & Kahneman, 1974) especially if the currently presented judgment is vastly over- or underestimating the correct value. Thereby, anchoring may defer convergence

to the correct answer or even result in convergence to a biased judgment. Moreover, research on advice taking showed that advice in the form of others' judgments is generally underweighted when forming judgments while individuals overweigh their own judgment (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). This could result in constant adjustments which worsen already highly accurate or even correct judgments and hinder convergence.

Nonetheless, judgments in sequential collaboration may benefit from contributors encountering the judgment of a previous contributor. Presenting judgments of others can serve a as frame of reference and thereby prevent extreme judgments (Bonner et al., 2007; Laughlin et al., 1999) which could help contributors in sequential collaboration to make more accurate adjustments. Moreover, as outlined above, presenting (aggregated) judgments of others was shown to improve subsequently provided individual judgments as well as their unweighted average (Becker et al., 2017; King et al., 2012; Minson et al., 2018). Thus, contributors judgments might as well profit from a presented judgment, especially if this judgment is already an aggregate of others' judgments or if it is the current best judgment in the sequential chain. However, these positive effects required either a first independent judgment which is not provided in sequential collaboration or an already highly accurate judgment of a previous participant to imitate which is unlikely to happen in early stages of sequential chains. Furthermore, sequential collaboration allows to opt out of providing a judgment which enables contributors to not provide a new judgment if they feel they cannot improve the previous judgment and their contribution is dispensable (N. L. Kerr & Bruun, 1983). Such an opt-out mechanism was shown to improve estimates obtained from independent individual judgments as individuals seem to select questions to answer based on metacognitive knowledge of their own expertise (Bennett et al., 2018). Thus, if contributors adequately distinguish between judgments to adjust and judgments to maintain based on metacognitive assessments of their expertise, this may also improve judgments and estimates in sequential collaboration. Thereby, contributors may incorporate the advantages of weighing judgments by expertise already into the judgment aggregation process. Lastly, Miller and Steyvers (2011) conducted a study closely resembling sequential collaboration and found that both group aggregates and individual performance increase when providing judgments sequentially rather than independently in a rank-ordering task.

3.1 Obtaining Accurate Judgments and Estimates in Sequential Collaboration

Based on the theoretical considerations above, we systematically investigated sequential collaboration to both show that this process is successful in aggregating individual

judgments and that the resulting estimates are highly accurate (Mayer & Heck, 2022b). To this end, we examined whether change probability and change magnitude decrease over the course of a sequential chain while judgment accuracy in turn increases which would imply convergence of contributors' judgments towards the correct answer. Furthermore, as unweighted averaging yields highly accurate estimates, it was used as a benchmark for comparing the accuracy of estimates obtained with sequential collaboration.

These hypotheses were tested in three experimental studies. In the first two experiments, participants were asked to provide judgments for 65 general knowledge questions either independently or sequentially in chains of four (Experiment 1) or six (Experiment 2) participants. To extend the results of the first two studies to a more complex and ecological valid task, participants located 57 European cities on respective maps building sequential chains of four in Experiment 3. Sequential chains were initiated by one participant who provided independent individual judgments. These judgments were presented to later participants in the sequential chain who decided to adjust or maintain the presented judgments which were updated for later participants in case of an adjustment. Contributors did not know their position in the sequential chain, did not see earlier judgments than the latest one, and did not know how often the presented judgment had already been adjusted. The latest judgments at the end of a sequential chain are considered sequential estimates which were compared to unweighted averaged judgments of same-sized groups of participants who provided independent judgments. Since this paradigm results in a nested data structure with judgments nested in participants and items, we estimated (generalized) linear mixed models to test our hypotheses. Position in a sequential chain served as predictor to examine change probability, change magnitude, and judgment accuracy over the course of a sequential chain while aggregation method, either unweighted averaging of independent individual judgments or sequential collaboration, was used to predict the accuracy of estimates obtained.

As expected, both change probability and change magnitude decreased over the course of a sequential chain in two experiments while there was only an insignificant trend for each effect in one experiment. Moreover, judgment accuracy increased over the course of a sequential chain in all three experiments. These results demonstrate that contributors approach some consensus on their judgments making sequential collaboration a successful strategy to aggregate individual judgments. Furthermore, comparing the error of estimates for sequential collaboration and unweighted averaging revealed more accurate estimates obtained with sequential collaboration in two of three experiments. To check the robustness of the estimate comparison between sequential collaboration and unweighted averaging, we examined other possible aggregates of both indepen-

dent individual judgments and judgments provided sequentially which showed that both unweighted averaging and sequential collaboration result in similar accurate estimates.

To summarize, sequential collaboration is a successful way to aggregate individual judgments and resulting estimates are similar accurate as estimates obtained with unweighted averaging. These findings highlight that dependent, incremental judgments can result in highly accurate estimates, thereby broadening the scope of not only wisdom of crowds in particular but collaboration methods in general.

Sequential collaboration might be suited better for more complex tasks as accuracy of estimates outperformed estimate accuracy of unweighted averaging most for the city-location task. Thereby, contributors may not necessarily change the whole presented judgment but can also decide to adjust only one of two dimensions, thereby maintaining parts of the previous judgment. However, estimating individuals' competence statistically and weighting independent individual judgments accordingly provides even more accurate estimates since estimates obtained with CCT-2D showed 12.35 pixels improvement while sequential collaboration only resulted in 7.22 pixels improvement. Nonetheless, sequential collaboration seems to at least partially enable contributors to correct each other implying that they implicitly weigh their judgments by expertise.

3.2 How Expertise Affects Judgments in Sequential Collaboration

Mayer, M., Broß, M., & Heck, D. W. (2022). *Expertise Determines Frequency and Accuracy of Contributions in Sequential Collaboration.* Manuscript under review. http://doi.org/10.31234/osf.io/s7vtg

While Mayer and Heck (2022b) showed that sequential collaboration is a successful way of aggregating individual judgments, the mechanism behind sequential collaboration is still unclear. Thus, we investigated two possible predictors influencing how frequent contributors adjust judgments and how much an adjustment improves or worsens the previous judgment, namely accuracy of the presented judgment and the contributors' expertise. Irrespective of individuals expertise, the more a presented judgment deviates from the correct answer and thus the more inaccurate it is, the easier it may be for contributors to detect whether they can provide a judgment to improve the presented one. Thus, contributors should adjust judgments more frequently and make larger improvements the more the presented judgment deviates from the correct answer.

Furthermore, as described above, both group decision making and wisdom of crowds

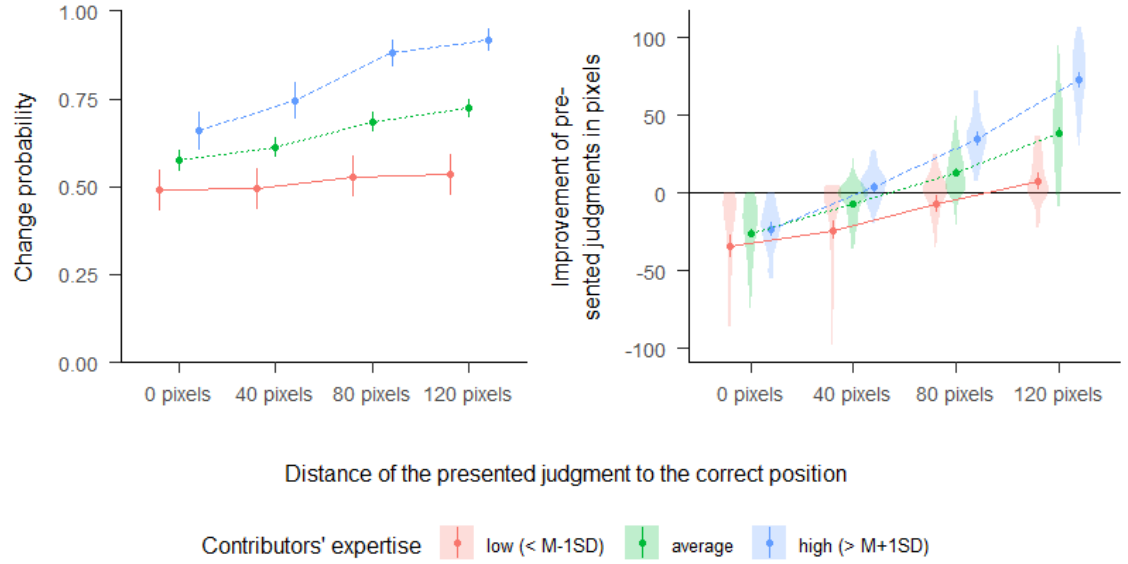
profit from weighing judgments by contributors' expertise. For wisdom of crowds, Bennett et al. (2018) demonstrated that such a weighing can also be achieved by allowing individuals to opt out of providing a judgment since they may do so according to some metacognitive assessment of their own expertise. Thus, highly accurate estimates obtained with sequential collaboration may be due to contributors using the possibility to opt in or opt out of providing a judgment to adjust or maintain judgments according to their individual expertise. Thereby, more knowledgeable contributors should be able to correct previous judgments already if these judgments only deviate slightly from the correct answer and provide highly accurate judgments which should result in large improvements. In contrast, less knowledgeable contributors can only correct previous judgments if these show larger deviations from the correct judgment. In addition, even if such contributors decide to adjust the presented judgment, they cannot to provide a new judgment that is as accurate as the provided judgment of a more knowledgeable contributor. Thus, we expect that individuals with higher expertise adjust previous judgments more frequently and can also improve those judgments more than individuals with lower expertise.

However, as contributors need to rely on their metacognitive assessment about their own expertise to evaluate both the accuracy of the presented judgment and their capacity to improve it, contributors with lower expertise may suffer from the Dunning-Kruger effect (Jansen et al., 2021; Kruger & Dunning, 1999). This phenomenon describes that low expertise may result in a miscalibartion concerning individuals' metacognitive knowledge about their own expertise which leads to an overestimation of one's own expertise. Thus, contributors with low expertise may provide improper adjustments of previous judgments in sequential collaboration as they overestimate their own expertise. Therefore, contributors with higher expertise may be better able to distinguish between previous judgments they can improve and previous judgments they cannot improve.

Based on these theoretical ideas, expertise and deviation of the previous judgment from the correct answer may interact such that contributors with higher expertise change highly accurate or already correct judgments less often and inaccurate judgments more often than contributors with lower expertise. Moreover, while contributors with higher expertise may be able to reach a certain level of accuracy for all their adjustments, contributors with lower expertise who cannot make such precise adjustments may be more influenced by the presented judgment. Such an anchoring effect (Mussweiler et al., 2004; Tversky & Kahneman, 1974) for contributors with lower expertise also results in an interaction effect of deviation of the presented judgment from the correct answer and expertise on improvement on the presented judgment. Thereby, the effect of presented deviation is smaller for contributors with lower compared to higher expertise as their capacity to improve judgments is undermined by an anchoring on the

presented judgment.

FIGURE 4: Change probability and improvement of presented judgments in Experiment 1.



Note. Points and vertical lines show empirical means and corresponding 99% confidence intervals. Violin plots indicate the distribution of the dependent variable aggregated across items for each participant.

These hypotheses were examined in three preregistered experiments (Mayer, Broß, & Heck, 2022). In the first experiment, we adapted the city-location task of Mayer and Heck's (2022b) Experiment 3 such that participants first provided independent judgments for 17 cities which were used as a measure of expertise. The remaining 40 cities were presented with a judgment having a preselected deviation from the correct judgment of 0, 40, 80, or 120 pixels. These judgments were introduced to participants as judgments of previous participants for which they could decide whether to adjust or maintain them. To analyze the hypotheses described above, we estimated (generalized) linear mixed models with individuals' expertise and presented deviation as predictors for frequency of adjustments and improvement of presented judgments and accounted for the nested structure of our data by including random effects for participants and items into the model. Figure 4 depicts the change probability and the improvement of presented judgments depending on the distance of presented judgments to the correct position and participants' expertise. As displayed, presented judgments were more frequently adjusted and more improved the higher participants' expertise was and the more the presented judgment deviated from the correct answer. Moreover, the higher participants expertise was, the stronger the positive effect of presented deviation was on both change probability and improvement of presented judgments. Contrary to the

expectations, more knowledgeable participants also adjusted already correct presented judgments more than less knowledgeable participants. Even though these results hint towards accuracy of the presented judgment and expertise strongly impact frequency of adjustments and improvement of presented judgments, inferences can only be weakly causal since expertise was measured in this study.

Therefore, expertise was manipulated in a second study using the random-dots estimation task. In this task, participants encounter images depicting randomly generated non-overlapping dots and are asked to provide a judgment on the number of presented dots. To manipulate expertise, participants were randomly assigned to either learn raster scanning, a technique to overlay the presented image with a 3×3 raster and count only the dots in one of the resulting areas before multiplying the result by nine, or to an control condition in which participants read an essay about the importance of accurate judgments. Afterwards, participants decided to adjust or maintain a presented judgment for the number of dots depicted in a presented image. Again, the presented judgment was framed as the judgment of a previous participant while the presented values were actually preselected to be either correct or deviate from the correct answer by +/-35% or +/-70% . The results were similar as in the previous experiment such that larger deviations of the presented judgment from the correct answer resulted in more frequent adjustments and more improvement. Moreover, participants assigned to the expertise-manipulation condition adjusted presented judgments more frequently and improved them more than participants in the control condition. Lastly, both presented deviation and condition interacted such that experts showed stronger effects of presented deviation on both change probability and improvement while they again also adjusted already correct judgments more frequently than novices.

Both experiments described above only examined the role of expertise in one step of sequential collaboration. Nonetheless, positive effects of expertise should not only be apparent in single adjustments but also in estimates obtained with sequential chains. Thus, we performed a third study again manipulating expertise using the random-dots estimation task but additionally assigned participants to sequential chains of two resulting in four different chain compositions, namely novice-novice, expert-novice, novice-expert, and expert-expert. While the first participant in each sequential chain again encountered preselected judgments with each image of random dots, the second participant was presented with updated judgments according to the adjustments of the first participant. On the level of one sequential step as in Experiment 1 and 2, we found similar results as in the two previous experiments even though there was no effect of expertise on change probability. On the level of sequential chains, experts adjusted judgments of novices more frequently than judgments of experts while novices adjusted judgments of previous experts and novices similar frequently. Moreover, judgments improved most

when experts corrected novices while novices worsened judgments of experts. Lastly, sequential chains yielded more accurate judgments the more experts they contained and the later these experts entered the sequential chain.

These findings shed a first light into the mechanism behind sequential collaboration showing that contributors weigh their judgments higher, the higher their expertise is which results in more frequent adjustments and more improvement of presented judgments. Ultimately, this has a positive effect on estimates obtained with sequential collaboration. As contributors organize the weighing of judgments by expertise themselves implicitly by opting in or out of providing a judgment, it is not necessary to explicitly assign expert roles or derive individuals' expertise from task performance and weigh judgments statistically.

3.3 Towards a Theory of Sequential Collaboration

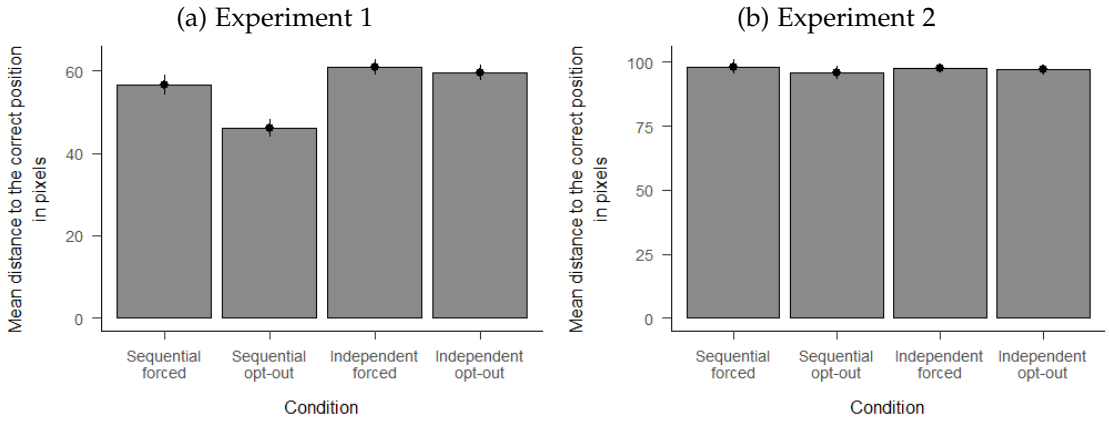
Even though sequential collaboration has yielded promising results in the studies discussed above, the mechanism behind sequential collaboration and possible boundary conditions are still unclear. While Mayer, Broß, and Heck (2022) showed that contributors' expertise determines change probability and improvement of presented judgments in one sequential step, and ultimately also the accuracy of estimates obtained in sequential chains, there are probably other constructs affecting whether and how accurate contributors adjust entries in sequential collaboration. One of these constructs closely related to expertise may be judgment confidence. Even though individuals' domain-specific expertise can be one source of judgment confidence (Zakay & Tuvia, 1998) if individuals do not hold erroneous beliefs (Koriat, 2008, 2011), judgment confidence is also affected by miscalibrated metacognitive knowledge about one's own expertise (Kruger & Dunning, 1999) or item-specific knowledge that enables individuals to answer a single item without having higher domain-specific expertise. However, both miscalibrated metacognitive knowledge about one's own expertise and item-specific knowledge may affect whether individual decide to adjust entries in sequential collaboration and whether their adjustments improve or worsen previous judgments. Thereby, judgment confidence may have two facets as contributors can make up to two decisions in the process of sequential collaboration. First, they decide whether to adjust or maintain a presented and also provide a new judgment if they decide to adjust the presented one. Thus, contributors are likely to hold a confidence for both of these judgments. Even though both of these judgment confidences are supposedly positively related as more competence contributors are likely to also be more confident, this might not always be the case. If a contributor with average expertise in European geography encounters a judgment of Rome positioned closely to Milan 400 kilometers north of the correct

position of Rome, they might be very confident that this judgment is incorrect. However, this does not mean that they know the exact location of Rome and can position the city very accurately resulting in only little to medium confidence in the judgment they provide. To gain a more comprehensive image of prerequisites for adjustments in sequential chains, future research should integrate expertise and judgment confidence into a theoretical framework of sequential collaboration.

Another important but still open question is whether providing judgments sequentially is beneficial over and above the possibility to opt out of providing a judgment in sequential collaboration. Research on unweighted averaging has demonstrated that allowing participants to select the questions to answer rather than to assign them to these questions improves resulting estimates as participants can provide judgments according to their metacognitive assessments (Bennett & Steyvers, 2022; Bennett et al., 2018). However, some of these questions have only been selected rarely or not even once which can make opting out of independent judgments inefficient even though resulting judgments and estimates are highly accurate. Similarly, contributors in sequential collaboration also adjust presented entries according to their metacognitive knowledge about their own expertise (Mayer, Broß, & Heck, 2022). This procedure is more efficient than allowing to opt out of independent individual judgments as sequential chains start off with a judgment and thus at least one judgment is provided for each question in a sequential chain. While opting out of providing a judgment is an important mechanism for sequential collaboration, it is not clear whether providing judgments sequentially also facilitates judgment accuracy. Thus, the effects of both of these features on sequential collaboration should be disentangled. To this end, we conducted a study using the city-location task already established in Mayer and Heck's (2022b) Experiment 3. In addition to requiring participants to provide independent individual judgments and the typical sequential-collaboration task allowing to opt out of providing a judgment, we also included a condition in which participants decided whether to provide an independent judgment or to opt out, and a condition in which participants were required to adjust the presented judgment of a previous participant. Thereby, we fully crossed judgment aggregation, either unweighted averaging of independent individual judgments or sequential collaboration, and whether participants had the possibility to opt out of provide a judgment. Figure 5 depicts preliminary results of two studies comparing the accuracy of estimates obtained with all four modes of judgment aggregation. In Experiment 1 (Panel A) we were able to replicate the result of Mayer and Heck (2022b) concerning estimate accuracy of sequential collaboration and unweighted averaging. Moreover, the findings hint towards the sequential process itself has some benefits over and above opting out as estimates are descriptively more accurate if participants provide judgments sequentially rather than independently even

though they have to answer all questions in both cases. However, this difference was not significant. Unfortunately, participants who had the possibility to opt out of providing an independent judgment mostly refrained from it which lead to similar accurate estimates in both conditions providing independent judgments. Thus, these results do not allow to conclude that sequential collaboration with the possibility to opt out of providing judgments results in more accurate estimates than unweighted averaging when opting out of providing judgments was allowed since both condition vastly differ in how much participants actually opted out of providing a judgment. In an attempt to increase opt-out rate for participants providing independent individual judgments, we increased item difficulty such that participants may more frequently refrain from providing an independent judgment. However, this lead to no differences in the accuracy of estimates between all conditions (Figure 5B). Thus a paradigm needs to be developed which allows to adequately disentangle the effects of providing judgments sequentially and opting out of providing a judgment to shed further light into the mechanism behind sequential collaboration.

FIGURE 5: Comparison of estimates obtained with unweighted averaging of independent judgments or sequential collaboration each for allowing participants to opt out and for requiring them to provide a judgment.



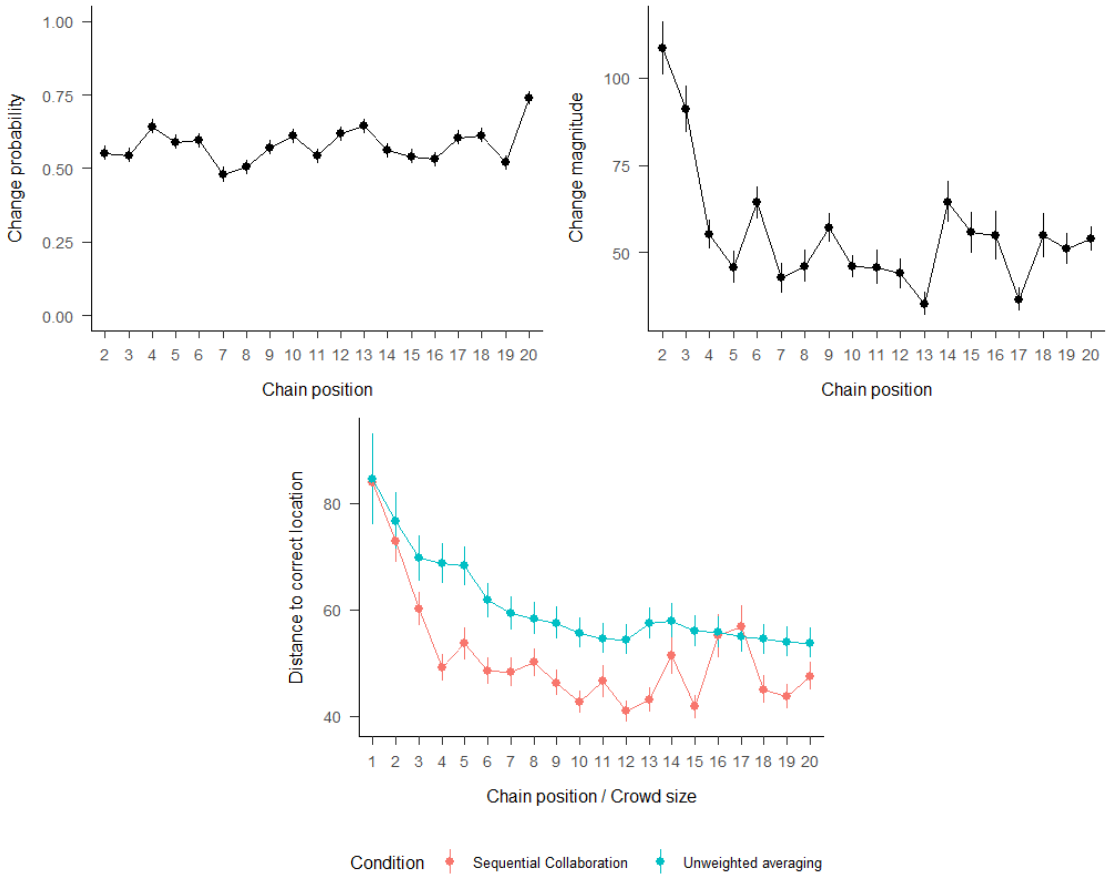
Moreover, all studies that investigated sequential collaboration so far only have built sequential chains of six participants at maximum. However, longer sequential chains may not be as beneficial as shorter ones for resulting estimates compared to unweighted averaging of independent individual judgments. While we found that change probability and change magnitude decrease and judgment accuracy increases for short sequential chains (Mayer & Heck, 2022b), the results of Mayer, Broß, and Heck (2022) however indicate that both experts and novices adjust and thereby worsen already correct and highly accurate judgments. Thus, it is more likely that judgments in long sequential chains can only reach a certain level of accuracy since correct and highly accurate

judgments are nonetheless regularly worsened before being corrected again. Preliminary results of a first study comparing estimate accuracy of long sequential of twenty contributors to equally large groups of participants providing independent individual judgments are displayed in Figure 6. While the upper left panel illustrates that change probability remained similar over the whole course of a sequential chain, change magnitude as displayed in the upper right panel decreased over the first three sequential steps before approaching a mean change magnitude of around 50 pixels in a sequential step. Lastly, the lower panel of Figure 6 shows the accuracy of estimates measured as distance of estimates to the correct location obtained with both sequential collaboration and unweighted averaging. For sequential collaboration, the estimate is the latest judgment in a sequential chain for an item. For unweighted averaging, the cumulative mean of location judgments for participants whose data was randomly grouped into crowds of twenty was computed as estimates for each item. Accuracy of estimates increased for both aggregation methods for the first ten participants who either entered the sequential chain or whose data was combined in a crowd However, estimates did not become more accurate in the following and still have considerable error after aggregating judgments of twenty participants. Nonetheless, estimates obtained with sequential collaboration remained more accurate than estimates obtained with unweighted averaging. Future research should look more closely into conditions under which long sequential chains perform well in sequential collaboration compared to unweighted averaging. Moreover, it is important to investigate how to improve contributors' capacity to accurately assess their possibilities to improve presented judgments such that contributors' judgments converge to the correct judgment over a long sequential chain.

Investigating the open research questions outlined above can help to develop a theory of sequential collaboration indicating prerequisites and boundary conditions of successful judgment aggregation. This may also help to develop a cognitive model of sequential collaboration that sheds further light into the cognitive processes guiding judgments and collaboration.

While sequential collaboration can be a beneficial method for judgment aggregation, it is also used to share information in online collaborative projects. Thereby, collaboration often resembles a natural hidden profile setting as contributors may share some information on the topic, for instance where the Eiffel Tower is located when editing the Wikipedia article about it, but contributors also hold unshared information, for instance concerning its construction process. Even though research on group decision making revealed that groups often fail to share information that is not available to all group members (Stasser & Titus, 1985), sequential collaboration may be a promising alternative for information sharing. As contributors do not interact directly with one another but rather are only presented with entries of previous contributors, informa-

FIGURE 6: Change probability, change magnitude, and estimate accuracy obtained with sequential collaboration and unweighted averaging for groups of up to 20 participants.



Note. Points and vertical lines show empirical means and corresponding 95% confidence intervals. Chain position 1 was omitted for both change probability and change magnitude since these participants could not opt out of answering. For unweighted averaging, cumulative means of the provided positions were computed to obtain estimates.

tion sharing in sequential collaboration may be much less affected by group biases or individual effects undermining information sharing in groups and may facilitate contributing according to one's expertise. Thus, future research should investigate whether information sharing is indeed facilitated by sequential collaboration compared to discussion in groups. Furthermore, if sequential collaboration proves to be more successful in information sharing than group discussion, a next step could be to examine whether little interaction among contributors in this process or the sequential design of providing judgments drives improved information sharing.

All propositions for future research described above call for experimental studies. However, sequential collaboration can also be examined by analyzing edits and entries

in large-scale online collaborative projects. While edits in Wikipedia need to be analyzed using text processing as Wikipedia articles consist of full-text paragraphs, OpenStreetMap offers both numeric spacial information such as longitude and latitude of objects, their length, and area they cover as well as thematic information organized in tags for each object. Thus, entries and changes in OpenStreetMap may be more easy to extract and process for further statistical analyses (Mayer, Heck, & Mocnik, 2022) and are suitable to test theories of how contributors adjust and maintain entries on projects using sequential collaboration everyday.

4 Conclusion

Research on collaboration has in the past mostly focused on aggregating independent individual judgments on the one hand and group decision making on the other hand. While unweighted averaging of independent individual judgments was shown to yield highly accurate estimates, groups often fail to integrate their individual information and hold onto their initial judgments resulting in biased group decisions. Nonetheless, both methods profit from considering expert judgment when aggregating individual judgments into either a statistical estimate or a group decision.

Going beyond previous approaches, we extended Cultural Consensus Theory, a model allowing to consider informants' expertise and item difficulty when aggregating individual judgments, to two-dimensional continuous data (Mayer & Heck, 2022a). This model extension allows to weigh geographical location judgments by informants' competence which resulted in more accurate location estimates than unweighted averaging for both simulated and empirical data.

However, apart from unweighted averaging of independent individual judgments and group decision making, with large-scale online collaborative projects such as Wikipedia and OpenStreetMap an asynchronous, sequential way of collaborating emerged within the last two decades. In sequential collaboration, contributors neither provide independent judgments nor a consensus group decision. Instead, after an entry is created by a first contributor independently, subsequent contributors who encounter this entry decide whether to adjust and maintain it sequentially. We assumed that this collaboration process is especially suitable to obtain highly accurate estimates as it allows contributors to opt out of providing a judgment enabling them to weigh judgments according to their expertise. A systematic investigation of judgment aggregation in sequential collaboration over three studies revealed that while adjustments become less frequent and smaller over the course of a sequential chain, judgments become increasingly accurate and estimates at the end of sequential chains are similarly accurate as estimates obtained with unweighted averaging (Mayer & Heck, 2022b). Moreover, adjusting and maintaining judgments was shown to heavily depend on the accuracy of the previous judgment as well as contributors expertise which ultimately affects the accuracy of estimates obtained with sequential collaboration (Mayer, Broß, & Heck, 2022). These findings support the notion that sequential collaboration allows contributors to implicitly weigh their judgments by expertise through opting in or out of providing a

judgments according to one's expertise.

Even though these results suggest that sequential collaboration is a successful way of aggregating individual judgments, these experiments are only a first step towards a theoretical framework of sequential collaboration. Considering judgment confidence, distinguishing the sequential providing of judgments from the possibility to opt out of providing a judgment, and the development of judgments in long sequential chains are further steps towards a better understanding of sequential collaboration. Findings on these open questions may help to develop a theory of judgment aggregation in sequential collaboration and to determine prerequisites as well as boundary conditions for highly accurate judgments and estimates. Moreover, sequential collaboration is not limited to being a method of judgment aggregation but may also be successful in supporting information sharing as detrimental group processes and individual biases may be diminished compared to group decision making.

To conclude, the present thesis further supports the notion that weighing individual judgments by expertise is beneficial for the resulting estimates. Moreover, I demonstrate that sequential collaboration is a successful method to aggregate individual judgments as it allows contributors to weigh their judgments by expertise. Thereby, I extend the scope of wisdom of crowds to dependent incremental judgments and shed light into a collaboration mechanism which is used every day in large-scale online collaborative projects like Wikipedia and OpenStreetMap.

5 Bibliography

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452–469. https://doi.org/10.1016/j.jmp.2013.01.004

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*, 151–181. https://doi.org/10.1007/s11336-013-9382-9

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13. https://doi.org/10.1016/j.jmp.2014.06.001

Arazy, O., Morgan, W., & Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in wikipedia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1025624

Aßfalg, A. (2018). Consensus theory for mixed response formats. *Journal of Mathematical Psychology*, *86*, 51–63. https://doi.org/10.1016/j.jmp.2018.08.005

Baeza-Yates, R., & Saez-Trumper, D. (2015). Wisdom of the crowd or wisdom of a few? an analysis of users' content generation. *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 69–74. https://doi.org/10.1145/2700171.2791056

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332. https://doi.org/10.1016/j.jmp.2012.06.002

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92. https://doi.org/10.1007/BF02294195

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

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*, E5070–E5076. https://doi.org/10.1073/pnas.1615978114

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

Bennett, S. T., & Steyvers, M. (2022). Leveraging metacognitive ability to improve crowd accuracy via impossible questions. *Decision*, *9*, 60–73. https://doi.org/10.1037/dec0000165

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

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., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes*, *103*, 121–133. https://doi.org/10.1016/j.obhdp.2006.05.001

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. *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, 337–340. http://mural.maynoothuniversity.ie/2476/

Clemen, T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*, 79–101. https://doi.org/10.1037/dec0000004

Dennis, A. R. (1996). Information exchange and use in small group decision making. *Small Group Research*, *27*, 532–550. https://doi.org/10.1177/1046496496274003

Dennis, A. R., Hilmer, K. M., & Taylor, N. J. (1998). Information exchange and use in GSS and verbal group decision making: Effects of minority influence. *Journal of Management Information Systems: JMIS; Armonk*, *14*(3), 61–88. https://doi.org/10.1080/07421222.1997.11518175

de Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, *115*, 2066–2071. https://doi.org/10.1073/pnas.1717632115

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

Faulmüller, N., Kerschreiter, R., Mojzisch, A., & Schulz-Hardt, S. (2010). Beyond group-level explanations for the failure of groups to solve hidden profiles: The individual preference effect revisited. *Group Processes & Intergroup Relations*, *13*, 653–671. https://doi.org/10.1177/1368430210369143

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

Friedman, A., Kerkman, D. D., & Brown, N. R. (2002). Spatial location judgments: A cross-national comparison of estimation bias in subjective North American geography. *Psychonomic Bulletin & Review*, *9*, 615–623. https://doi.org/10.3758/BF03196321

Friedman, A., Kerkman, D. D., Brown, N. R., Stea, D., & Cappello, H. M. (2005). Cross-cultural similarities and differences in North Americans' geographic location judgments. *Psychonomic Bulletin & Review*, *12*, 1054–1060. https://doi.org/10.3758/BF03206443

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451. https://doi.org/10.1038/075450a0

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, *65*, 959–974. https://doi.org/10.1037/0022-3514.65.5.959

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*, 149–167. https://doi.org/10.1037/0033-2909.121.1.149

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

Greitemeyer, T., & Schulz-Hardt, S. (2003). Preference-consistent evaluation of information in the hidden profile paradigm: Beyond group-level explanations for the dominance of shared information in group decisions. *Journal of Personality and Social Psychology*, *84*, 322–339. https://doi.org/10.1037/0022-3514.84.2.322

Grüning, D. J., & Krueger, J. (2021). *Vox peritorum: Capitalizing on confidence and projection to characterize expertise*. https://doi.org/10.31234/osf.io/6vndh

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

Henningsen, D. D., & Henningsen, M. L. M. (2004). The effect of individual difference variables on information sharing in decision-making groups. *Human Communication Research*, *30*, 540–555. https://doi.org/10.1111/j.1468-2958.2004.tb00744.x

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40–46. https://doi.org/10.1016/0030-5073(78)90037-5

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.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*, 756–763. https://doi.org/10.1038/s41562-021-01057-0

Kerr, D. S., & Murthy, U. S. (2009). The effectiveness of synchronous computer-mediated communication for solving hidden-profile problems: Further empirical evidence. *Information & Management*, *46*, 83–89. https://doi.org/10.1016/j.im.2008.12.002

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*, 78–94. https://doi.org/10.1037/0022-3514.44.1.78

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

King, A. J., Cheng, L., Starke, S. D., & Myatt, J. P. (2012). Is the true 'wisdom of the crowd' to copy successful individuals? *Biology Letters*, *8*, 197–200. https://doi.org/10.1098/rsbl.2011.0795

Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37–46. https://doi.org/10.1145/1460563.1460572

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453–462. https://doi.org/10.1145/1240624.1240698

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*, 117–139. https://doi.org/10.1037/a0022171

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*, 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. https://doi.org/10.1037/0022-3514.77.6.1121

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Social judgment and decision making* (1 ed., pp. 227–242). Taylor & Francis. https://doi.org/10.4324/9780203854150-23

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., Bonner, B. L., Miner, A. G., & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80*, 103–117. https://doi.org/10.1006/obhd.1999.2848

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*, 322–333. http://sjdm.org/~baron/journal/18/18331/jdm18331.pdf

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 : JAMIA*, *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

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*, 9020–9025. https://doi.org/10.1073/pnas.1008636108

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, *16*, 54–75. https://doi.org/10.1177/1088868311417243

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*, 276–299. https://doi.org/10.1037/a0036677

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., Broß, M., & Heck, D. W. (2022). *Expertise determines the frequency and accuracy of contributions in sequential collaboration*. https://doi.org/10.31234/osf.io/s7vtg

Mayer, M., & Heck, D. W. (2022a). *Cultural consensus theory for two-dimensional data: Expertise-weighted aggregation of location judgments*. https://doi.org/10.31234/osf.io/unhvc

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

Mayer, M., Heck, D. W., & Mocnik, F.-B. (2022). *Using OpenStreetMap as a data source in psychology and the social sciences*. https://doi.org/10.31234/osf.io/h3npa

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. *Social Computing, Behavioral-Cultural Modeling and Prediction*, 236–243. https://doi.org/10.1007/978-3-642-19656-0_34

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, *94*, 535–546. https://doi.org/10.1037/a0013773

Meyen, S., Sigg, D. M. B., Luxburg, U. v., & Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. *Cognitive Research: Principles and Implications*, *6*, 18. https://doi.org/10.1186/s41235-021-00279-0

Miller, B. J., & Steyvers, M. (2011). The wisdom of crowds with communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*. https://escholarship.org/uc/item/4jt6q62c

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, *64*, 4177–4192. https://doi.org/10.1287/mnsc.2017.2823

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. In R. F. Pohl (Ed.), *Cognitive illusions* (1st, pp. 183–199). Psychology Press.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 126–132. https://doi.org/10.1038/s41562-017-0273-4

Niederer, S., & van Dijck, J. (2010). Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New Media & Society*, *12*, 1368–1387. https://doi.org/10.1177/1461444810365297

OpenStreetMap Contributors. (2021, January 21). *DE:statistik – OpenStreetMap wiki*. https://wiki.openstreetmap.org/wiki/DE:Statistik

OpenStreetMap Contributors. (2022). *OpenStreetMap*. https://www.openstreetmap.org/about

Palley, A., & Satopää, V. (2022). *Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions* (SSRN Scholarly Paper No. ID 3504286). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3504286

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using Gibbs sampling.

Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *Journal of personality and social psychology*, *80*, 918–930. https://doi.org/10.1037//0022-3514.80.6.918

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*, 532–535. https://doi.org/10.1038/nature21054

Quinn, S., & Bull, F. (2019). Understanding threats to crowdsourced geographic data quality through a study of OpenStreetMap contributor bans. In N. A. Valcik (Ed.), *Geospatial information system use in public organizations - how and why GIS should be used by the public sector* (1., pp. 80–96). Routledge. https://doi.org/10.4324/9780429272851-6

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338. https://www.jstor.org/stable/677564

Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2015). Internet [https://ourworldindata.org/internet]. *Our World in Data*.

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

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*, 1–28. https://doi.org/10.1016/0749-5978(89)90055-1

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes*, *45*, 66–84. https://doi.org/10.1016/0749-5978(90)90005-T

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Steinwehr, U., & Bushuev, M. (2021). Faktencheck: Wie verlässlich ist wikipedia? *Deutsche Welle*. https://www.dw.com/de/faktencheck-wie-verl%C3%A4sslich-ist-wikipedia/a-56212126

Stewart, D. D., & Stasser, G. (1995). Expert role assignment and information sampling during collective recall and decision making. *Journal of Personality and Social Psychology*, *69*, 619–628. https://doi.org/10.1037/0022-3514.69.4.619

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. (2009). The wisdom of crowds in the recollection of order information. *Advances in Neural Information Processing Systems*, *22*. https://proceedings.neurips.cc/paper/2009/hash/4c27cea8526af8cfee3be5e183ac9605-Abstract.html

Surowiecki, J. (2005). *The wisdom of crowds* (1. ed). Anchor Books.

The Signpost. (2021). *Ban on IPs on ptwiki, paid editing for tatarstan, IP masking*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_Signpost/2020-11-01/News_and_notes&oldid=1012005073

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wagner, C., & Vinaimont, T. (2010). Evaluating the wisdom of crowds. *Issues in Information Systems*, *11*, 724–732. https://iacis.org/iis/2010/724-732_LV2010_1546.pdf

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

Waubert de Puiseau, B., Greving, S., Aßfalg, A., & Musch, J. (2017). On the importance of considering heterogeneity in witnesses' competence levels when reconstructing crimes from multiple witness testimonies. *Psychological Research*, *81*, 947–960. https://doi.org/10.1007/s00426-016-0802-1

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field Methods*, *19*, 339–368. https://doi.org/10.1177/1525822X07303502

Wikipedia Contributors. (2022a). *Wikimedia statistics - all wikipedias*. https://stats.wikimedia.org/#/all-wikipedia-projects

Wikipedia Contributors. (2022b). *Wikipedia:about*. https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1091417331

Wittenbaum, G. M., Hubbell, A. P., & Zuckerman, C. (1999). Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of Personality and Social Psychology*, *77*, 967–978. https://doi.org/10.1037/0022-3514.77.5.967

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281. https://doi.org/10.1006/obhd.2000.2909

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*, 452–470. https://doi.org/10.1111/j.1551-6709.2011.01223.x

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, *98*, 103–115. https://doi.org/10.1016/S0001-6918(97)00037-1

Zielstra, D., & Zipf, A. (2010). Quantitative studies on the data quality of OpenStreetMap in germany. *AGILE 2010. The 13th AGILE International Conference on Geographic Information Science*. https://www.giscience2010.org/pdfs/paper_187.pdf

A Acknowledgements

"I don't know half of you half as well as I should like, and I like less than half of you half as well as you deserve."

from *The Lord of the Rings: The Fellowship of the Ring*

Unlike I planned and expected when starting my dissertation in 2019, I spent more time during the last three years working from home than I ever imagined and got to know my advisors and colleagues less than I had hoped for. Nonetheless, everybody was always happy to meet digitally and thereby numerous people supported me even though we were not able to meet personally.

First and foremost, I am grateful to Daniel for encouraging me to start this dissertation in the first place. From him, I learned the importance of precise theories, methodological rigor, and clear writing. He gave me the freedom to pursue my ideas and provided feedback and guidance whenever I needed them.

I would also like to thank Edgar, Arndt, and Benni who supported me with helpful feedback, new perspectives, and interesting insights along the way. Moreover, I am grateful to Thorsten who helped out when I desparately needed a third thesis reviewer.

Many thanks to Marcel, Julian, and Barbara for their valuable advice and feedback, many interesting discussions, and recently also lunch breaks that sometimes lasted longer than expected. I would also like to thank Tobias, Gloria, Annika, and all of SMiP for all the interesting and stimulating workshops, retreats, and informal meetings we spent together.

Particular thanks to my husband, Christian, who celebrated my successes with me and always encouraged in me whenever I did not believe in myself. He saw me in my best and my worst times of this dissertation projects and I am deeply indebted to him for his support.

Lastly, I would like to thank all my family and friends for their encouragement and moral support over the past three years and especially in the last few months. Thank you all.

Maren Mayer
Mannheim, June 2022

B Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title *Expertise-Weighing of Judgments in Wisdom of Crowds: Investigating Independent Judgments and Sequential Collaboration* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not yet present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby confirm the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date:

C Co-Authors' Statements

Co-Author: Daniel W. Heck

With this statement, I confirm that the following articles included in the presented thesis were primarily conceived and written by Maren Mayer.

Mayer, M., & Heck, D. W. (2022). Cultural Consensus Theory for Two-Dimensional Data: Expertise-Weighted Aggregation of Geographic Location Judgments. Manuscript under review. http://doi.org/10.31234/osf.io/unhvc

Mayer, M., & Heck, D. W. (2022). *Sequential Collaboration: The Accuracy of Dependent, Incremental Judgments. Decision*. Advanced online publication. https://doi.org/10.1037/dec0000193

Mayer, M., Broß, M., & Heck, D. W. (2022). *Expertise Determines Frequency and Accuracy of Contributions in Sequential Collaboration.* Manuscript under review. http://doi.org/10.31234/osf.io/s7vtg

Maren Mayer developed the theoretical background, conceptualized the experimental designs for all studies reported in the manuscripts, performed the simulations and data analysis, wrote the first drafts, and revised them. I contributed to the theoretical background and the statistical analyses and revised the manuscripts. Moreover, I co-developed the statistical model described in (Mayer & Heck, 2022a) and provided recommendations for the simulation settings.

Prof. Dr. Daniel W. Heck
Marburg, June 2022

Co-Author: Marcel Broß

With this statement, I confirm that the following articles included in the presented thesis were primarily conceived and written by Maren Mayer.

Mayer, M., Broß, M., & Heck, D. W. (2022). *Expertise Determines Frequency and Accuracy of Contributions in Sequential Collaboration.* Manuscript under review. http://doi.org/10.31234/osf.io/s7vtg

Maren Mayer designed and conducted two of three experiments described in this manuscript. She performed all data analyses reported and wrote the first draft of the article as well as revisions of the manuscript. I contributed to one experiment by developing the study design and investigating the hypotheses, and revised the manuscript.

Marcel Broß
Marburg, June 2022

D Copies of Articles

Cultural Consensus Theory for Two-Dimensional Data: Expertise-Weighted Aggregation of Location Judgments

Maren Mayer[1,2] & Daniel W. Heck[3]

[1] University of Mannheim

[2] Heidelberg Academy of Sciences and Humanities

[3] University of Marburg

Author Note

Maren Mayer, Department of Psychology, School of Social Sciences, University of Mannheim, Germany. https://orcid.org/0000-0002-6830-7768

Daniel W. Heck, Department of Psychology, University of Marburg, Germany. https://orcid.org/0000-0002-6302-9252

Data and R scripts for the analyses are available at the Open Science Framework (https://osf.io/jbzk7/).

The present work was presented at the SJDM Annual Meeting 2020 (Virtual Conference) and at the 15th Conference of the Section 'Methods and Evaluation' (2021) of the German Psychological Society (DGPs). The present manuscript has not yet been peer reviewed. A preprint was uploaded to PsyArXiv and ResearchGate for timely dissemination (version: May 7, 2022).

This work was funded by the WIN programme of the Heidelberg Academy of Sciences and Humanities, financed by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg and also supported by the Research Training Group "Statistical Modeling in Psychology" funded by the German Research Foundation (DFG grant GRK 2277).

The authors made the following contributions. Maren Mayer: Conceptualization, Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing; Daniel W. Heck: Conceptualization, Methodology, Writing - Review & Editing.

Abstract

Cultural consensus theory is a model-based approach for analyzing responses of informants when correct answers are unknown. The model provides aggregate estimates of the latent consensus knowledge at the group level while accounting for heterogeneity both with respect to informants' competence and items' difficulty. We develop a specific version of cultural consensus theory for two-dimensional continuous judgments as obtained when asking informants to locate a set of unknown sites on a geographic map. The new model is fitted using hierarchical Bayesian modeling, with a simulation study indicating satisfactory parameter recovery. We also assess the accuracy of the aggregate location estimates by comparing the new model against simply computing the unweighted average of the informant's judgments. A simulation study shows that, due to weighting judgments by the inferred competence of the informants, cultural consensus theory provides more accurate location estimates than unweighted averaging. This result is also supported in an empirical study in which individuals judged the location of European cities on maps.

Keywords: wisdom of crowds, group decision making, Bayesian modeling, test theory, psychometrics

Cultural Consensus Theory for Two-Dimensional Data: Expertise-Weighted Aggregation of Location Judgments

1 Introduction

In many domains in the social sciences and particularly in psychological research, participants often provide responses to questions for which correct answers are not known. For instance, researchers may ask whether one agrees or disagrees with a set of statements about a certain topic such as beliefs about AIDS (Trotter et al., 1999). Cultural consensus theory (CCT, Romney et al., 1986) is a method for analyzing responses from several informants when correct answers are unknown. The model infers the latent cultural consensus of a group while considering variance both in the competence of informants and in the difficulty of items. Hence, CCT has also been described as "test theory without an answer key" (Batchelder & Romney, 1988).

The fact that true answers are unknown complicates the aggregation of informants' responses because it is not clear which of the informants are most competent in the sense that they provide judgments close to the unknown cultural truth. As a remedy, CCT allows researchers to identify the latent cultural truth while simultaneously estimating the cultural competence of each informant. The main principle of CCT is that informants with more cultural knowledge, and thus, higher competence regarding the latent consensus, are likely to show similar answer patterns across the set of questions asked (Romney et al., 1986). Based on the correlation of answer patterns, the method jointly estimates the cultural truth at the group level and the informants' competence at the individual level. This requires that multiple informants provide judgments to a set of items from the same knowledge domain (Weller, 2007).

1.1 Applications and Extensions of Cultural Consensus Theory

CCT was first developed in anthropological research for questionnaires about cultural topics with a dichotomous response format (Batchelder & Romney, 1988; Romney et al., 1986). For instance, one of the first applications investigated the

intracultural variability of beliefs about whether illnesses are contagious (Romney et al., 1986). The method has since been applied in various contexts such as aggregating eyewitness reports (Waubert de Puiseau et al., 2017; Waubert de Puiseau et al., 2012), obtaining forecasts for various events (Anders et al., 2014; Merkle et al., 2020), or estimating social networks where individuals provide information about social relations among different people (Batchelder et al., 1997; Batchelder, 2009).

The original version of CCT was applicable only to dichotomous data with one latent cultural truth to which all informants belong. As it may be possible that not all informants share a single, common consensus, Anders and Batchelder (2012) extended CCT to multiple cultural truths (see also Aßfalg & Klauer, 2020). Essentially, such extended models assume that informants belong to separate latent classes which differ with respect to the assumed cultural truth. For instance, medical professionals and lay people may differ with respect to medical beliefs resulting in different latent cultural truth if the group membership is not known.

CCT was also extended to other response formats than binary answers. Extensions have been developed for continuous data (Anders et al., 2014; Batchelder & Anders, 2012), ordinal responses (Anders & Batchelder, 2015), and mixed response formats (Aßfalg, 2018), and have been used to aggregate ratings about the grammatical acceptability of English phrases as well as judgments about the importance of various health behaviors. Statistical inference for such extended CCT models has often relied on hierarchical Bayesian modeling in which parameter estimates are obtained via Markov chain Monte Carlo sampling (Anders et al., 2014; Anders & Batchelder, 2012; Aßfalg & Klauer, 2020). Overall, all these extensions have enabled researches to adapt the CCT approach to various types of data while assuming a certain structure of cultural truths underlying informants' answers.

CCT is also applicable to scenarios in which correct answers are not known during the time of data collection, but may become available later. Such applications are especially interesting because the performance of different aggregation methods can be directly compared against each other. In fact, prior research in judgment and

decision making showed that aggregating independent individual judgments with an unweighted average of all judgments results in highly accurate group estimates for various tasks and contexts (Hueffer et al., 2013; Larrick & Soll, 2006; Steyvers et al., 2009; Surowiecki, 2005). This is surprising because all judgments are weighted equally without considering or estimating informants' competence with respect to the corresponding domain. In contrast, the aggregation of judgments in CCT is weighted by the estimated competence of informants, thereby assigning more weight to informants closer to the cultural truth. Merkle et al. (2020) recently showed that a CCT-inspired aggregation mechanism indeed outperforms unweighted averaging. Similarly, the accuracy of aggregated eyewitness testimonies increases when accounting for the witnesses' competence levels (Waubert de Puiseau et al., 2017). This illustrates that CCT is a useful tool for aggregating judgments when the ground truth becomes available only at a later time.

While CCT has been adapted to several types of response formats and applications, an extension to two- or higher-dimensional continuous judgments has not been developed yet. Such an extension is especially useful for the aggregation of geographical judgments about the unknown location of several sites on a map. Possible applications for such an extension are, for instance, two-dimensional location judgments in research on geographic knowledge and representation (Friedman, Brown, et al., 2002; Friedman et al., 2012, 2005; Friedman, Kerkman, et al., 2002; Thorndyke & Hayes-Roth, 1982), location judgments for objects hidden by obstacles (Yarbrough et al., 2002), or the search of optimal locations for public facilities (e.g., park-and-ride facilities, Faghri et al., 2002). Especially when comparing the geographical knowledge of different cultural groups with respect to location judgments on maps (Friedman, Brown, et al., 2002; Friedman et al., 2005), a two-dimensional extension of CCT allows researchers to aggregate individual judgments while identifying individuals' competence as a possible source of variance in judgments. Furthermore, a two-dimensional extension of CCT may be useful for locating unknown sites based on expert judgments in scenarios such as finding a lost submarine (Surowiecki, 2005), ancient archaeological

sites (Casana, 2014), natural resources (e.g., water harvesting sites, Al-shabeeb, 2016), or suitable areas for ecotourism (Mahdavi et al., 2015).

In the following, we thus extend CCT to two-dimensional location judgments based on Anders' (2014) CCT model for one-dimensional continuous responses. We check the validity and performance of the proposed CCT model and its Bayesian implementation in JAGS (Plummer, 2003) by investigating parameter convergence and recovery in a Monte Carlo simulation. Moreover, we use simulations to examine under which conditions CCT's weighting of judgments by individuals' competence improves the accuracy of location estimates at the group level. Empirically, we apply the new model to reanalyze location judgments of European cities on maps (Mayer & Heck, 2021) and compare the accuracy of the aggregate location estimates to those obtained with unweighted averaging. Overall, the results of our simulation studies and the empirical reanalysis show that CCT's weighting of individual location judgments by informants' competence improves the estimation accuracy compared to weighting all judgments equally.

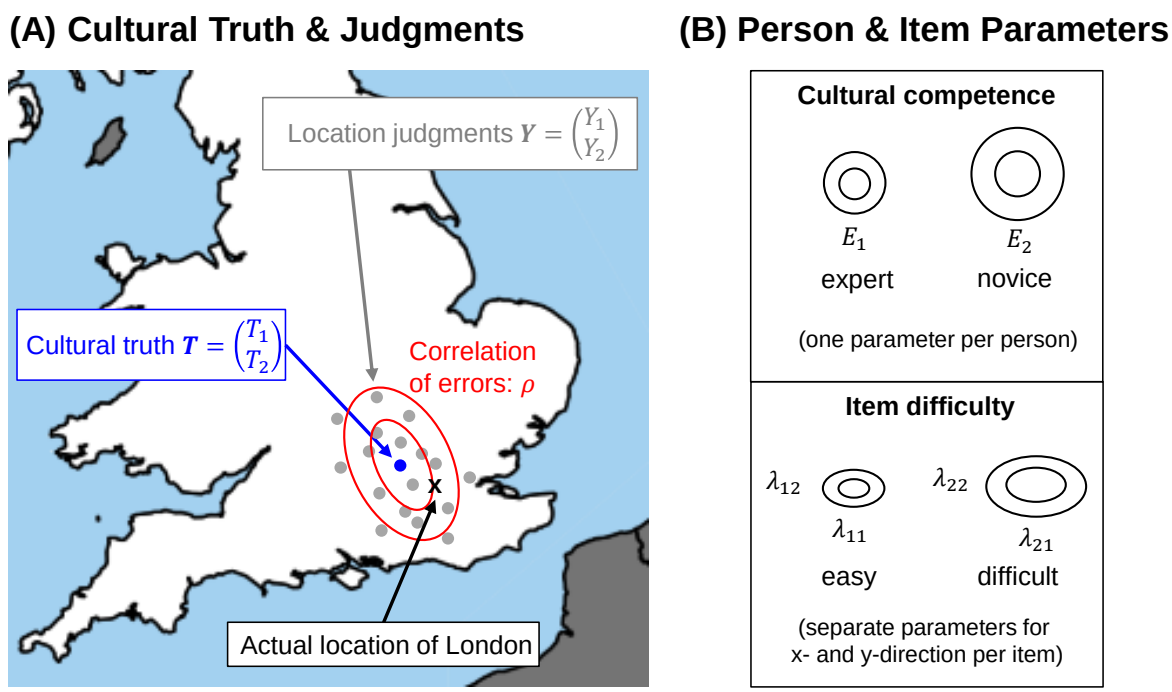
2 Model extension for two-dimensional continuous responses

2.1 Data structure

We extend the CCT model for one-dimensional continuous responses by Anders et al. (2014) to two-dimensional continuous judgments. As in all CCT models, the model requires that multiple informants provide judgments for a set of items from the same competence domain (Weller, 2007). For instance, as illustrated in Figure 1A, several informants could be asked to locate different European cities such as London on geographic maps (Mayer & Heck, 2021). Locations can be measured in different units depending on the application. For instance, one may use pixels of the presented image as in our empirical study below or geographical coordinates such as longitude and latitude, but other two-dimensional judgments are also feasible.

Figure 1

Data structure and CCT parameters for location judgments of London.



Regarding the notation, we assume that $i = 1, \ldots, N$ informants answer $k = 1, \ldots, M$ items by providing continuous, two-dimensional location judgments

$$\boldsymbol{Y}_{ik} = \begin{pmatrix} Y_{ik1} \\ Y_{ik2} \end{pmatrix}. \tag{1}$$

This means that each location judgment contains two components with Y_{ik1} referring to the first dimension (e.g., the x-axis or longitude on a map) and Y_{ik2} referring to the second dimension (e.g., the y-axis or latitude).

2.2 Model specification

The CCT model for two-dimensional judgments (CCT-2D) assumes that all respondents share a single latent cultural truth $\boldsymbol{T}_k$ for each item k. In our example, the latent-truth parameters refer to the group's consensus knowledge about the location of London and other European cities on a map. Note that our example concerns a case where the true locations are in principle available, but of course, the model also applies

to scenarios in which this is not the case.

As displayed in Figure 1A, we assume that the observed judgments $\boldsymbol{Y}_{ik}$ can be modeled by two additive components, the shared cultural truth and an unsystematic judgment error,

$$\boldsymbol{Y}_{ik} = \boldsymbol{T}_k + \boldsymbol{\varepsilon}_{ik}. \tag{2}$$

This additive structure of a true score and an error term is not only common for CCT models (Anders et al., 2014; Anders & Batchelder, 2012), but also at the core of classical test theory (Lord et al., 1968). Similar to other CCT models (Anders et al., 2014) and item response theory in general (Embretson & Reise, 2000), we assume that the errors $\boldsymbol{\varepsilon}_{ik}$ are conditionally independent given the person competence E_i and the item difficulty $\boldsymbol{\lambda}_k$. Moreover, since judgments are continuous, we assume a bivariate normal distribution of errors,

$$(\boldsymbol{\varepsilon}_{ik} \mid E_i, \boldsymbol{\lambda}_k) \overset{\text{iid}}{\sim} \text{MV-Normal}\left(\mathbf{0}, \boldsymbol{\Sigma}_{ik}\right). \tag{3}$$

The covariance matrix $\boldsymbol{\Sigma}_{ik}$ of judgment errors is modeled as a function of the informant's competence and the item's difficulty. The error variances in the x- and y-direction (i.e., the diagonal elements of $\boldsymbol{\Sigma}_{ik}$) are assumed to be smaller for persons with higher cultural competence and for items that are easier, meaning that in such cases the observed judgments are closer to the cultural truth. For instance, when asked to locate cities in the United Kingdom, informants with high competence will position these cities close to the shared cultural knowledge about the location. Formally, this idea is implemented by defining the person competence E_i and the item difficulty λ_{kd} as multiplicative factors which jointly determine the standard deviation of informants' judgments around the cultural truth in the d-th dimension,

$$\sigma_{ikd} = E_i \, \lambda_{kd} \tag{4}$$

Since cultural competence is modeled as a multiplicative factor affecting the standard deviation, the parameter E_i is restricted to be positive ($E_i > 0$). Figure 1B illustrates how the parameter E_i affects the variance of the distribution of errors. Essentially,

smaller values of E_i reflect a higher competence since judgments are closer to the cultural truth.

Recent versions of CCT (e.g., Anders et al., 2014) also assume that items vary in difficulty such that more difficult items result in a larger variance of judgments around the cultural truth. For the present case of location judgments, we define a vector-valued item-difficulty parameter $\boldsymbol{\lambda}_k$ for each item with two components $\lambda_{k1} > 0$ and $\lambda_{k2} > 0$ for the x- and y-dimension, respectively. We model the difficulty of each item with two instead of only one value because the x- and y-dimension may differ in difficulty.

2.3 Model assumptions specific to location judgments

Two-dimensional location judgments have some unique features which require special consideration in model development. Imagine that informants are asked to locate London, Birmingham, Glasgow, Liverpool, and Dublin on a map of the United Kingdom and Ireland similar to Figure 1A. The CCT-2D model outlined above accounts for such two-dimensional continuous responses by assuming that all informants answer according to the same underlying cultural truth. Here, the latent truth $\boldsymbol{T}_k$ refers to the group's shared knowledge about the positions of city k on the map. The model assumes that the location judgments of an informant are closer or further away from the shared consensus knowledge depending on their competence level. Importantly, the parameter E_i refers to the *general* competence of an informant irrespective of the x- or y-direction. Hence, when an informant knows that London is located in the south of the United Kingdom, it is also likely that they know whether it is located more to the west or to the east. This restriction simplifies the interpretation of the competency parameter E_i as a one-dimensional trait or construct.

Whereas competence is modeled as a one-dimensional parameter, the model assumes that each city has separate and possibly different difficulties λ_{k1} and λ_{k2} in the x- and y-direction, respectively. Due to geographical features of a map such as borders, lakes, coasts, or other anchor points, informants may be naturally restricted in the positioning of a location in the vertical direction but not in the horizontal direction or

vice versa. For instance, when positioning Liverpool and Dublin, informants are limited by the coastline to the West and the East, respectively, which may in turn result in a reduced variance of judgments in the x-direction (longitude) compared to the y-direction (latitude).

More generally, certain features of geographic maps such as coastlines may also lead to spatially correlated errors of location judgments. For instance, a positive correlation may emerge when positioning cities on a map which are closely located to a "diagonal" coastline (e.g., Aberdeen which is located close to a coast going from South-West to North-East). In other cases, however, informants are not restricted by nearby coasts (e.g., Birmingham), meaning that judgment errors in x- and y- direction may be uncorrelated. Overall, these considerations lead us to allow for a stochastic dependence of the judgment errors ε_{ik1} and ε_{ik2} in the x- and y-direction, respectively. We thus assume that, for each item k, the normally-distributed errors may correlate between the two dimensions with correlation ρ_k (as illustrated by the tilted red ellipses in Figure 1A). This results in the following covariance matrix of the two-dimensional judgment errors in Equation 3:

$$\mathbf{\Sigma}_{ik} = \begin{pmatrix} (E_i\lambda_{k1})^2 & \rho_k E_i^2 \lambda_{k1}\lambda_{k2} \\ \rho_k E_i^2 \lambda_{k1}\lambda_{k2} & (E_i\lambda_{k2})^2 \end{pmatrix}. \tag{5}$$

Hence, the errors may be correlated between the two dimensions within each item for each informant, which does, however, not imply that the errors are correlated across items or informants. Hence, the CCT-2D model still satisfies the conditional-independence assumption with respect to the two-dimensional vector of errors $\boldsymbol{\varepsilon}_{ik}$.

2.4 Model simplifications

Compared to the CCT model for one-dimensional continuous data developed by Anders et al. (2014), we simplified the CCT-2D model for two-dimensional judgments with respect to several aspects. First, we do not assume multiple cultural truths. In our example of positioning cities on a map of the United Kingdom and Ireland, multiple

cultural truths would imply that there are two or more latent classes of informants with each group having a different consensus of where the cities are located (Anders & Batchelder, 2012). When inferring the position of unknown locations such as natural resources, missing victims, or ancient archaeological sites, we assume that informants often use similar information and background knowledge to form their judgment. Thus, a multimodal distribution of distinct patterns of location judgments is possible but rather unlikely. In other scenarios such as the city-location task, a single correct position on the map does exist but is not available to the informants. In such cases, CCT is most useful when it provides a single, competence-weighted group-level estimate for each item which can then be compared to the accuracy of other aggregation approaches such as unweighted averaging (Merkle et al., 2020; Waubert de Puiseau et al., 2017).

Second, we do not assume a systematic response bias of location judgments. A bias for one-dimensional responses means that informants generally shift all their answers up or down to a certain degree as reflected by an additive component for each informant (Anders et al., 2014). When positioning cities on a map of the United Kingdom, a response bias would imply that informants shift all their location judgments in a certain direction by a fixed distance (e.g., horizontally, vertically, or diagonally). However, such a general shift of location judgments for all items seems to be unlikely given that certain cues provided by the map (e.g., the borders, coasts, or other geographic features) constrain the possible responses for each item in different ways. For instance, when positioning cities on a map of the United Kingdom and Ireland, a bias to the east would simply result in slightly biased judgments for some cities (e.g. London, Birmingham, and Manchester) but to judgments located in the ocean for others (e.g., Glasgow and Dublin). Hence, the CCT-2D model does not assume a response bias.

Lastly, the CCT-2D model does not include a scaling-bias parameter. For one-dimensional continuous data, a scaling bias refers to a multiplicative bias (i.e., a "stretching factor") for each informant which is assumed to affect the judgments of all items (Anders et al., 2014). When giving location judgments, a scaling bias would mean that informants' judgments on each axis and for all items are scaled by a multiplicative

component resulting in location judgments that are, for instance, positioned at about half of the correct latitude. Since informants do not give their judgments numerically but geographically, a scaling bias would depend on where the origin of the coordinate system is located, which is usually unknown to the informants. Moreover, a possible bias should not depend on the underlying coordinate system. We thus did not implement a scaling bias in the CCT-2D model.

2.5 Hierarchical Bayesian modeling

To fit the CCT-2D model to data and estimate its parameters, we adopt the hierarchical Bayesian modeling approach by Anders et al. (2014). Hierarchical modeling allows researchers to specify a population distribution for a set of model parameters such as person abilities or item difficulties (Lee & Wagenmakers, 2014). This provides many benefits such as a partial pooling of the information between the individual and the group level, which in turn results in shrinkage of the estimates (e.g., Heck, 2019; Singmann & Kellen, 2019). In our case, we assume separate population distributions of the competence parameters E_i across informants and of the item difficulty parameters $\boldsymbol{\lambda}_k$ across questions.

Besides specifying hierarchical distributions, the Bayesian framework also requires to define prior distributions. In the following, we adopt the common notation of distributions of the software JAGS (Plummer, 2003) which is used to fit the CCT-2D model below. The normal distribution is thus not parameterized by the mean μ and the standard deviation σ, but rather by the mean μ and the precision parameter $\tau = 1/\sigma^2$ (i.e., the inverse of the variance). Similarly, for the t distribution, the second parameter refers to the precision and not to the scale parameter.

Often, normal distributions are assumed as hierarchical group-level distributions. Concerning the latent truth for each item k, we assume that the cultural truth coordinates T_{kd} (with dimension index $d = 1, 2$) are located on the real line and are normally distributed across items,

$$T_{kd} \sim \text{Normal}(\mu_T, \tau_T). \tag{6}$$

In contrast, the parameters E_i and λ_{kd} are constrained to be positive. As a remedy, we first apply a log transformation to obtain parameters on the real line for which we can assume unbounded normal distributions (Anders et al., 2014). Taking the dimensionality of the parameters into account, the CCT-2D model assumes a one-dimensional hierarchical distribution of the informants' competence,

$$\log E_i \sim \text{Normal}(\mu_{\log E}, \tau_{\log E}), \tag{7}$$

and a two-dimensional distribution (with dimensions $d \in \{1, 2\}$) of the items' difficulty,

$$\log \boldsymbol{\lambda}_k \sim \text{MV-Normal}(\boldsymbol{\mu}_{\log \lambda}, \boldsymbol{\Sigma}^{-1}_{\log \lambda}). \tag{8}$$

For Bayesian inference, it is necessary to specify prior distributions for the hyperparameters of the hierarchical group-level distributions (e.g., for $\mu_{\log E}$ and $\boldsymbol{\mu}_{\log \lambda}$). Our main goal is to estimate the parameters reflecting cultural truth, competence, and item difficulty. Since we are not interested in testing hypotheses with theoretically informed prior distributions (e.g., via Bayes factors, Heck et al., 2022), we rely on prior distributions that are only weakly informative. Moreover, some hyperparameters are fixed to constants to ensure the identifiability of the resulting model similar as in item response theory (Embretson & Reise, 2000). For the correlation of judgment errors in the x- and y-direction for item k, we assume the following prior:

$$\rho_k \sim \text{Uniform}(-1, 1). \tag{9}$$

For the mean and precision of the latent truth coordinates, we assume

$$\mu_T \sim \text{Normal}(0, 0.25) \tag{10}$$

$$\tau_T \sim \text{Half-}t_{\text{df}=1}(0, 1). \tag{11}$$

For the mean and standard deviation of the (log) competence, the prior is

$$\mu_{\log E} = 0 \tag{12}$$

$$\sigma_{\log E} \sim \text{Half-}t_{\text{df}=1}(0, 1). \tag{13}$$

For the mean and standard deviation of the (log) difficulty parameters, we assume

$$\mu_{\log \lambda, d} = 0 \tag{14}$$

$$\sigma_{\log \lambda, d} \sim \text{Half-}t_{\text{df}=1}(0, 3). \tag{15}$$

Finally, the prior for the correlation of the (log) difficulty in x- and y-direction across items is

$$\rho_{\log \lambda} \sim \text{Uniform}(-1, 1). \tag{16}$$

A positive correlation $\rho_{\log \lambda}$ means that if positioning a city is difficult with respect to one axis, it is also difficult with respect to the other axis.

3 Simulation study

We performed a simulation study to examine general properties of the CCT-2D model. First, we want to assess how well the model can recover the true, data-generating parameters in various, realistic scenarios. Second, we compare the accuracy of location estimates obtained with the CCT model for two-dimensional continuous data to location estimates obtained with the unweighted aggregation of judgments. Simulated data and R scripts are available at https://osf.io/jbzk7/.

3.1 Method

In the simulation study, the following factors were varied in a fully crossed design using 100 replications per cell:

- Number of informants: $N = 10, 20, 50, 100$
- Number of items: $M = 5, 10, 25, 50$
- Standard deviation of log informants' competence: $\sigma_{\log E} = 0, 0.25, 0.5, 1$
- Standard deviation of log item difficulty: $\sigma_{\log \lambda} = 0, 0.25, 0.5, 1$

We chose a wide range for the sample size N to illustrate the effect of having few or many informants on parameter recovery and on the relative performance of CCT-2D compared to unweighted averaging. However, informants' competence can only be

estimated precisely if the number of items is sufficiently large. Hence, we also varied the number of items M on a large range. Overall, these settings reflect the fact that CCT is useful for a wide range of scenarios with both smaller and larger numbers of informants who answer more or less questions (e.g., Waubert de Puiseau et al., 2012).

Furthermore, we varied the standard deviation of the logarithm of informants' competence ($\sigma_{\log E}$) and the standard deviation of the logarithm of item difficulty ($\sigma_{\log \lambda}$) on a large range, including conditions with no variance at all. The standard deviations refer to the logarithm of these parameters since informants' competence and item difficulty must be positive, which also reflects the model's assumption that the log-transformed parameters follow unbounded normal distributions. While both types of variances can be expected to affect parameter recovery of their respective parameters, $\sigma_{\log E}$ is especially relevant for the comparison of the accuracy of estimates obtained with CCT-2D and unweighted averaging. Without any variance in informants' competence, CCT and unweighted averaging are expected to perform approximately equally well because equal weighting of judgments leads to optimal performance (Davis-Stober et al., 2014). However, if the variance in informants' location judgments partially emerges due to differences in informants' competence, CCT-2D is expected to result in more accurate estimates than unweighted averaging because it assigns larger weights to competent informants (Merkle et al., 2020).

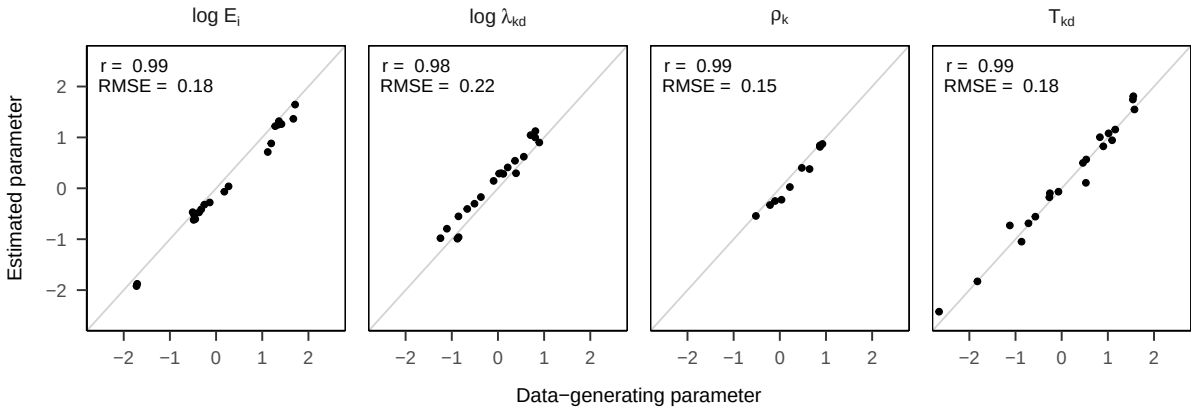
All simulations were conducted with the software JAGS (Plummer, 2003) in R using the packages `rjags` and `runjags` (Denwood, 2016; Plummer, 2021). For parameter estimation, we used 8,000 Markov chain Monte Carlo (MCMC) samples from six chains with 1,000 adaptions, 1,500 burn-in iterations, and a thinning factor of 3. These MCMC settings were selected to achieve a potential scale reduction factor of $\hat{R} < 1.1$ for all parameters. For this purpose, we first performed a small-scale simulation study with only few informants, few items, and a small variance in informants' competence and item difficulty to adjust the setting for JAGS. In the main simulation study, only 56 simulations (0.22%) did not converge with more than 10% of parameters having a potential scale reduction factor of $\hat{R} > 1.1$ and were, thus, excluded from the

analysis. For the remaining simulations, the average potential scale reduction factor was $\hat{R} = 1.002$ (99% quantile = 1.02). The model code for JAGS can be found in Appendix A.

3.2 Parameter recovery

Figure 2

Parameter recovery of the CCT-2D model for a single simulated data set.

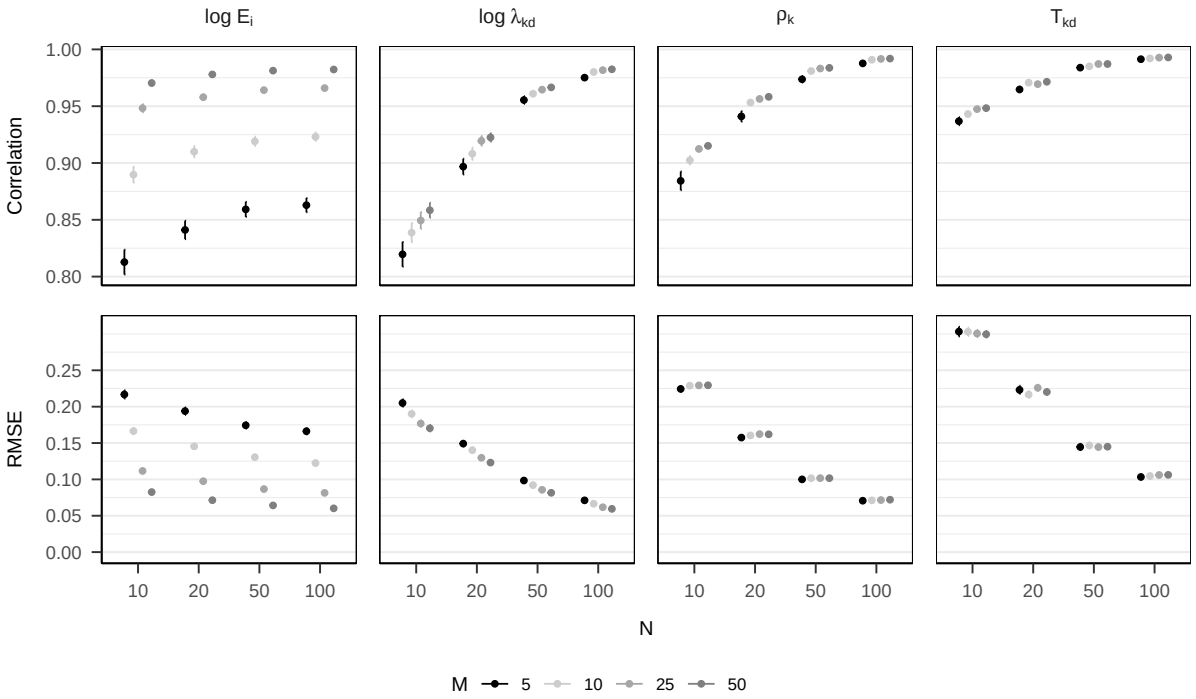


Note. Parameter recovery for a single simulated data set with $N = 20$ informants, $M = 10$ items, $\sigma_{\log E} = 1$, and $\sigma_{\log \lambda} = 0.5$. The first two panels show the logarithm of informants' competence ($\log E_i$) and item difficulty ($\log \lambda_{kd}$).

To examine parameter recovery in our extended CCT model, we first investigate parameter recovery using a single simulated data set. For this example, we chose a model with $N = 20$ informants, $M = 10$ items, a standard deviation of informants' competence of $\sigma_{\log E} = 1$, and a standard deviation of item difficulty of $\sigma_{\log \lambda} = 0.5$. Figure 2 shows the data-generating and estimated parameters for $\log E_i$, $\log \lambda_{kd}$, ρ_k, and T_{kd} including the correlation of data-generating and estimated parameters and the root-mean-square error (RMSE). For the vector-valued parameters $\boldsymbol{\lambda}_k$ and $\boldsymbol{T}_k$, the data-generating and estimated values for the x- and y-dimension are displayed jointly in the respective panels. All correlations are above .98 with the RMSE of the estimates ranging between 0.15 and 0.22. This indicates that the CCT-2D model performs quite well even with a moderate number of informants and items.

Figure 3

Parameter recovery across 25,544 replications.



Note. Average correlations of data-generating and estimated parameters and RMSEs are displayed with 95% confidence intervals. For simulations with $\log \sigma_E = 0$ and $\log \sigma_\lambda = 0$, no correlations could be computed for the parameters $\log E_i$ and $\log \lambda_{kd}$, respectively.

To judge the performance of the CCT-2D model for various scenarios, we assess the parameter recovery by computing the average correlation and RMSE of the data-generating and the estimated parameters across all 25,544 replications. Again, we display the correlation and RMSE for $\log \lambda_{kd}$ and T_{kd} for both dimensions in one panel. For all simulations with $\sigma_{\log E} = 0$ or $\sigma_{\log \lambda} = 0$, the correlation of generated and posterior values for $\log E_i$ and $\log \lambda_{kd}$, respectively, cannot be computed. This affected 11,188 replications for which either $\sigma_{\log E}$, $\sigma_{\log \lambda}$, or both were zero.

Figure 3 displays the average correlation and RMSE for all combinations of N and M. The item parameters $\log \lambda_{kd}$, ρ_k, and T_{kd} were clearly affected by the number of informants (N). This is due to the item parameters requiring a certain number of informants who answer these items to yield reliable parameter estimates. In contrast, the person parameters E_i were more strongly affected by the number of items (M).

This shows that the estimation of person parameters requires a certain number of items to be reliable. Of all parameters, RMSEs of the cultural truth T_{kd} were somewhat more affected by varying levels of N than those of all other parameters with RMSEs as high as 0.30. However, correlations of data-generating and estimated parameters of $\log \lambda_{kd}$ and $\log E_i$ were more strongly affected by varying levels of N and M respectively with correlations just above .80 for both parameters.

Furthermore, Figure 4 displays the parameter recovery of $\log E_i$ (Panel A) and $\log \lambda_{kd}$ (Panel B) for varying levels of $\sigma_{\log E}$ and $\sigma_{\log \lambda}$, respectively. While RMSEs are very small when there is no variance in either of the parameters, the recovery of E_i and λ_{kd} is worse for low levels of $\sigma_{\log E}$ and $\sigma_{\log \lambda}$, respectively, with correlations between data-generated parameters and estimated parameters as low as .64 for $\log E_i$ and .65 for $\log \lambda_{kd}$. However, as already observed in Figure 3, with increasing M, parameter recovery for $\log E_i$ improves, and with increasing N, parameter recovery for $\log \lambda_{kd}$ improves.

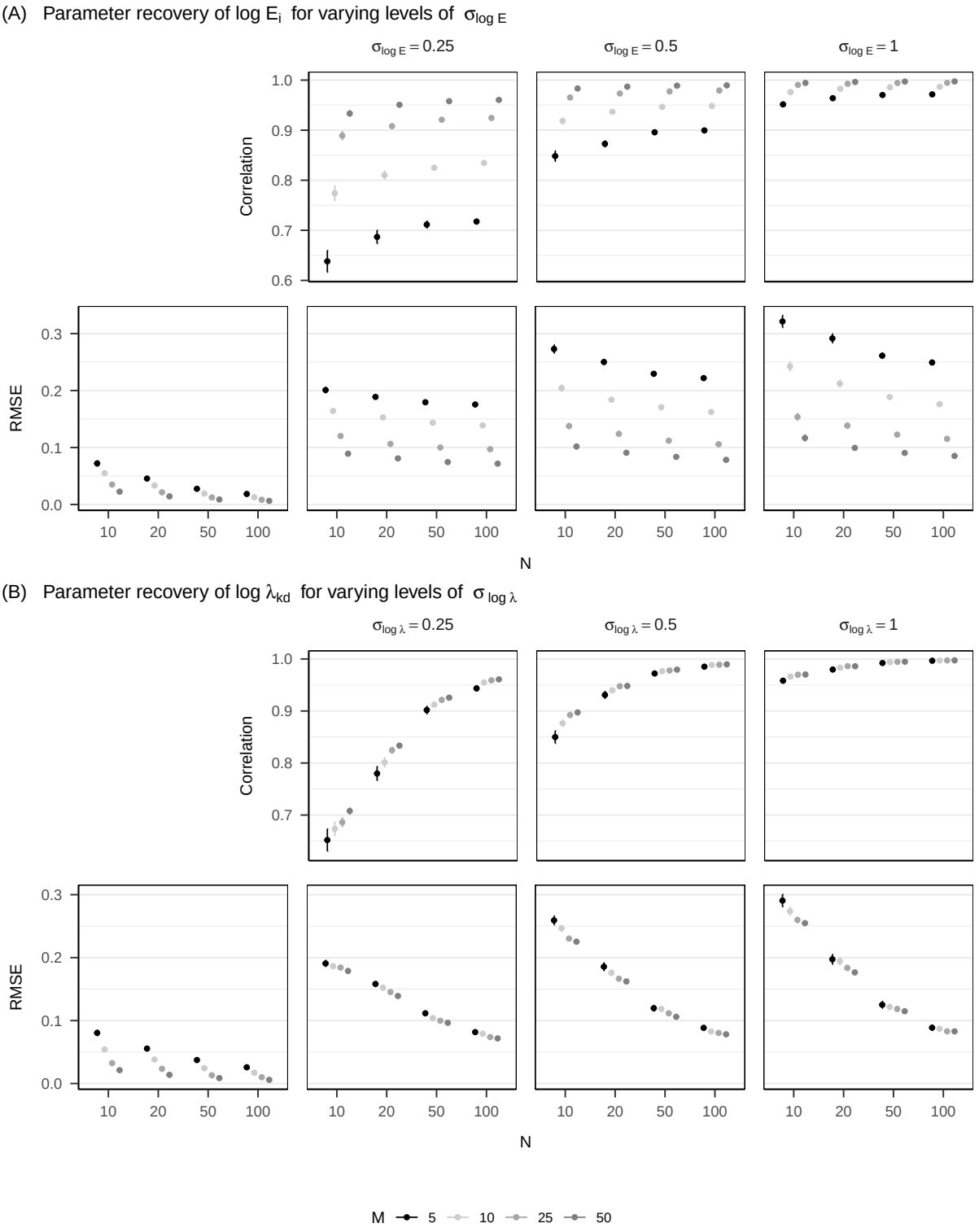
Overall, parameter recovery is acceptable for small N and M as well as low levels of $\sigma_{\log E}$ and $\sigma_{\log \lambda}$. As expected, all parameters show better recovery the larger N and M are and the larger the variances in informants' competence and item difficulty are. Accordingly, if N and M are small while there is little variance in $\sigma_{\log E}$ and $\sigma_{\log \lambda}$, the parameters $\log E_i$ or $\log \lambda_{kd}$ cannot be estimated reliably.

3.3 Comparing the accuracy of CCT-2D and unweighted averaging

In the following, we compare the accuracy of aggregating two-dimensional location judgments either with the CCT-2D model or with unweighted averaging. To obtain unweighted group-level estimates, we simply computed the unweighted mean of all location judgments for each item (separately for the x- and the y-coordinate). As a measure of accuracy, we use the Euclidean distance to the correct position for each item. Figure 5 displays the mean Euclidean distances across all items between the correct values and the CCT-2D estimates (gray points) and between the correct values and the estimates obtained with unweighted averaging (black points). To facilitate

Figure 4

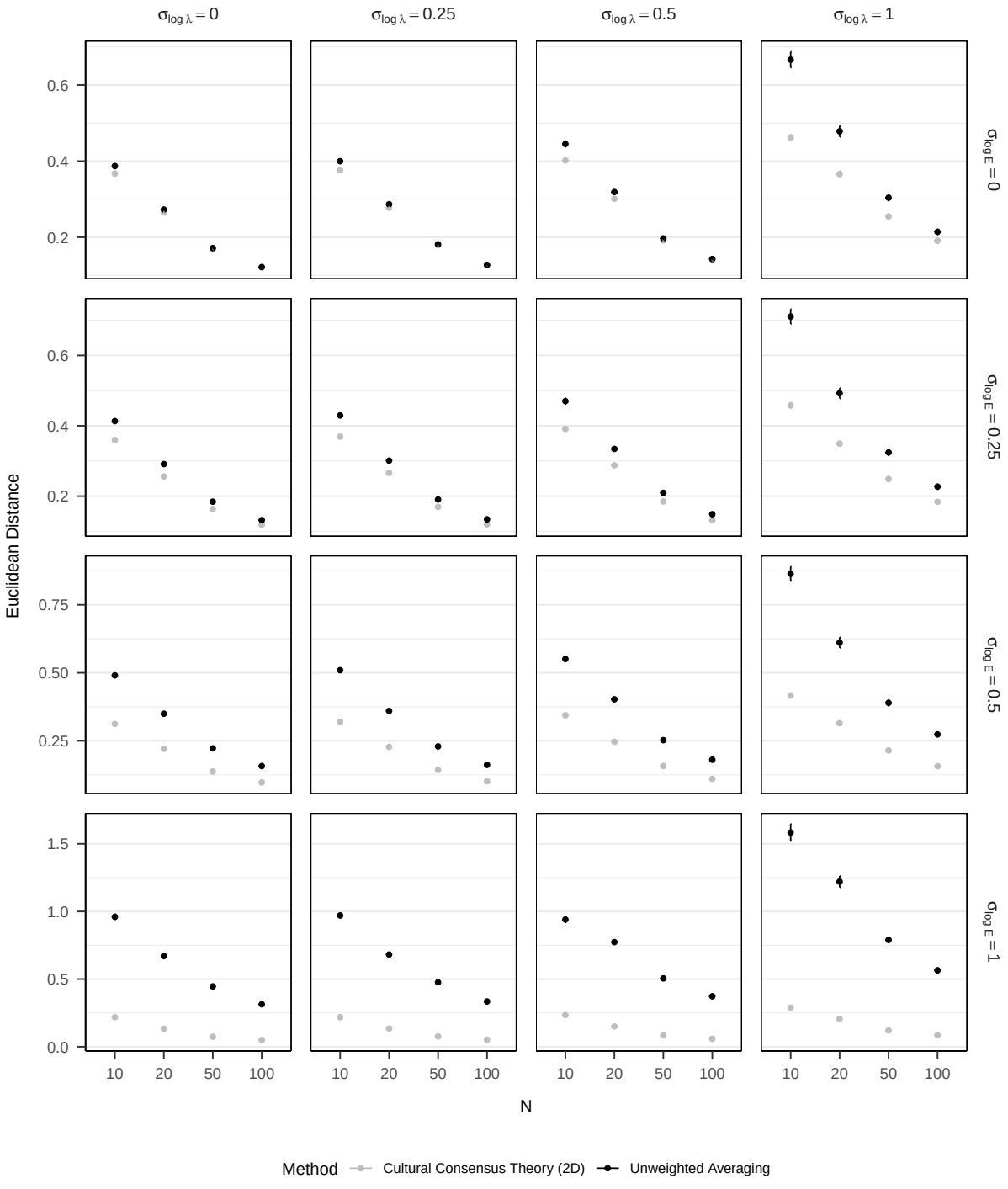
Average parameter recovery for different $\sigma_{\log E}$ or $\sigma_{\log \lambda}$.



Note. Mean correlations and RMSEs are displayed with 95% confidence intervals. For simulations with $\sigma_{\log E} = 0$ and $\sigma_{\log \lambda} = 0$ no correlations could be computed for $\log E_i$ and $\log \lambda_{kd}$, respectively.

Figure 5

Marginal accuracy of aggregate location estimates.



Note. The scaling of the y-axis differs across rows to improve readability. Mean accuracy is displayed with 95% confidence intervals.

interpretation of the results, we aggregated across replications with varying numbers of items.

As expected, Figure 5 shows that aggregating location judgments with CCT-2D yielded more accurate estimates than aggregating judgments with unweighted averaging. However, without any variance in informants' competence ($\sigma_{\log E} = 0$) or item difficulty ($\sigma_{\log \lambda} = 0$), both methods lead to equally accurate location estimates (upper left panel). In line with the principles of averaging out individual errors, Figure 5 shows that both unweighted averaging and CCT generally provided more accurate estimates the larger the sample of informants was. However, increasing sample size was more beneficial for unweighted averaging than for CCT estimates. Furthermore, estimates obtained with unweighted averaging became worse the larger the variance in informants' competence became. This was expected since increasing the heterogeneity of informants' competence yields larger variation in judgments, which in turn results in larger Euclidean distances to the correct position. The CCT model accounts and corrects for this additional variance in the observed location judgments, thereby resulting in a better recovery of the latent truth.

Even in the absence of differences in competence (first row in Figure 5), CCT-2D resulted in more accurate location estimates than unweighted averaging. This effect is due to shrinkage of the item parameters in the Bayesian hierarchical model. More precisely, the CCT-2D model assumes a hierarchical group-level distribution of the cultural-truth parameters $\boldsymbol{T}_k$ across items. Shrinkage of these random-effect parameters results in estimates closer to the mean μ_T compared to estimates based on assuming independent item parameters (i.e., fixed effects, Heck, 2019). As a consequence, extreme estimates are avoided especially when there are only few judgments for each item (i.e., if the sample size N is small). In Figure 5, this results in a higher accuracy of CCT-2D compared to unweighted averaging even in the absence of differences in competence. However, with increasing numbers of judgments per item (i.e., for larger N), shrinkage is reduced as the item parameters can be estimated more precisely. In turn, this results in a similar accuracy for CCT-2D and unweighted averaging. Overall, our comparison

shows that CCT-2D can increase the accuracy of aggregated location judgments by accounting for heterogeneity in competence and item difficulty.

4 Empirical study

In addition to the simulation study, we also apply the CCT-2D model to empirical data of participants who located various European cities on geographic maps (Mayer & Heck, 2021). Additionally, we compare the accuracy of aggregated location judgments of CCT-2D and unweighted averaging. Since multiple informants provided judgments for multiple items from the same knowledge domain (i.e., locations of European cities), the data fulfills the necessary requirements for an analysis with CCT-2D. All data and R scripts are available at https://osf.io/jbzk7/.

4.1 Methods

In the following, we reanalyze the data of a study by Mayer and Heck (2021) in which participants had to judge the location of 57 European cities on 7 different maps. We recruited 417 adult participants via a commercial German panel provider for an experiment on collaboration. 235 of these participants completed a condition in which they provided independent location judgments for all the presented items which makes their data suitable for an reanalysis with both CCT-2D and unweighted averaging. However, we excluded 7 participants who positioned more than 10% of the cities outside of the countries of interest (which were highlighted in white color), resulting in a total of 228 participants. In the remaining sample of participants, the mean age was 46.68 (SD = 15.23) and 46.9% of the participants were female. Most participants had a college degree (34.2%) or a high-school diploma (25.9%), while 24.1% had vocational education, and 15.8% had a lesser educational attainment.

A comprehensive overview of all presented cities and maps can be found in Appendix B1. All maps were scaled to 1:5,000,000 and were presented as images with 800×500 pixels. At this scaling, the influence of earth's curvature is small and can be neglected in further analyses. The maps only showed oceans which were colored in blue, landmasses which were colored in white for countries of interest and in gray for all other

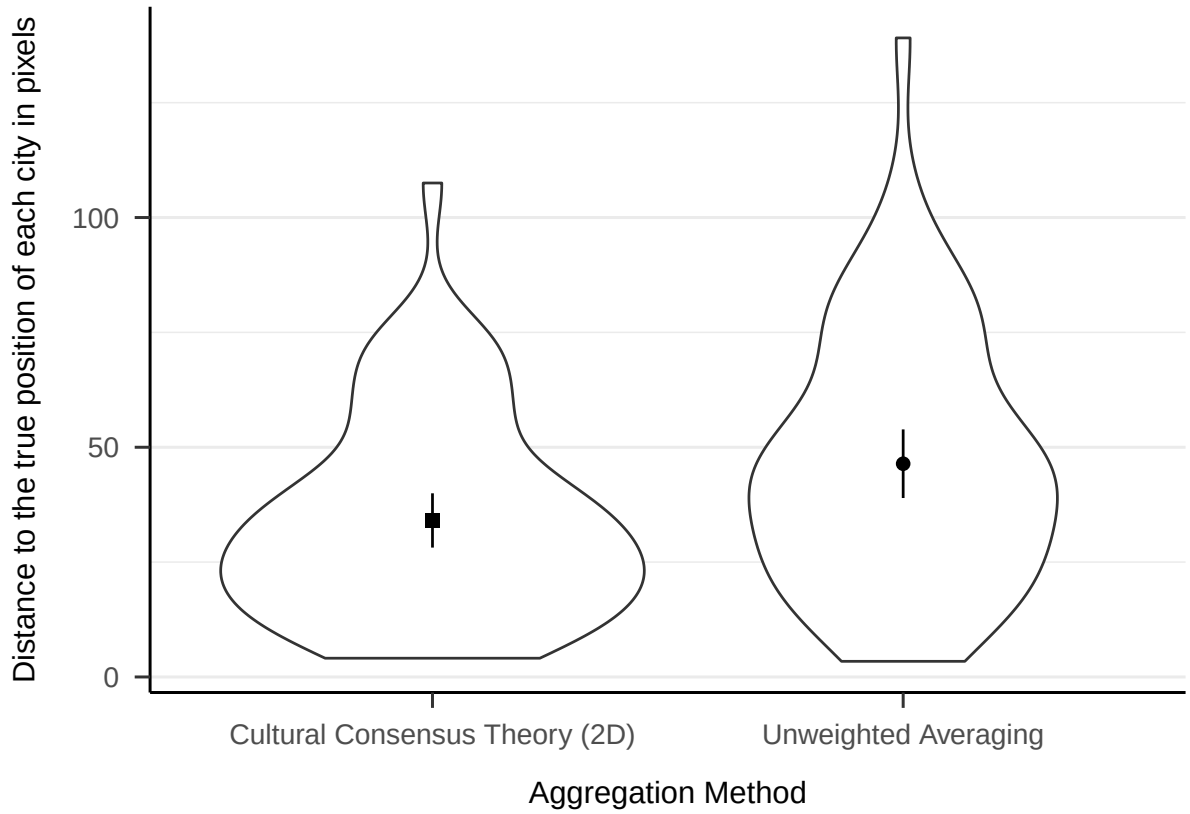
countries, and national borders as black lines as shown in Figure 7.

While completing the study, participants indicated the position of each of the 57 cities independently in separate trials. Maps and cities clustered within maps were presented in random order. Since the study was conducted online, we implemented a maximum time limit of 40 seconds for each item to prevent looking up the correct locations of the cities (for details, see Mayer & Heck, 2021).

4.2 Results

Figure 6

Accuracy of location estimates for 57 European cities.



Note. Reanalysis based on $N = 228$ participants from the data by Mayer and Heck (2021).

To compare the accuracy of CCT-2D and unweighted averaging, we first computed the group-level estimates for all locations of the 57 cities. For unweighted averaging, we simply aggregated the independent location judgments for each city by

taking the mean in the x- and the y-direction. For the CCT-2D model, we extracted the posterior-mean estimates of the two-dimensional cultural-truth parameters $\boldsymbol{T}_k$. We then computed the accuracy of the estimated locations by the Euclidean distance to the actual location of the presented cities.

Figure 6 displays the mean Euclidean distances across the 57 cities for the aggregate location estimates of CCT-2D and unweighted averaging. The results show that aggregating location judgments with CCT-2D resulted in more accurate estimates than unweighted averaging. To illustrate the advantage of CCT-2D for aggregating location judgments, Figure 7 displays the estimated locations of both methods as well as the correct locations for the five cities on the map of the United Kingdom and Ireland. CCT-2D shows more accurate estimates than unweighted averaging for four of the five cities (i.e., Birmingham, Dublin, Glasgow, and London) and an equally accurate estimate for one city (Liverpool). Notably, for some cities such as London, the distance between the true and the estimated location is approximately half as large for CCT-2D compared to unweighted averaging. The supplementary material provides plots of all seven European maps used in the study, each displaying the location estimates obtained with unweighted averaging and CCT-2D as well as the cities' actual positions (https://osf.io/jbzk7/).

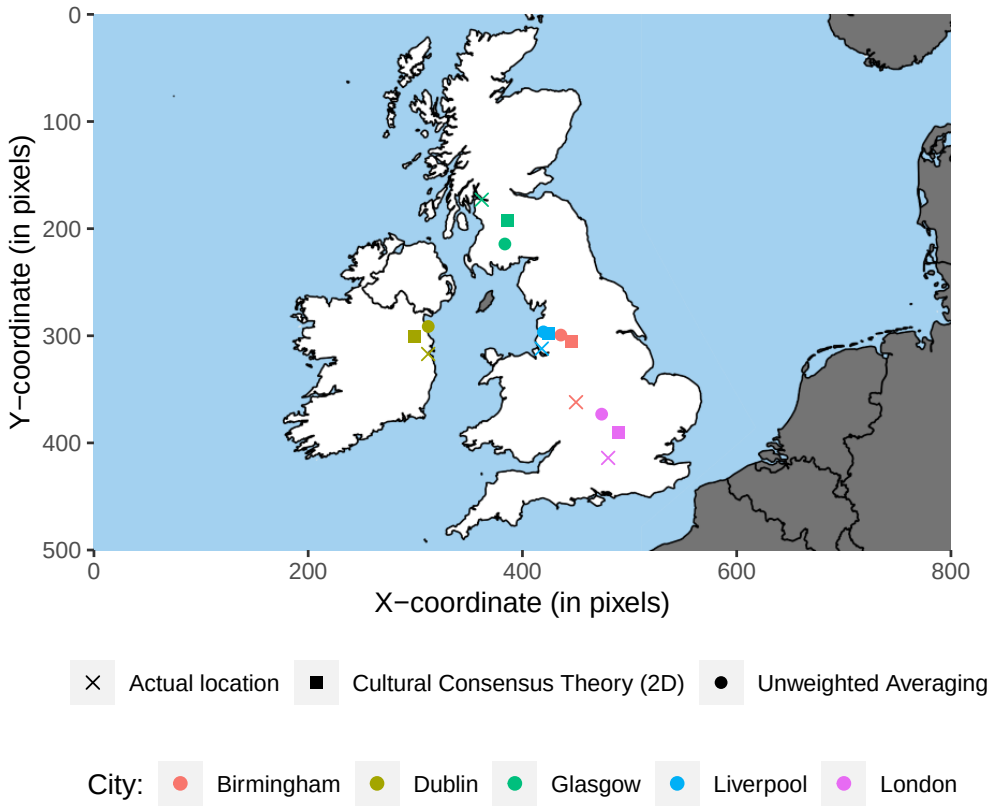
The descriptive patterns shown in Figures 6 and 7 were also supported by a statistical analysis. A paired-sample t-test showed that the accuracy of the CCT-2D estimates was significantly higher than that of estimates obtained with unweighted averaging ($t(56) = 10.43, p < .001$). Notably, Cohen's d indicated a large effect size of $d = 1.38$. Across all cities, estimates were on average 12.35 pixels closer to the correct position, resembling the improvement for Glasgow in Figure 7 which was 15.63 pixels.

To further examine the validity of the CCT-2D model, we also computed the correlation between the estimated competence parameters $\log E_i$ and individuals' education level. Individuals with a higher education level should have more geographic knowledge and thus provide more accurate judgments which are closer to the cultural truth. Since smaller values of the competence parameter indicate higher individual

competence (i.e., reflecting a smaller variance of judgments around the cultural truth), we expect a negative correlation between the estimated competence and education level. When encoding the education level as an ordinal variable, a Spearman rank correlation indeed showed a medium negative correlation of $-.35$ ($p < .001$), thus strengthening the validity of the CCT-2D model and the $\log E_i$ parameters.

Figure 7

Estimated versus actual locations of five cities.



5 Discussion

We proposed a novel model of Cultural Consensus Theory for two-dimensional location judgments (CCT-2D). The model is based on the hierarchical Bayesian CCT model by Anders et al. (2014) for one-dimensional data. The CCT-2D model estimates the latent cultural truths of the presented items, that is, the group's consensus knowledge concerning the (unknown) positions of the items. To do so, the model infers the informants' competence based on the distance of their response patterns to the shared consensus, as well as the difficulty of the items. To account for the spatial

structure of the two-dimensional data, the model assumes that judgment errors are correlated between the two dimensions for each item.

We successfully applied the new model both to simulated and empirical data. Using simulations, we showed that the CCT-2D model has a very good parameter recovery for a large range of numbers of informants and numbers of items. Moreover, the simulations showed that the CCT-2D group-level estimates for the latent truths of the locations were more accurate in terms of the Euclidean distance to the true locations than the estimates obtained with unweighted averaging of individual judgment. This is due to the fact that the CCT-2D model considers additional information obtained by inferring differences in the items' difficulty and the informants' competence. Furthermore, a reanalysis of an empirical study in which informants located 57 European cities on seven maps showed a large effect concerning an increase in accuracy of CCT-2D compared to unweighted averaging. These findings conceptually replicate the results of Merkle et al. (2020) who found that a CCT-inspired mechanism of weighting informants' judgments by their expertise outperformed unweighted averaging for one-dimensional forecasting judgments (i.e., for point spread forecasts of the Australian Football League).

5.1 Limitations and future research

While our results provide preliminary evidence for the usefulness of the proposed CCT-2D model, the model has several limitations that should be addressed in the future. First, it is possible that response biases may lead to a general shift of location judgments away from the borders into the interior regions of the presented maps. A similar effect may also occur due to certain geographic features such as coastlines or national borders (Friedman, Brown, et al., 2002; Friedman et al., 2005). Note that a simple, additive shift of all location judgments into a certain direction by a certain distance similar as in the one-dimensional CCT model by Anders et al. (2014) cannot describe such a complex, nonlinear bias towards inner regions. However, it may generally be difficult to disentangle complex, item-independent response biases from

distortions of the latent consensus knowledge about the locations of specific items both empirically and conceptually.

Second, the proposed CCT-2D model assumes bivariate normal distributions of the observed location judgments and of the latent truths concerning the positions of the presented items. However, locations on maps are naturally constrained by the borders of the map and by geographic features such as coasts or national borders (Friedman et al., 2005). It is thus likely that our assumption that location judgments and latent truths follow bivariate normal distributions with unbounded support is violated. As a remedy, the CCT-2D model of location judgments may be improved by implementing a truncation of the support in the two-dimensional space by respecting geographic features of the map. For instance, when estimating the location of Dublin, one may exclude observed judgments that position the city in the Atlantic Ocean, while also implementing a corresponding truncation for the support of the bivariate normal distribution of observed judgments (Gelfand et al., 1992). For the application of our model to empirical data, we simply excluded participants who positioned more than 10% of their judgments outside the highlighted countries of interest to more adequately fulfill this assumption.

In principle, it is also possible to truncate the support of the bivariate distribution of latent truths to landmasses only. Thereby, one ensures that all posterior samples of the inferred locations in MCMC sampling are actually located on land and away from the sea. However, implementing complex, nonlinear, two-dimensional truncations in JAGS or other software is not straightforward. Even when considering only a set of simple, linear order constraints, tailored MCMC algorithms are usually required to ensure that all posterior samples satisfy the constraints (Heck & Davis-Stober, 2019). Moreover, these methods often assume that the truncated parameter space is convex which is not the case for landmasses on geographic maps. Thus, we leave it to future research to implement the truncation of distributions in the CCT-2D model.

Besides aggregating location judgments on geographic maps, our extension of

CCT to two-dimensional continuous data can also be applied to other types of judgments such as continuous ratings of both the emotional arousal and valence of pictures on two visual analogue scales (Funke & Reips, 2012; Reips & Funke, 2008). When using such response scales, it is reasonable to include response-bias shifts and scaling biases as in Anders et al. (2014) to account for different response styles. The CCT-2D model can also easily be extended to d-multivariate responses on an arbitrary number of judgment dimensions. Such an approach could be useful, for instance, when rating faces with respect to several dimensions such as trustworthiness, attractiveness, and symmetry on continuous scales (Oosterhof & Todorov, 2008).

5.2 Conclusions

The proposed CCT-2D model extends the scope of applications of cultural consensus theory to two-dimensional continuous data. Researchers can now analyze and aggregate geographical location judgments consisting of x- and y-coordinates or longitude and latitude to infer the group's cultural knowledge about the unknown locations. In doing so, the model weighs the observed judgments both by the informants' competence and by the items' difficulty. Concerning the study design, it is necessary to recruit multiple informants who provide judgments for multiple items from the same knowledge domain. We showed that the CCT-2D model provides good parameter recovery and, in cases where the factual truth is known, provides aggregate group-level estimates that are more accurate than those obtained by the unweighted averaging of location judgments.

6 References

Al-shabeeb, A. R. (2016). The use of AHP within GIS in selecting potential sites for water harvesting sites in the Azraq Basin—Jordan. *Journal of Geographic Information System*, *8*(1), 73–88. https://doi.org/10.4236/jgis.2016.81008

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*, 151–181. https://doi.org/10.1007/s11336-013-9382-9

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452–469. https://doi.org/10.1016/j.jmp.2013.01.004

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13. https://doi.org/10.1016/j.jmp.2014.06.001

Aßfalg, A. (2018). Consensus theory for mixed response formats. *Journal of Mathematical Psychology*, *86*, 51–63. https://doi.org/10.1016/j.jmp.2018.08.005

Aßfalg, A., & Klauer, K. C. (2020). Consensus theory for multiple latent traits and consensus groups. *Journal of Mathematical Psychology*, *97*, 102374. https://doi.org/10.1016/j.jmp.2020.102374

Batchelder, W. H. (2009). Cultural consensus theory: Aggregating expert judgments about ties in a social network. *Social Computing and Behavioral Modeling*, 1–9. https://doi.org/10.1007/978-1-4419-0056-2_5

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332. https://doi.org/10.1016/j.jmp.2012.06.002

Batchelder, W. H., Kumbasar, E., & Boyd, J. P. (1997). Consensus analysis of three-way social network data. *The Journal of Mathematical Sociology*, *22*, 29–58. https://doi.org/10.1080/0022250X.1997.9990193

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92. https://doi.org/10.1007/BF02294195

Casana, J. (2014). Regional-scale archaeological remote sensing in the age of big data: Automated site discovery vs. Brute force methods. *Advances in Archaeological Practice*, *2*(3), 222–233. https://doi.org/10.7183/2326-3768.2.3.222

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*, 79–101. https://doi.org/10.1037/dec0000004

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Faghri, A., Lang, A., Hamad, K., & Henck, H. (2002). Integrated knowledge-based geographic information system for determining optimal location of park-and-ride facilities. *Journal of Urban Planning and Development*, *128*, 18–41. https://doi.org/10.1061/(ASCE)0733-9488(2002)128:1(18)

Friedman, A., Brown, N. R., & Mcgaffey, A. P. (2002). A basis for bias in geographical judgments. *Psychonomic Bulletin & Review*, *9*, 151–159. https://doi.org/10.3758/BF03196272

Friedman, A., Kerkman, D. D., & Brown, N. R. (2002). Spatial location judgments: A cross-national comparison of estimation bias in subjective North American geography. *Psychonomic Bulletin & Review*, *9*, 615–623. https://doi.org/10.3758/BF03196321

Friedman, A., Kerkman, D. D., Brown, N. R., Stea, D., & Cappello, H. M. (2005). Cross-cultural similarities and differences in North Americans' geographic location judgments. *Psychonomic Bulletin & Review*, *12*,

1054–1060. https://doi.org/10.3758/BF03206443

Friedman, A., Mohr, C., & Brugger, P. (2012). Representational pseudoneglect and reference points both influence geographic location estimates. *Psychonomic Bulletin & Review*, *19*, 277–284. https://doi.org/10.3758/s13423-011-0202-x

Funke, F., & Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, *24*, 310–327. https://doi.org/10.1177/1525822X12444061

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532. https://doi.org/10.2307/2290286

Heck, D. W. (2019). Accounting for estimation uncertainty and shrinkage in Bayesian within-subject intervals: A comment on Nathoo, Kilshaw, and Masson (2018). *Journal of Mathematical Psychology*, *88*, 27–31. https://doi.org/10.1016/j.jmp.2018.11.002

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2022). A review of applications of the Bayes factor in psychological research. *Psychological Methods*. In press. https://doi.org/10.1037/met0000454

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87. https://doi.org/10.1016/j.jmp.2019.03.004

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.

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

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Mahdavi, A., Niknejad, M., & Karami, O. (2015). A fuzzy multi-criteria decision method for ecotourism development locating. *Caspian Journal of Environmental Sciences*, *13*(3), 221–236. https://cjes.guilan.ac.ir/article_1373.html

Mayer, M., & Heck, D. W. (2021). *Sequential collaboration: Comparing the accuracy of dependent, incremental judgments to wisdom of crowds*. https://doi.org/10.31234/osf.io/w4xdk

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

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087–11092. https://doi.org/10.1073/pnas.0805664105

Plummer, M. (2003). *JAGS: A program for analysis of bayesian graphical models using Gibbs sampling*.

Plummer, M. (2021). *rjags: Bayesian graphical models using MCMC*. https://CRAN.R-project.org/package=rjags

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in internet-based research: VAS generator. *Behavior Research Methods*, *40*, 699–704. https://doi.org/10.3758/BRM.40.3.699

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus:

A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338. https://www.jstor.org/stable/677564

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology*. Psychology Press.

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. (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). Curran Associates, Inc.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Thorndyke, P. W., & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, *14*, 560–589. https://doi.org/10.1016/0010-0285(82)90019-6

Trotter, R., Weller, S., Baer, R., Pachter, L., Glazer, M., Alba-García, J., & Klein, R. (1999). Consensus theory model of AIDS/SIDA beliefs in four latino populations. *AIDS Education and Prevention*, *11*, 414–426.

Waubert de Puiseau, B., Aßfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied*, *18*, 390–403. https://doi.org/10.1037/a0029801

Waubert de Puiseau, B., Greving, S., Aßfalg, A., & Musch, J. (2017). On the importance of considering heterogeneity in witnesses' competence levels when reconstructing crimes from multiple witness testimonies. *Psychological Research*, *81*, 947–960. https://doi.org/10.1007/s00426-016-0802-1

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field Methods*, *19*, 339–368. https://doi.org/10.1177/1525822X07303502

Yarbrough, G. L., Wu, B., Wu, J., J. He, Z., & Leng, T. (2002). Judgments of object location behind an obstacle depend on the particular information

selected. *Journal of Vision*, *2*, 625. https://doi.org/10.1167/2.7.625

Appendix A

JAGS code for the CCT-2D model of two-dimensional location judgments

```
model{
  for(i in 1:n){
    for(k in 1:m){
      sigma[i,k,1] <- E[i]*lam[k,1]
      sigma[i,k,2] <- E[i]*lam[k,2]
      Sigma[i,k,1,1] <- pow(sigma[i,k,1], 2)
      Sigma[i,k,2,2] <- pow(sigma[i,k,2], 2)
      Sigma[i,k,1,2] <- rho[k] * sigma[i,k,1] * sigma[i,k,2]
      Sigma[i,k,2,1] <- rho[k] * sigma[i,k,1] * sigma[i,k,2]
      Tau[i,k,1:2,1:2] <- inverse(Sigma[i,k,1:2,1:2])
      Y[i,k,1:2] ~ dmnorm(T[k,1:2], Tau[i,k,1:2,1:2])
    }
  }

# Parameters
  for (i in 1:n){
    Elog[i] ~ dnorm(Emu,Etau)
    E[i] <- exp(Elog[i])
  }
  lamSigma[1,1] <- pow(lamsigmax, 2)
  lamSigma[2,2] <- pow(lamsigmay, 2)
  lamSigma[1,2] <- lamrho * lamsigmax * lamsigmay
  lamSigma[2,1] <- lamSigma[1,2]
  for (k in 1:m){
    T[k,1] ~ dnorm(Tmu,Ttau)
    T[k,2] ~ dnorm(Tmu,Ttau)
    lamlog[k,1:2] ~ dmnorm.vcov(lammu[1:2], lamSigma[1:2,1:2])
```

```
    lam[k,1] <- exp(lamlog[k,1])
    lam[k,2] <- exp(lamlog[k,2])
  }

# Hyperparameters
  Tmu ~ dnorm(0,0.25)
  Ttau ~ dt(0,1,1)T(0,)
  lammu[1] <- 0
  lammu[2] <- 0
  lamsigmax ~ dt(0,3,1)T(0,)
  lamsigmay ~ dt(0,3,1)T(0,)
  lamrho ~ dunif(-1, 1)
  Emu <- 0
  Etau <- pow(Esigma, -2)
  Esigma ~ dt(0,1,1)T(0,)

  for(k in 1:m){
    rho[k] ~ dunif(-1, 1)
  }
}
```

Appendix B

European cities used in the reanalysis

Table B1

European cities and maps from the study by Mayer and Heck (2021).

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

AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

https://doi.org/10.1037/dec0000193

Sequential Collaboration: The Accuracy of Dependent, Incremental Judgments

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

[1] Department of Psychology, School of Social Science, University of Mannheim

[2] Heidelberg Academy of Sciences and Humanities, Heidelberg, Baden-Württemberg, Germany

[3] Department of Psychology, University of Marburg

Online collaborative projects in which users contribute to extensive knowledge bases such as Wikipedia or OpenStreetMap have become increasingly popular while yielding highly accurate information. Collaboration in such projects is organized sequentially, with one contributor creating an entry and the following contributors deciding whether to adjust or to maintain the presented information. We refer to this process as *sequential collaboration* since individual judgments directly depend on the previous judgment. As sequential collaboration has not yet been examined systematically, we investigate whether dependent, sequential judgments become increasingly more accurate. Moreover, we test whether final sequential judgments are more accurate than the unweighted average of independent judgments from equally large groups. We conducted three studies with groups of four to six contributors who either answered general knowledge questions (Experiments 1 and 2) or located cities on maps (Experiment 3). As expected, individual judgments became more accurate across the course of sequential chains, and final estimates were similarly accurate as unweighted averaging of independent judgments. These results show that sequential collaboration profits from dependent, incremental judgments, thereby shedding light on the contribution process underlying large-scale online collaborative projects.

Keywords: wisdom of crowds, teamwork, mass collaboration, group decision-making

Supplemental materials: https://doi.org/10.1037/dec0000193.supp

Maren Mayer https://orcid.org/0000-0002-6830-7768

Daniel W. Heck https://orcid.org/0000-0002-6302-9252

This work was funded by the WIN programme of the Heidelberg Academy of Sciences and Humanities, which is financed by the Ministry of Science, Research, and the Arts of the State of Baden-Württemberg, and also supported by the Research Training Group "Statistical Modeling in Psychology," funded by the German Research Foundation (DFG Grant GRK 2277).

Maren Mayer played lead role in formal analysis, investigation, and writing of original draft and equal role in conceptualization, methodology, and writing of review and editing. Daniel W. Heck played equal role in conceptualization, methodology, and writing of review and editing.

This work was presented at the 62nd Conference of Experimental Psychologists (Virtual TeaP, 2021). The present version of the article (May 16, 2022) has not yet been peer reviewed. A preprint was uploaded to PsyArXiv and Research Gate for timely dissemination.

The data are available at https://osf.io/96nsk/

The experimental materials are available at https://osf.io/96nsk/

The preregistered design (transparent changes notation) is available at https://aspredicted.org/8vn9m.pdf; https://aspredicted.org/9j9de.pdf

Correspondence concerning this article should be addressed to Maren Mayer, Department of Psychology, School of Social Science, University of Mannheim, B6, 30-32, 68159 Mannheim, Germany. Email: maren.mayer@students.uni-mannheim.de

Collaborative online projects have become a popular source for information gathering over the last 20 years. The most prominent example is Wikipedia, an online encyclopedia that allows users to contribute semantic information to various topics in the form of structured articles (Wikipedia Contributors, 2021). Giles (2005) showed that information on Wikipedia is very accurate in general. Moreover, certain topics, such as information on cancer or certain drugs, are similarly accurate as official health information or text books (Kräenbring et al., 2014; Leithner et al., 2010). Another example of online collaboration is OpenStreetMap, a collaborative project that aims at generating a comprehensive, open, and free-to-use map of the world (OpenStreetMap Contributors, 2021). OpenStreetMap does not only comprise geographical numeric information about the locations of objects, such as coordinates, but also semantic information such as names of streets, areas, buildings, and other useful information (e.g., addresses or websites of shops and restaurants). Comparing the accuracy of OpenStreetMap with commercial map providers or governmental sources also revealed a comparable accuracy (e.g., Girres & Touya, 2010; Zielstra & Zipf, 2010).

The high accuracy of Wikipedia and other online collaborative projects has often been attributed to the wisdom of crowds (e.g., Arazy et al., 2006; Kittur & Kraut, 2008; Niederer & van Dijck, 2010), which refers to the aggregation of judgments from different informants (Galton, 1907; Surowiecki, 2004). The term "wisdom of crowds" is a broad concept encompassing various methods of eliciting and aggregating judgments. As a measure for crowd wisdom, prior work often examined the unweighted mean or median of independent individual judgments (e.g., Budescu & Chen, 2014; Davis-Stober et al., 2014; Galton, 1907; Hueffer et al., 2013; Larrick & Soll, 2006; Merkle et al., 2020). The high accuracy of these judgments is due to the central limit theorem, which ensures that errors in independent, individual judgments cancel out (Hogarth, 1978) and has been demonstrated for various tasks and in various contexts (e.g., Hueffer et al., 2013; Steyvers et al., 2009). The accuracy of unweighted averaging of independent individual judgments increases when judgments bracket the true answer (Larrick & Soll, 2006; Simmons et al., 2011) and are negatively correlated and unbiased (Davis-Stober et al., 2014; Keck & Tang, 2020).

The collection and aggregation of judgments in online collaborative projects can be regarded as a certain type of wisdom of crowds. However, in online collaboration, judgments are usually not collected independently and then aggregated mechanically but rather elicited in a dependent and sequential manner. Instead of providing independent judgments, contributors encounter already existing entries and decide whether to change the presented information reflecting the latest version of an entry or whether to maintain the presented version. We refer to this way of collaborating as sequential collaboration. Because unweighted averaging is known to result in highly accurate estimates for various tasks and contexts, we will use it as a benchmark for assessing the accuracy of sequential collaboration.

In the following, we first define sequential collaboration and distinguish it from other forms of collaboration and aggregating judgments. Next, we discuss prior research on dependent judgments, which has shown both positive and detrimental effects of dependency. Our main goal is to compare sequential collaboration to unweighted averaging. We investigate why and under which conditions the elicitation of incremental, dependent judgments can benefit accuracy compared to taking the unweighted average of independent individual judgments. In three studies, we used general knowledge questions and maps on which cities should be positioned to test whether sequential collaboration within small groups of four to six contributors yields improved judgments. Moreover, we tested whether the final judgments of a sequential chain are more accurate than estimates obtained by aggregating independent individual judgments. In line with our hypotheses, we found that judgment accuracy increased over the course of sequential chain and that sequential collaboration yielded similarly accurate results as unweighted averaging.

Sequential Collaboration

As outlined above, collaboration in online projects is often organized sequentially by making incremental changes to the latest available information. Sequential collaboration starts with one contributor creating an initial, independent entry. The next contributors encountering this entry then decide whether to adjust or maintain the presented information. Whenever the entry is changed, the information is updated such that only the latest

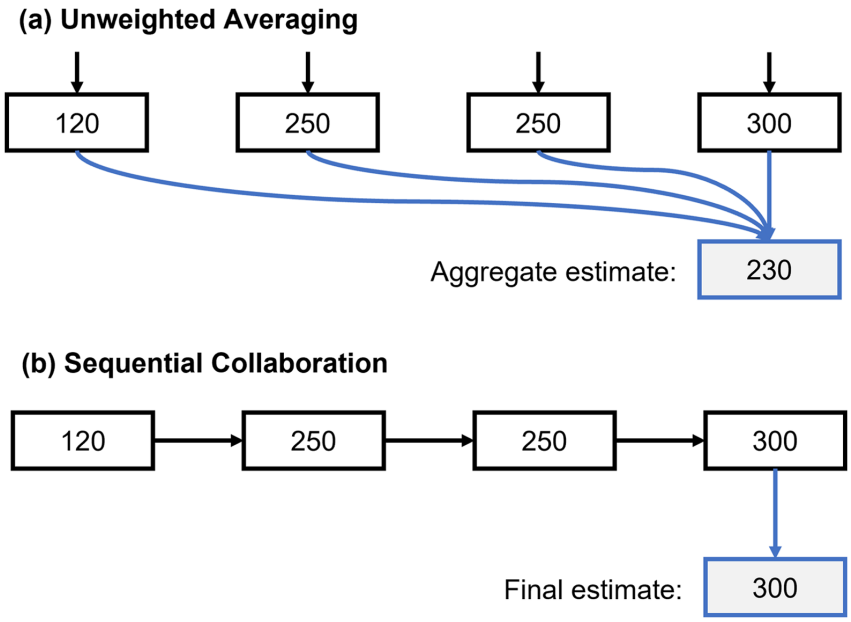
version of the entry is presented to the next contributor. For example, the first contributor might respond to the question "How tall is the Eiffel Tower?" with 420 m. The second contributor encountering this judgment could simply maintain it, whereas the third contributor might adjust the height to 290 m. After several contributors have adjusted and maintained the judgment, the correct height of 300 m may be entered. The sequence of decisions of whether to maintain or adjust entries made by a previous contributor forms a sequential chain. Figure 1 displays how group estimates are generated with unweighted averaging and in sequential collaboration. In the former, the aggregated estimate is obtained by averaging independent individual judgments; in the latter, the estimate is the last judgment in a sequential chain generated by adjusting and maintaining previous judgments.

Even though sequential collaboration is performed by a group of individuals and shares some aspects with other forms of group decision-making, it also has some unique features distinguishing it from other forms of collaboration. In research on group decision-making, group work usually takes place simultaneously (Kerr & Tindale, 2004; Lu et al., 2012; Stasser & Titus, 1985), even though interactions do not necessarily take place in person (Dennis, 1996; Dennis et al., 1998; Lu et al., 2012). In a paradigm organized like this, all members of the group have the opportunity to listen to all judgments and opinions, to ask questions to other group members, and to share justifications and other information. In sequential collaboration, however, information is shared only by adding or correcting the judgment of a previous contributor, which implies that the dependency between judgments is limited to the displayed information. Furthermore, direct interactions with other contributors are neither necessary nor possible in sequential collaboration, and additional information such as the number of adjustments already made to this information or reasons why information was adjusted is initially not available.

A form of collaboration similar to sequential collaboration is the Delphi method (Dalkey & Helmer, 1963; Geist, 2010; Jeste et al., 2010). The Delphi method was designed to obtain judgments on a given topic from a group of experts who do not interact directly. First, experts provide independent judgments and justifications for these judgments, which are then combined in a report by a moderator. This report is sent to all experts, who can then revise their judgments based on the judgments and information included in the report. When experts have reached a sufficient consensus, the individual judgments are aggregated to a final

Figure 1

Illustration of Group Estimates in Unweighted Averaging and Sequential Collaboration



Note. See the online article for the color version of this figure.

result. The Delphi method is similar to sequential collaboration in that individuals do not directly interact with each other. However, in sequential collaboration, contributors do neither receive judgments of several other contributors nor justifications of these judgments. Moreover, contributors are not necessarily required to provide a judgment, and even if they do, they may not notice if their judgment is in turn adjusted by others. Finally, the Delphi method focuses on eliciting judgments by a group of experts, whereas in sequential collaboration, neither the specific contributors nor the number of contributors have to be predefined.

Possible Issues and Benefits of Sequential Collaboration

Even though sequential collaboration seems to be a successful way of integrating judgments of various individuals, the process of sequentially deciding whether to adjust or maintain a previous judgment has not been systematically examined yet. Nonetheless, research on related phenomena allows us to derive testable predictions.

Possible issues for the accuracy of sequential collaboration may arise from the anchoring effect (Tversky & Kahneman, 1974). Anchoring describes the robust phenomenon that a presented numerical value influences subsequent, often unrelated numerical judgments (Mussweiler et al., 2004). This effect may undermine the accuracy of sequential collaboration such that adjustments made to a previous judgment are systematically biased toward the previous judgment. Especially, when the previous judgment heavily over- or under-estimates the correct value, anchoring might affect later judgments such that arriving at accurate, unbiased estimates is prolonged or hindered.

The conditions under which information provided by others is considered in forming a judgment have been extensively studied in the advice-taking literature (Bonaccio & Dalal, 2006). Egocentric discounting describes the phenomenon that advice is generally underweighted relative to one's own initial judgment, in turn resulting in less accurate judgments compared to equally weighing the advice and one's own judgment (Yaniv & Kleinberger, 2000). In sequential collaboration, egocentric discounting could lead contributors to adjust the presented previous judgment mainly according to their prior beliefs, which in turn could be detrimental to accuracy as the chain may not converge to the correct answer. However, advice taking improves when no initial individual judgment is formed before receiving advice (Koehler & Beauregard, 2006). This resembles the situation in sequential collaboration more closely, since contributors are directly confronted with the previous judgment and do not have to form an initial, independent judgment. Hence, compared to the standard advice-taking paradigms, contributors in sequential collaboration may be more likely to accept a presented judgment.

Prior research also provides preliminary evidence in favor of the accuracy of sequential collaboration. Providing participants with a frame of reference improves subsequent judgments, especially because it prevents extreme judgments (Bonner et al., 2007; Laughlin et al., 1999). Previous judgments in a sequential chain may serve as a frame of reference that prevents extreme judgments and fosters to reach an accurate estimate earlier. However, especially at the beginning of a sequential chain, judgments by previous contributors may not provide an accurate frame of reference.

Providing judgments of other individuals can also improve the accuracy of aggregation methods based on unweighted averaging. Imitating successful individuals leads to more accurate judgments (King et al., 2012), and discussions in dyads also improve judgments, but only when initial independent judgments are formed (Minson et al., 2017). Moreover, Becker et al. (2017) showed that information about others' judgments is beneficial when this information equally weighs all other judgments (as opposed to overweighing judgments of a single, highly influential individual). Given that individual judgments can be improved by providing judgments of others, sequential collaboration may lead to more accurate judgments. Especially, the finding that imitating successful individuals improves accuracy (King et al., 2012) is relevant for sequential collaboration as contributors may often be presented with the currently best judgment in the sequential chain, which can easily be imitated by not making a change. However, while King et al. (2012) selected the currently most accurate judgment from a large pool of independent judgments, the judgments presented in a sequential chain are not necessarily very accurate, especially if only a few contributors have encountered and edited it.

Sequential collaboration may also benefit from the fact that in group work, not all group members contribute to a given task equally and some do not contribute at all (free-rider effect; Bray et al., 1978) and that group members often contribute less the more they feel that their contribution is dispensable (Kerr & Bruun, 1983). Such effects may also be observed in sequential collaboration, since contributors can maintain a previous judgment when thinking that they cannot substantially improve it. This opt-out mechanism could in turn improve accuracy since giving respondents the possibility to select the questions to be answered improves accuracy of unweighted averaged judgments (Bennett et al., 2018). The fact that contributors can self-select which judgments to adjust may thus lead to a higher accuracy of the resulting judgments. However, this requires that contributors can accurately distinguish which judgments to maintain (assuming they cannot substantially contribute to them) and which judgments to adjust (assuming they can improve the present state of an entry).

Miller and Steyvers (2011) performed a study closely resembling sequential collaboration in a rank-ordering task. Participants were presented either with judgments of previous participants (resulting in a sequential chain of judgments) or with randomly generated rank orders (resulting in independent judgments). They could opt out from answering by accepting the presented rank order, which was the order made by the previous participant. Miller and Steyvers (2011) found that both the group aggregate and the average subject's performance increased for sequential compared to independent judgments. Moreover, the last judgment in such a sequential chain was more accurate than the group aggregate of independent judgments. Forming sequential chains may, thus, lead to improved judgment accuracy and even outperform the aggregation of independent judgments. Moreover, these results suggest that providing an opportunity to opt out can be beneficial both for individual and for group accuracy.

Hypotheses

Based on prior research, we expect that sequential collaboration is an effective method of eliciting and aggregating individual judgments. Our first two hypotheses concern basic assumptions about sequential collaboration:

> *Hypothesis 1:* Over the course of a sequential chain, (1a) the probability of changing a judgment and (1b) the magnitude of change decrease.

> *Hypothesis 2:* Over the course of a sequential chain, the accuracy of the most recent judgment increases.

Given its high accuracy, unweighted averaging can be used as a benchmark for other forms of collaboration. As discussed above, sequential collaboration may profit from the possibility that contributors are not required to adjust the presented information (Bennett et al., 2018), but rather can decide to maintain the presented judgment if they perceive their own judgments to be dispensable (Kerr & Bruun, 1983). Since providing information about judgments from others can improve accuracy (Becker et al., 2017; King et al., 2012; Minson et al., 2017), especially when such information is organized in a sequential chain (Miller & Steyvers, 2011), accuracy of sequential collaboration may exceed that of unweighted averaging:

> *Hypothesis 3:* Sequential collaboration yields more accurate group estimates than unweighted averaging.

Since unweighted averaging is known to yield highly accurate estimates, it is also plausible that sequential collaboration does not yield better but merely similarly accurate estimates. This would already be an important and relevant finding because sequential collaboration may not profit from the central limit theorem for means of independent random variables the same way as unweighted averaging does (Hogarth, 1978). Sequential collaboration does not involve the computation of a mean of individual judgments, and thus, the central limit theorem is not directly applicable (Zhang et al., 2022). However, contributors may integrate the presented information when forming their own judgment (Bonaccio & Dalal, 2006). Such an implicit averaging of the presented and the internally generated judgment could result in increased accuracy due to error cancelation.

To test our hypotheses, we conducted three online experiments (two of which were preregistered) using chains of four to six contributors in sequential collaboration and corresponding

group sizes for unweighted averaging.[1] The materials comprised general knowledge questions with numerical judgments in the first two experiments and geographic maps on which participants had to position cities in the third experiment.

Experiment 1

Method

Materials

We presented 65 difficult general knowledge questions such as "How tall is the Eiffel Tower?" or "When was Leonardo da Vinci born?" to the participants. The questions were taken from an item pool on general knowledge questions (Pohl, 1998) and updated with contemporary information whenever necessary. The median of correctly answered questions was 0.53% (median absolute deviation = 0.78%), indicating that the questions were indeed difficult to answer correctly for participants. All items, their correct numerical answers, and the unit in which the answer had to be given are provided in Appendix Table A1.

Participants

For this online study, 310 German college students participated via a German panel provider. The compensation ranged between 0.60EUR and 1EUR according to the time for participation. To control data quality already during data collection, participants who changed their browser window or switched to other programs more than five times were excluded during participation. Based on the results of the pilot study, we suspected participants to look up answers when more than 10% of the questions were answered correctly. This was the case for three participants who were not considered for building sequences and whose data were excluded for the analysis. Two participants were excluded due to irregular answer patterns in more than 10% of the questions (i.e., answering with number series such as "23456" or answering "0"). Last, two participants were excluded since the same position in a sequential chain was assigned to two participants due to a technical issue. We kept the data of the participant whose data were used throughout the rest of this sequential chain. Our final sample comprised 303 participants, of whom 76.6% were female, 22.8% were male, and 0.7% identified as diverse. The mean age of the sample was 23.8 years ($SD = 3.2$).

Design and Procedure

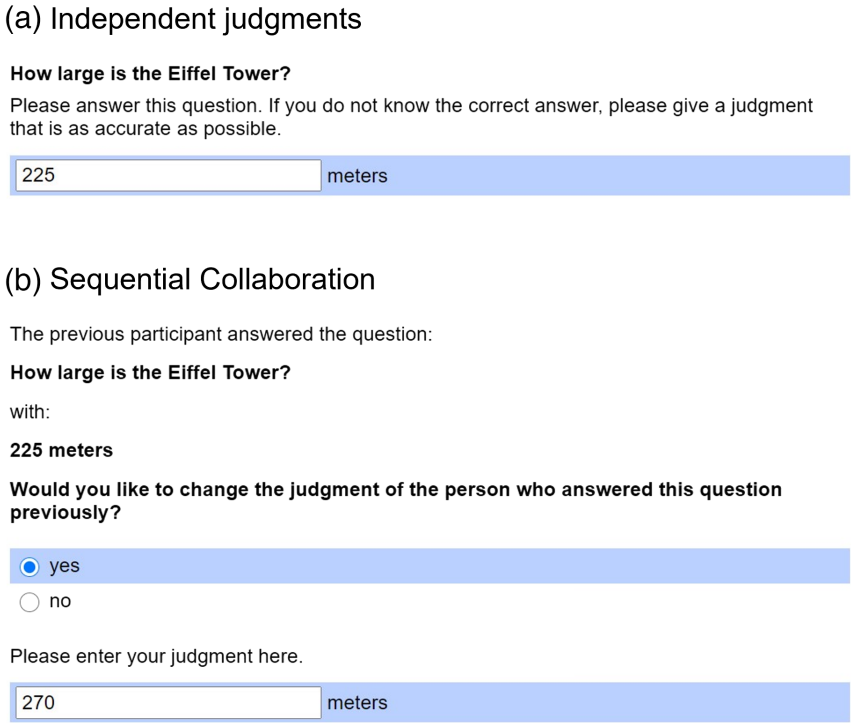
Participants were randomly assigned to the independent-judgments questionnaire (192 participants) or the sequential-collaboration questionnaire (111 participants). After consenting to the study, they were introduced to the corresponding tasks illustrated in Figure 2. In the independent-judgments questionnaire, one general knowledge question per trial was presented to participants who had to type their judgment into a text box before proceeding to the next question. In the sequential-collaboration questionnaire, participants also saw one general knowledge question per trial, but additionally, the answer of a previous participant was shown below the question. Participants decided whether to adjust or maintain the presented judgment. Only in the former case, the text box appeared in which the new judgment could be entered before proceeding to the next question. Participants were informed that the presented judgments were from one or more of the previous participants and did not know their position in the sequential chain.

General knowledge questions were presented in random order, and the unit in which the judgment had to be given was provided directly after the text box. To prevent looking up answers, we implemented a time limit of 30 s for entering a judgment in both questionnaires. Additionally, we implemented a minimum waiting time of 2 s for the sequential-collaboration questionnaire to prevent clicking through the study. After answering all questions, participants provided demographic information, were thanked for participation, and debriefed.

In the unweighted-averaging condition, 155 participants completed the independent-judgments questionnaire, and their judgments were then averaged. In the sequential-collaboration condition, independent initial judgments are required to start sequential chains. Hence, we initialized sequential chains by 37 participants who answered the independent-judgments questionnaire. We used a sequence length of four, meaning that each sequential chain consists of one participant

[1] Prior to conducting the three experiments reported in the present article, we also conducted a pilot study to pretest and improve the experimental paradigm.

Figure 2
Questionnaires Used in Experiments 1 and 2



Note. See the online article for the color version of this figure.

completing the independent-judgments questionnaire followed by three participants completing the sequential-collaboration questionnaire consecutively. For each participant, only the latest judgment in the sequential chain was presented. This procedure resulted in a sample size of 148 participants in the sequential-collaboration condition.

Results

Before analyzing the data, we excluded 314 judgments that were timed out after 30 s. Since sequential chains containing such a judgment were excluded completely, this resulted in the exclusion of 809 judgments in total. Hence, 18,886 judgments remained for analysis.

Since judgments are given on vastly different scales (e.g., from single digits for the length of a soccer goal up to millions for the number of students enrolled in German universities), a standardization of raw judgments is necessary. To this end, we subtracted the correct answer for each question from the raw judgments before dividing the result by the standard deviation of all judgments obtained with the independent-judgments questionnaire. The resulting *standardized errors* are equal to zero for correct judgments while being negative and positive in case of under- and overestimation, respectively. Moreover, we used *absolute standardized errors* for testing hypotheses concerning the accuracy of judgments and estimates.

After standardizing the judgments, we removed the 1% most extreme values from the data as these judgments may distort the results. Excluding outliers across conditions is also recommended by André (2022), who demonstrated that excluding outliers separately within each condition can increase false-positive rates. We identified 189 extreme judgments with this procedure. Again, we excluded both these judgments and the corresponding sequences, resulting in a final sample of 18,626 judgments.

Confirmatory Analyses

Hypothesis 1a states that change probability decreases over the course of a sequential chain. To test this prediction, we only considered data of participants completing the sequential-collaboration questionnaire who could decide whether to change or maintain a presented judgment (Position 2, 3, or 4 in a sequential chain). We modeled the decision of whether to adjust or maintain a judgment as a function of the chain position in the sequence by

fitting a generalized linear mixed model with the R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). Since the dichotomous dependent variable can only be 1 (adjust) or 0 (maintain), we used a logit link function. As every participant answered the same 65 items, and we added random intercepts for items and participants to account for the nested structure of our data (Pinheiro & Bates, 2000). We added crossed random effects of items and participants to all hierarchical models reported since all reported studies have a data structure in which individual judgments are nested in items and participants. Last, we set polynomial contrasts to test for a decline in change probability with increasing chain position.[2]

Figure 3 shows the mean change probability for each chain position with corresponding between-subjects error bars and violin and box plots indicating the distribution of change probabilities for participants (aggregated across items). Even though descriptively in line with Hypothesis 1a, the linear trend of the effect of chain position on change probability was not significant ($\beta = -0.311$, CI = [−0.685, 0.063], $z = -1.629$, $p = .103$).

Hypothesis 1b states that change magnitude decreases over the course of a sequential chain. Again, we only used judgments from the sequential-collaboration questionnaire (Chain Positions 2, 3, and 4) and excluded all trials in which participants did not change the presented judgment of a previous participant. Change magnitude was computed as the absolute difference between the standardized error of a judgment and the standardized error of the previous judgment. We fitted a linear mixed model with change magnitude as a continuous dependent variable and chain position as independent variable and included polynomial contrasts for the factor chain position. In line with Hypothesis 1b, the model revealed a significant negative linear trend of chain position, $\beta = -0.042$, CI = [−0.074, −0.011], $t(82.093) = -2.633$, $p = .010$. Figure 4 shows the empirical means of change magnitude for each chain position with between-subjects error bars, as well as violin and box plots of the distribution of change magnitudes for participants (aggregated across items).

Hypothesis 2 states that judgments become more accurate over the course of a sequential chain. To test this prediction, we only considered data from the sequential-collaboration condition. We fitted a linear mixed model with absolute standardized errors as dependent variable and chain position as independent variable. Since fixed-effect coefficients in linear mixed models have been shown to be robust when residuals are not normally distributed (LeBeau et al., 2018; Schielzeth et al., 2020), we did not transform the dependent variable accordingly. Furthermore, we set polynomial contrasts for the factor chain position.

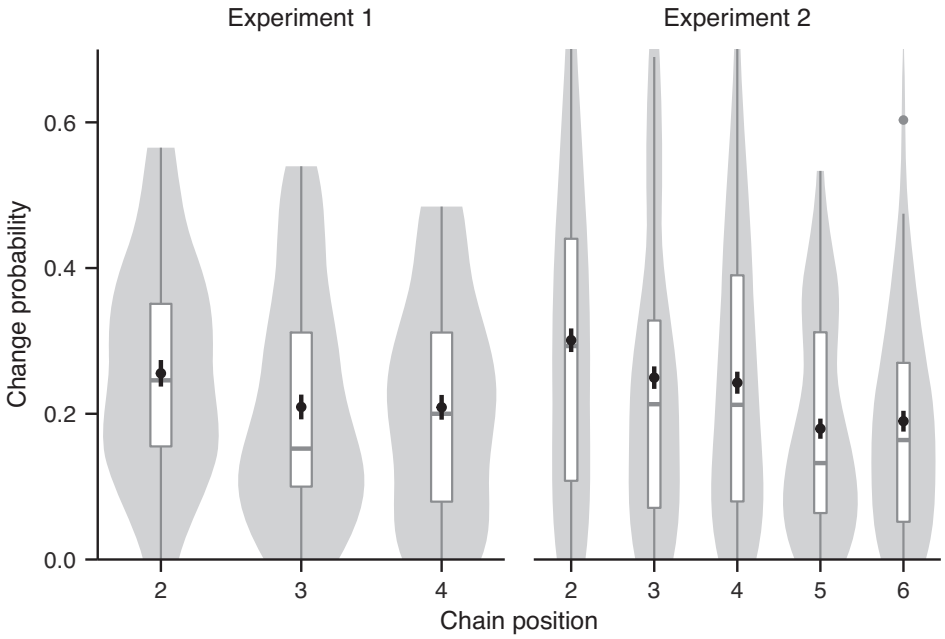
Figure 5 shows the mean absolute standardized errors for each chain position with corresponding violin and box plots. In line with Hypothesis 2, judgments become more accurate as the distance to the correct answer declines over the course of a sequential chain. This pattern was also confirmed by the linear mixed model showing a significant negative linear trend, $\beta = -0.024$, CI = [−0.036, −0.011], $t(143.55) = -3.800$, $p < .001$.

Hypothesis 3 states that sequential-collaboration estimates are more accurate than estimates obtained with unweighted averaging. Before testing this hypothesis, we checked whether the randomization worked as intended, meaning that participants completing the independent-judgments questionnaire had the same judgment accuracy irrespective of whether they were included in the unweighted-averaging condition or served as starters for sequential chains (Position 1) in the sequential-collaboration condition. We only considered data obtained with the independent-judgments questionnaire and fitted a linear mixed model with the absolute standardized error as dependent and condition as independent variable. We did not find a significant effect of condition on the absolute standardized error, $\beta = 0.005$, CI = [−0.012, 0.023], $t(198.37) = 0.598$, $p = .550$, indicating that the randomization worked as intended.

Next, we computed the accuracy of the group estimates for each condition. For sequential collaboration, the estimate for each chain is the judgment at the last chain position (Position 4). Accuracy of these estimates is defined as the absolute standardized error of the last judgment in a sequential chain. For unweighted averaging, the group estimate is the mean of judgments for groups of four participants. We computed the

[2] Whenever testing polynomial contrasts, we only report results for the linear trend since we are interested in a decrease in change probability (Hypothesis 1a), in change magnitude (Hypothesis 1b), and in absolute error (Hypothesis 2). Other trends are only reported when being statistically significant.

Figure 3
Change Probability Within a Sequential Chain



Note. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of change probabilities for the participants (aggregated across items).

absolute standardized errors of these estimates by (a) randomly assigning participants to virtual groups of four, (b) averaging the four standardized errors for each question, and (c) computing the absolute value of the average.[3] Since the number of participants in the unweighted-averaging condition (i.e., 155 participants) is not a multiple of four, we randomly selected one participant whose data were duplicated before generating the virtual groups. This procedure resulted in absolute standardized errors of estimates of the 65 items presented in the study for 39 groups of participants in the unweighted-averaging condition and 37 sequences of participants in the sequential-collaboration condition.

To test Hypothesis 3, we fitted a linear mixed model with absolute standardized errors as dependent variable and condition as independent variable. Figure 6 displays the mean absolute standardized error and corresponding 95% between-subjects confidence intervals with violin and box plots referring to the distribution of group accuracy (aggregated across items). Contrary to Hypothesis 3, estimates in the sequential-collaboration showed a slightly higher mean absolute error of estimates, meaning that these judgments were descriptively less accurate. However, the linear mixed model did not show a significant effect of condition, $\beta = 0.008$, CI = [−0.005, 0.022], $t(72.15) = 1.237$, $p = .220$.

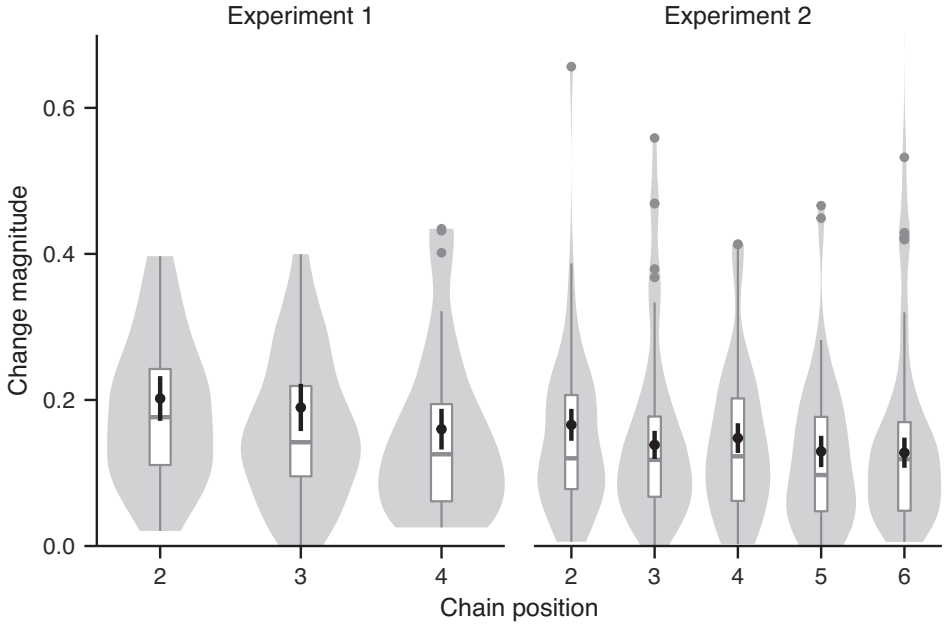
To explore the robustness of our results, we conducted additional analyses examining the influence of outlier exclusions and transformations of judgments on the accuracy of unweighted averaging and sequential collaboration. We also performed nonparametric, descriptive assessments of the hypotheses using the common-language effect size and observation-oriented modeling. The additional analyses led to similar conclusions as those reported in the main text and can be found in the Supplemental Material at https://osf.io/96nsk/.

Exploratory Comparison of Different Aggregation Methods

We additionally assessed the effect of various aggregation methods for obtaining group estimates. For independent judgments, we computed

[3] To check the robustness of the grouping, we computed the mean difference in absolute standardized errors of estimates between conditions and the corresponding linear mixed model for 100 different random groupings in the unweighted-averaging condition. The mean difference in absolute estimates was 0.007 ($SD = 0.001$) for Experiment 1 and 0.01 ($SD = 0.001$) for Experiment 2. The results of the linear mixed model remained the same for all 100 comparisons in both experiments.

Figure 4
Change Magnitude Within a Sequential Chain



Note. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of change magnitude for the participants (aggregated across items).

the mean and the median of the four judgments within each group. For sequential collaboration, we did not only consider the last value in a chain but also aggregated the four individual judgments in a chain using the mean and the median. Furthermore, we computed a weighted mean with the weights 1/10, 2/10, 3/10, and 4/10 for judgments at Chain Positions 1, 2, 3, and 4, respectively.

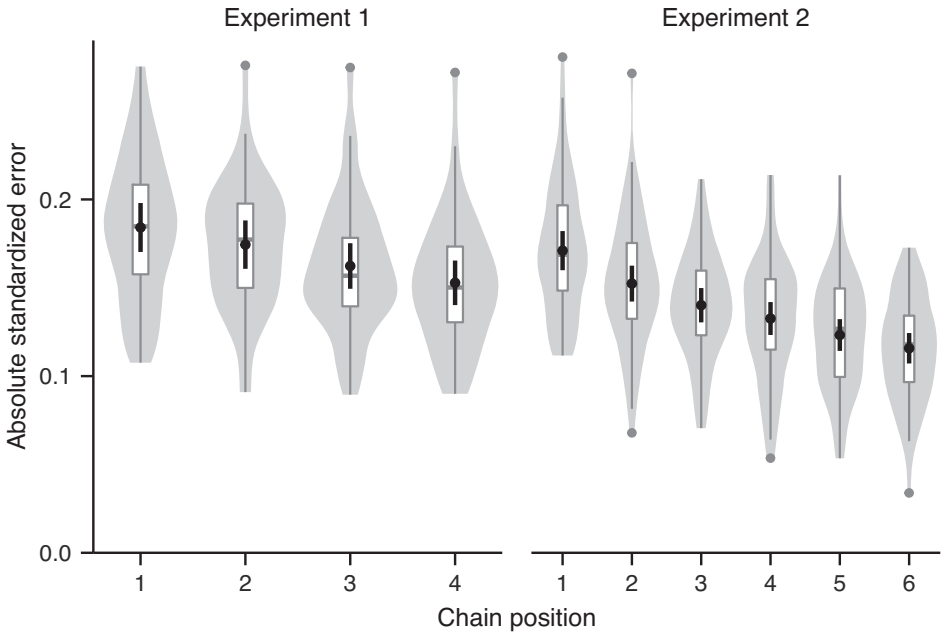
For all aggregation methods and both conditions, Table 1 shows the mean absolute error and the mean squared error of the group estimates. The most accurate estimates were obtained when using median aggregation for independent judgments. Similar to the confirmatory analysis of Hypothesis 3, estimates obtained with unweighted averaging were descriptively more accurate than those of sequential collaboration irrespective of the aggregation method for both absolute and squared errors. For sequential collaboration, all aggregation methods yielded similar accurate results, which were overall less accurate than aggregating independent individual judgments.

Discussion

Overall, Experiment 1 yielded mixed results. The results indicate a basic assumption about sequential collaboration holds, namely, that contributors can improve the accuracy of previous judgments by correcting each other sequentially. There was, however, no evidence for Hypothesis 3, as unweighted averaging and sequential collaboration yielded similarly accurate group estimates. Furthermore, the exploratory moderator analyses showed that the accuracy of sequential collaboration was not much influenced by different aggregation methods, whereas independent judgments showed more accurate estimates when taking the median rather than the mean. This may be due to reducing the effect of extreme judgments, which may distort the mean.

While Experiment 1 provides first insights into sequential collaboration, it has some limitations that restrict the generalizability of our results. First, the sample was restricted to college students who typically have a similar age and educational background. The limited diversity in individuals' expertise might have reduced the chances of improving previous judgments in sequential collaboration since the distribution of knowledge might have been too homogeneous. Second, we implemented a rather short chain length of four individuals. Results might differ when using longer chains since additional contributors might improve the sequential-collaboration estimate substantially.

Figure 5
Accuracy of Judgments Within a Sequential Chain



Note. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of errors for the participants (aggregated across items).

Experiment 2

To address the limitations of Experiment 1, we conducted a second experiment using the same general knowledge questions while increasing the chain length from four to six and collecting an adult sample with no restrictions in age or education. Thereby, we test the robustness of our findings,

Table 1
Comparison of Different Aggregation Methods for Independent Judgments and Sequential Collaboration

Experiment	Method	Aggregation	Absolute error	*SE*	Squared error	*SE*
Experiment 1	Sequential collaboration	Last value	0.16	0.01	0.11	0.01
		Mean	0.16	0.01	0.11	0.01
		Median	0.17	0.01	0.12	0.01
		Weighted mean	0.16	0.01	0.11	0.01
	Independent judgments	Mean	0.16	<0.01	0.08	<0.01
		Median	0.14	<0.01	0.07	<0.01
Experiment 2	Sequential collaboration	Last value	0.12	<0.01	0.07	0.01
		Mean	0.13	<0.01	0.08	0.01
		Median	0.13	<0.01	0.08	0.01
		Weighted mean	0.13	<0.01	0.08	0.01
	Independent judgments	Mean	0.13	<0.01	0.07	<0.01
		Median	0.13	<0.01	0.08	<0.01
Experiment 3	Sequential collaboration	Last value	46.17	1.01	4120.77	208.22
		Mean	46.00	0.86	3615.98	152.01
		Median	44.90	0.95	3820.80	184.06
		Weighted mean	43.33	0.85	3338.07	151.41
	Independent judgments	Mean	53.74	0.71	4205.06	114.84
		Median	47.95	0.74	3723.38	131.24

Note. *SE* = standard error.

especially the improvement of judgments within a sequential chain, and apply the paradigm to a more diverse sample and a longer sequential chain. The design and confirmatory analyses were preregistered.[4] Note that we did not preregister Hypotheses 1a and 1b concerning the change probability and change magnitude within a sequential chain. Moreover, we improved the exclusion criteria for extreme judgments as already applied in Experiment 1 and added an exploratory analysis comparing different aggregation methods for independent judgments and sequential collaboration. All analyses concerning the accuracy of estimates (Hypotheses 2 and 3) are again based on absolute standardized errors similar as in Experiment 1.

Method

Materials, Design, and Procedure

Experiment 2 used the same design and questions as Experiment 1 (see Figure 2 and Table A1, respectively) with some minor adjustments. Since the sample was not restricted in age, we extended the time limit for answering a question from 30 to 40 s. Moreover, we implemented a chain length of six, meaning that the first participant in a sequential chain answered the independent-judgments questionnaire and was then followed by five participants answering the sequential-collaboration questionnaire.

Participants

A German panel provider sampled 686 participants for this study. Participants were compensated between 0.75EUR and 1EUR depending on the time for study completion. During data collection, 21 participants were suspected to look up answers, as they entered more than 10% correct answers, and were thus excluded both for building sequential chains and for the analysis. Moreover, eight participants had irregular answer patterns and were excluded. One participant was excluded since the position in the sequential chain was allocated to two different participants. After excluding these participants and, if necessary, participants in the corresponding sequential chains, the final sample comprised 646 participants. Half of the participants were females (49.9%), the mean age was 48.1 years (SD = 19.5). Most participants had a college degree (27.2%), followed by a high-school diploma (25.5%), and vocational education (23.1%), whereas 24.2% of the participants had a lesser education attainment.

Results

As preregistered, we first excluded judgments (and corresponding chains) that were timed out after 40 s. After this exclusion, 40,324 out of 41,990 judgments remained. The judgments were then standardized itemwise similarly as in Experiment 1 by subtracting the correct answer and dividing by the standard deviation of all independent judgments. Finally, the 1% most extreme judgments (and corresponding chains) were excluded from the data, resulting in 39,699 judgments for the analysis. We conducted the same confirmatory and exploratory analyses as for Experiment 1.

Confirmatory Analyses

To test whether the change probability decreases over the course of a sequential chain (Hypothesis 1a), we fitted a generalized linear mixed model with the decision whether to adjust or maintain a judgment as dependent variable. Figure 3 displays the distribution of change probabilities for each chain position. As hypothesized, the change probability decreased over the course of a sequential chain as indicated by a significant negative linear trend ($\beta = -0.581$, CI = [−0.982, −0.181], $z = -2.843$, $p = .004$).

Next, we fitted a linear mixed model to test whether change magnitude decreases over the course of a sequential chain (Hypothesis 1b). Figure 4 shows that change magnitude slightly decreased over the course of a sequential chain. However, we did not find a significant negative linear trend, $\beta = -0.016$, CI = [−0.040,0.009], $t(174.228) = -1.249$, $p = .214$.

To test whether accuracy of judgments increases over the course of a sequential chain (Hypothesis 2), we estimated a linear mixed model with absolute standardized errors as dependent variable. As predicted, the model revealed a significant negative linear trend of chain position, $\beta = -0.045$, CI = [−0.054, −0.036], $t(290.96) = -9.528$, $p < .001$. Accordingly, Figure 5 shows a decrease of

[4] The preregistration form is available at https://aspredicted.org/8vn9m.pdf.

absolute standardized errors over the course of a sequential chain.

Before testing whether sequential collaboration yielded more accurate estimates than unweighted averaging (Hypothesis 3), we checked whether the randomized assignment to conditions was successful. We analyzed only data obtained with the independent-judgments questionnaire and fitted a linear mixed model with absolute standardized error of individual judgments as dependent variable and condition as independent variable. The model did not show a significant effect of condition on accuracy, $\beta = 0.003$, CI = [−0.011,0.017], $t(390.99) = 0.385$, $p = .701$. This indicates that no condition had an a priori advantage in judgment accuracy.

The error of group estimates in both conditions was computed as already established in Experiment 1. Since sequential chains had six participants, virtual groups for unweighted averaging were also composed of six participants (see Footnote 3). We then fitted a linear mixed model with the absolute standardized error of the group estimate as dependent variable. Figure 6 shows that sequential-collaboration estimates were slightly more accurate than those obtained with unweighted averaging. This impression was confirmed by a linear mixed model, showing a significant difference of the absolute standardized error across conditions, $\beta = 0.014$, CI = [0.005,0.023], $t(100.71) = 3.067$, $p = .003$.

Exploratory Comparison of Different Aggregation Methods

Table 1 shows a comparison of different aggregation methods for sequential collaboration and independent judgments. In contrast to Experiment 1, for absolute errors, sequential collaboration provided the most successful aggregation of individual judgments when using the last judgment in a sequence. However, when focusing on squared instead of absolute errors, taking the mean in unweighted averaging was similarly accurate as the last value of a sequential chain. Overall, the differences between conditions and aggregation methods are small, and there seems to be no clear advantage in the accuracy of estimates between unweighted averaging and sequential collaboration. Further robustness analyses can be found in the Supplemental Material.

Discussion

Experiment 2 also showed that a basic assumption of sequential collaboration holds, namely, that the accuracy of judgments increases through incremental changes. Moreover, sequential-collaboration estimates were more accurate than estimates obtained with unweighted averaging in the confirmatory analysis. However, the comparison of different aggregation methods across conditions revealed that there are only small differences in accuracy between sequential collaboration and unweighted averaging.

Experiments 1 and 2 investigated the accuracy of sequential collaboration when eliciting quantitative judgments. However, both experiments used general knowledge questions, which limit the generalizability of the results and may pose some issues. First, the questions are prone to extreme judgments. For instance, one participant answered 120,000,000,000,000,000 km to the question "How long is the mean distance between Earth and Moon?" for which the correct answer is 384,400 km. Having extreme judgments in the data might especially hurt the performance of unweighted averaging. Furthermore, general knowledge questions occur rather seldom in online collaboration projects, thus limiting the ecological validity of the conclusions. Thus, we conducted a conceptual replication using different materials, which are less prone to extreme judgments and more closely resemble actual online collaboration projects.

Experiment 3

Experiment 3 is a conceptual replication of Experiment 1. Both studies used a similar design with some minor changes due to the different materials. Instead of general knowledge questions, participants were presented with geographic maps on which they had to locate the positions of different cities. We thus focus on two-dimensional location judgments (i.e., *x*- and *y*-coordinates) rather than one-dimensional numerical judgments. In contrast to general knowledge questions, two-dimensional location judgments on geographical maps are naturally constrained by the size of the map (more precisely, by the maximum distance between the correct location and all possible judgments), which limits the maximum range of extreme judgments. We preregistered Experiment

3 at www.aspredicted.org.[5] Going beyond the preregistration, we also tested whether change probability (Hypothesis 1a) and change magnitude (Hypothesis 1b) decrease over the course of a sequential chain. Furthermore, we adjusted the outlier analysis as described in Experiments 1 and 2 and added an exploratory analysis comparing different aggregation methods.

Method

Participants

We recruited 417 adult participants via a commercial German panel provider, which compensated participants according to the time for completing the study. Since participants were presented with maps, they were supposed to only participate using a computer. Due to issues in the recruitment of participants by the panel provider, 39 participants were nonetheless able to access and complete the study using mobile devices. We excluded all participants using mobile devices and all sequences including these participants, thus excluding 70 participants in total. Additionally, four participants were able to access and complete the study a second time. Therefore, we excluded the data collected at the second participation. Since two of these participants were assigned to the sequential-collaboration condition for their second participation, and sequences were built based on their judgments, we excluded another 10 participants in total. We also checked whether participants looked up the correct answers or whether participants clicked at a similar position for all items and identified one participant who was suspected to look up answers who was thus excluded. The final sample comprised 333 participants, of whom 46.0% were females. The mean age was 45.5 years ($SD = 15.2$). Participants had a diverse educational background, with 35.4% holding a college degree, 24.9% having a high-school diploma, 24.0% having vocational education, and 18.3% having a lesser educational attainment.

Materials

As stimulus material, we selected seven maps displaying different European countries: (a) Italy; (b) France; (c) Germany; (d) United Kingdom and Ireland; (e) Austria and Switzerland; (f) Spain and Portugal; and (g) Poland, Czech, Hungary, and Slovenia. All maps were on a scale of 1:5,000,000 with an image resolution of 800 × 500 pixels. Regarding the available geographic information, the maps only showed land mass, oceans, and country borders. The countries of interest were colored white, whereas all other countries were colored gray; oceans were colored blue and country borders were represented as black lines. Overall, we selected 57 cities across all seven maps. For each map, we selected between 4 and 17 cities while considering the expected geographic knowledge of German participants. Appendix Table B1 provides a comprehensive overview of the materials, and all maps are also available in the Supplementary Material (https://osf.io/96nsk/).

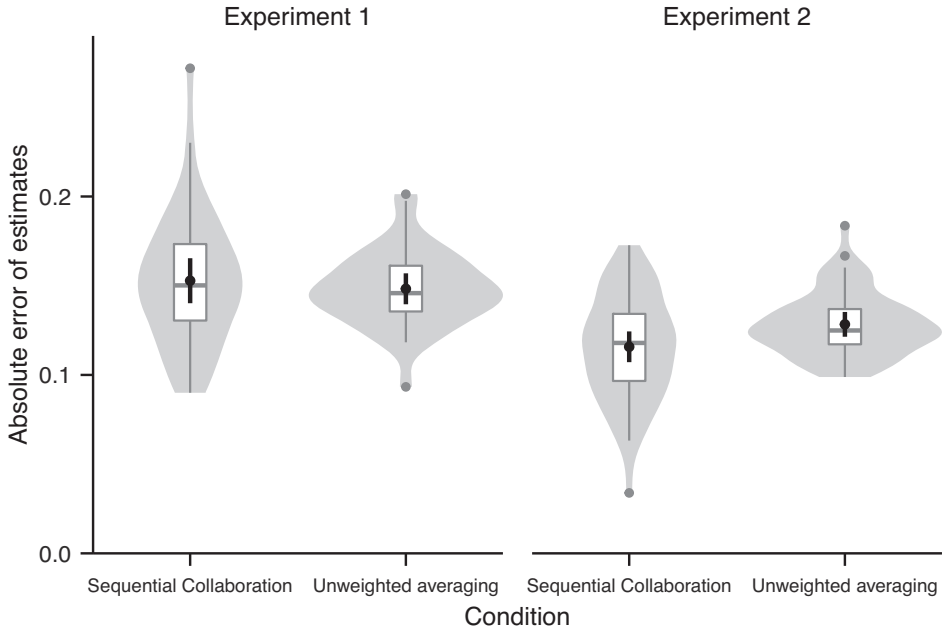
Design and Procedure

We randomly assigned participants to either the sequential-collaboration questionnaire (112 participants) or the independent-judgments questionnaire (221 participants). As in Experiment 1, we formed sequences of four participants, meaning that one participant who answered the independent-judgments questionnaire started a sequential chain followed by three participants who completed the sequential-collaboration questionnaire. This resulted in 183 participants in the unweighted-averaging condition and 150 participants in the sequential-collaboration condition.

After being informed about the aim of the study and providing informed consent, participants were instructed about the task. In the independent-judgments questionnaire, participants had to indicate the position of the given cities on the presented map as accurately as possible. In the sequential-collaboration questionnaire, participants were provided with the location judgment of a city given by a previous participant. Subsequently, they could choose either to modify the given position by indicating a new position or to directly continue to the next city without changing the presented location judgment. The order in which the seven maps were presented was randomized as was the order of the presented cities within each map. Furthermore, each trial asked about the position of only one city, such that participants provided only a single location judgment before continuing to the next city.

[5] The preregistration form is available at https://aspredicted.org/9j9de.pdf.

Figure 6
Accuracy of Unweighted Averaging and Sequential Collaboration



Note. Estimates for sequential collaboration pertain to the last judgment in a sequential chain. Black points display the empirical means for each condition, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the corresponding distribution of participants (aggregated across items).

Participants were given 40 s to indicate the city's position or to decide to not change the presented position. Additionally, participants completing the sequential-collaboration questionnaire had a waiting period of 2 s before they could continue to the next city. Finally, participants provided demographic information, were debriefed, and were thanked for participation.

Results

As dependent variable for Hypotheses 2 and 3, we computed the Euclidean distance to the correct answer for each judgment.[6] Next, we excluded 225 judgments (and corresponding sequential chains) that were timed out after 40 s, meaning that 18,433 out of 18,981 judgments remained for analysis. We again excluded the 1% most extreme judgments (i.e., 184 judgments) as defined by the distance to the correct answer. After the exclusion of sequential chains that contained outliers, 18,161 judgments remained for analysis.

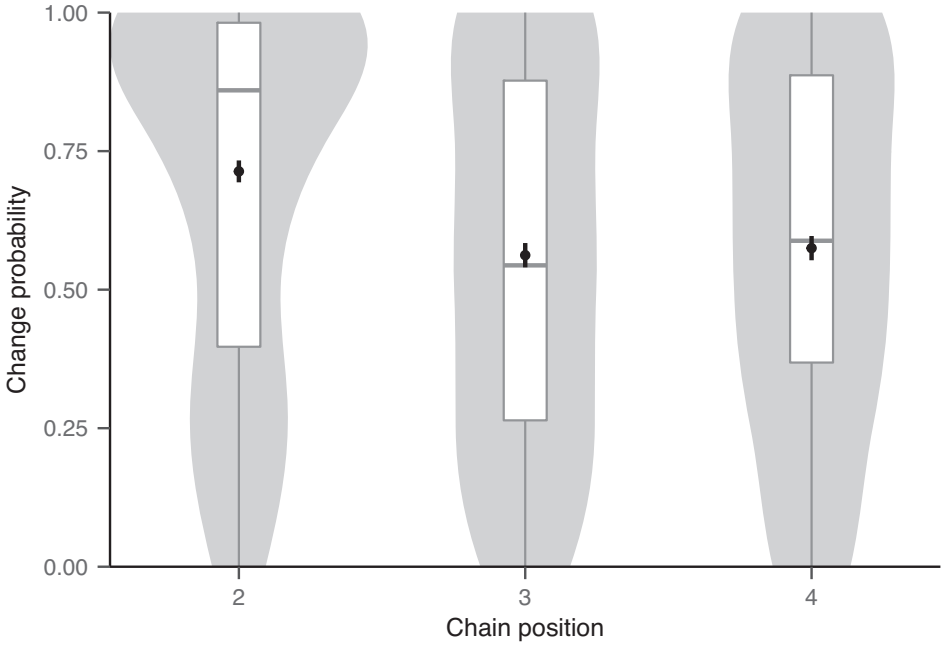
Confirmatory Analyses

To test whether change probability decreases over the course of a sequential chain (Hypothesis 1a), we fitted a generalized linear mixed model with the decision whether a judgment was adjusted or maintained as dependent variable. Figure 7 displays the change probability for each chain position with error bars as well as violin and box plots illustrating the distribution of change probabilities for participants (aggregated across items). In line with Hypothesis 1a, the plot shows a decreasing change probability with increasing chain position. This visual impression was confirmed by the model, which revealed a significant negative linear trend of chain position ($\beta = -0.937$, CI = [−1.845, −0.028], $z = -2.021$, $p = .043$).

To test whether the magnitude of changes decreases over the course of a sequential chain (Hypothesis 1b), we fitted a linear mixed model with change magnitude as dependent variable. Figure 8 shows the empirical mean and distribution of the change magnitude across chain positions. In line with our hypothesis, we found a significant negative linear trend of chain position, $\beta = -14.952$, CI = [−24.139, −5.765],

[6] All hypotheses were also analyzed using the *x*- and *y*-coordinate separately as dependent variables. These analyses yielded the same results as the analysis using Euclidean distances as dependent variable.

Figure 7
Change Probability Within a Sequential Chain in Experiment 3



Note. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of change probabilities for the participants (aggregated across items).

$t(108.971) = -3.190$, $p = .002$. Furthermore, we also found a significant positive quadratic trend, $\beta = 9.507$, CI = [0.279, 18.736], $t(108.984) = 2.019$, $p = .046$, which indicates a larger difference between Positions 2 and 3 than between Positions 3 and 4 (cf. Figure 8).

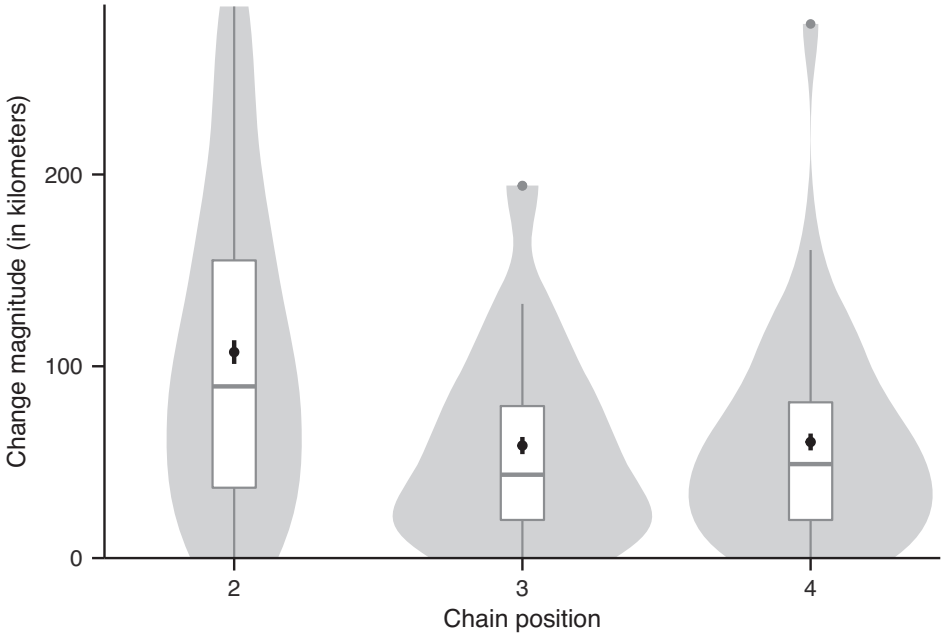
To test whether judgments become more accurate over the course of a sequential chain (Hypothesis 2), we fitted a linear mixed model with chain position as independent variable and Euclidean distance of each judgment to the true position of a city as dependent variable. The model revealed a significant linear trend between chain position and distance, $\beta = -17.610$, CI = [−24.801, −10.419], $t(145.17) = -4.777$, $p < .001$. Furthermore, the quadratic trend was also significant, $\beta = 8.316$, CI = [1.125, 15.507], $t(145.17) = 2.256$, $p = .026$. In combination with the negative linear trend, the positive quadratic trend indicates that accuracy improved more between Positions 2 and 3 compared to Positions 3 and 4, a pattern also displayed in Figure 9. An overview of the judgments given by the participants for each city can be found in the Supplementary Material.

Before comparing the accuracy of sequential collaboration and unweighted averaging (Hypothesis 3), we again performed a randomization check. To compare judgment accuracy in the independent-judgments questionnaire across conditions, we estimated a linear mixed model with the Euclidean distance to the true position as dependent variable. We did not find a significant difference in accuracy between judgments in the unweighted-averaging condition and judgments that were used to start sequential chains in the sequential-collaboration condition, $\beta = -0.768$, CI = [−10.691, 9.160], $t(220.25) = -0.152$, $p = .879$, thus indicating that the randomization was successful.

We computed group estimates for each condition similar as in Experiments 1 and 2. Estimates obtained with independent judgments were based on random groups of four participants. We averaged the four location judgments separately for each coordinate, thereby computing the geometric center. For sequential collaboration, estimates pertain to the last judgment in each chain. As a dependent variable, we computed the Euclidean distance between the resulting mean estimate and the true position of each city.

Figure 10 displays the mean distance to the true position with corresponding error bars and violin and box plots indicating the distribution of mean distances for different the two experimental

Figure 8
Change Magnitude Within a Sequential Chain in Experiment 3



Note. Distances were converted from pixels to kilometers for this figure. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of change magnitude for the participants (aggregated across items).

groups (aggregated across items). In line with Hypothesis 3, sequential-collaboration estimates resulted in a smaller distance to the true position than estimates obtained with unweighted averaging. This impression was supported by a linear mixed model with the Euclidean distance as dependent variable, showing that sequential collaboration yielded more accurate estimates than unweighted averaging, $\beta = -7.411$, CI = [−14.532, −0.301], $t(80.70) = -2.049$, $p = .044$.

Figure 11 illustrates the high accuracy of sequential collaboration for five cities on the map of Italy. The figure shows the mean estimates of the two methods as well as the actual positions of the five cities. For Florence, Milan, Rome, and Venice, sequential collaboration yielded more accurate estimates than unweighted averaging, while both methods yielded similarly accurate estimates for Naples. Across all maps and cities, sequential collaboration resulted in location estimates that were 16.03 km closer to the actual position than those obtained via unweighted averaging (see Figure 10). In Figure 11, this value resembles the difference of the average distance of the two estimates for the location of Rome (13.55 km). Similar plots for the other six maps are available in the Supplemental Material.

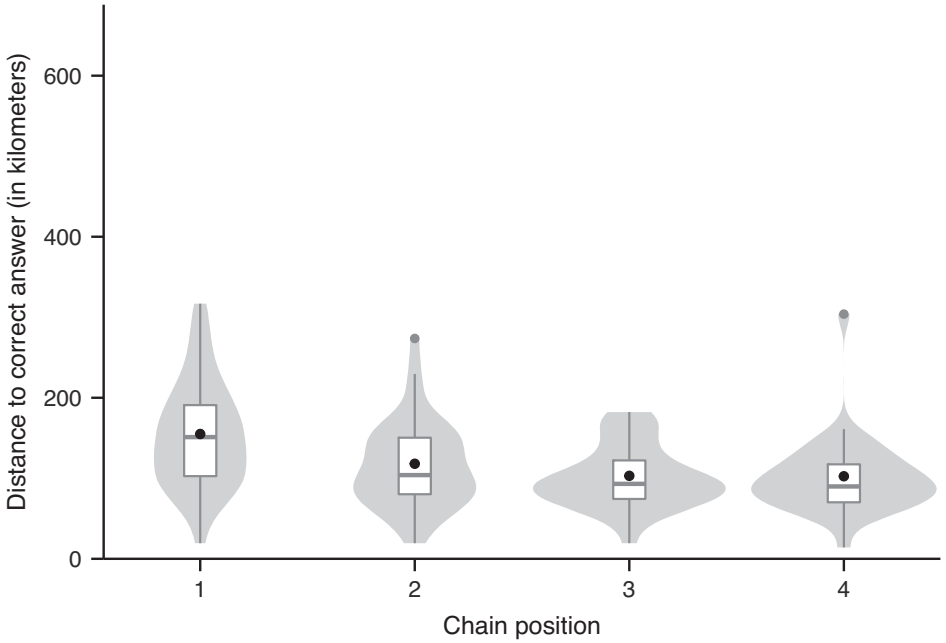
Exploratory Comparison of Different Aggregation Methods

Similarly, as in Experiments 1 and 2, we compared different aggregation methods for sequential collaboration and unweighted averaging. Table 1 shows that sequential collaboration generally yielded more accurate estimates than unweighted averaging when focusing on the Euclidean distance. However, sequential-collaboration estimates yielded similar accurate estimates when using the mean, median, or the weighted mean. When measuring accuracy in terms of squared Euclidean distance, median aggregation of independent judgments was more accurate than the last judgment in sequential collaboration, but all other aggregations for sequential collaboration remained more accurate.

Discussion

Experiment 3 replicated the results of Experiments 1 and 2 using geographic maps instead of general knowledge questions. Sequential collaboration yielded more accurate estimates over the course of a sequential chain, whereas change probability and change magnitude of judgments

Figure 9
Accuracy of Judgments Within a Sequential Chain in Experiment 3



Note. Distances were converted from pixels to kilometers for this figure. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of distances to the correct answers for the participants (aggregated across items).

decreased. Additionally, sequential collaboration yielded more accurate results than unweighted averaging in the confirmatory analysis. However, the exploratory analysis revealed that various methods of aggregating judgments into estimates for sequential collaboration yielded similar accurate estimates.

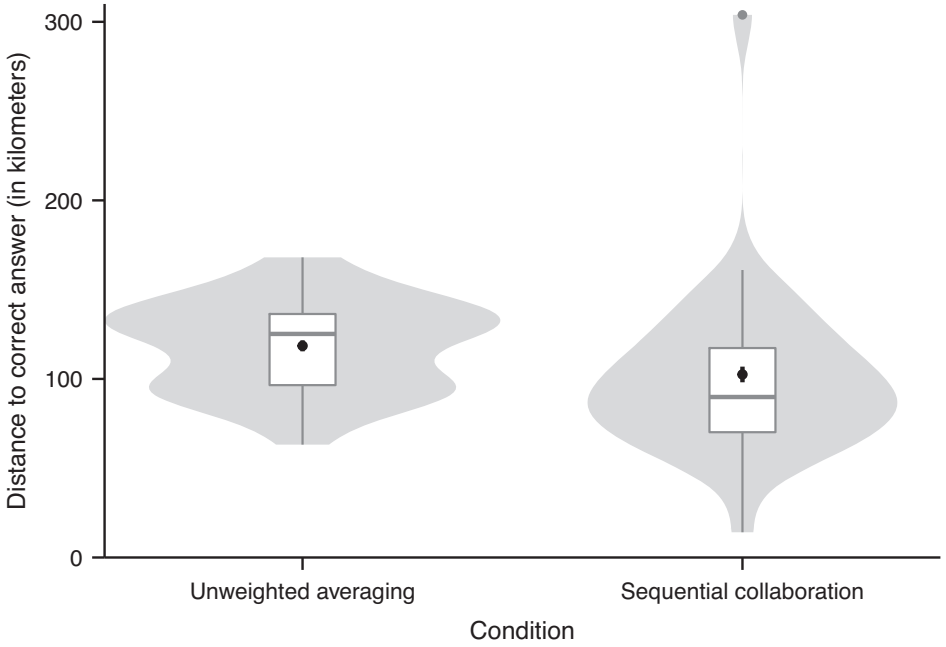
General Discussion

Sequential collaboration describes a collaboration method in which contributors form a sequential chain of judgments by deciding whether to adjust or maintain the latest judgment provided by a previous contributor. In three online studies using general knowledge questions and geographic maps, we examined whether change probability and change magnitude decrease over the course of a sequential chain (Hypotheses 1a and 1b, respectively), whereas judgment accuracy increases (Hypothesis 2). As a benchmark, we compared the accuracy of estimates obtained with sequential collaboration to estimates obtained with unweighted averaging (Hypothesis 3). All three experiments provided evidence that accuracy increased within sequential chains of judgments, whereas a decrease in change probability was observed only in Experiments 2 and 3, and a decrease in change magnitude only occurred in Experiments 1 and 3. Sequential collaboration outperformed unweighted averaging only in Experiments 2 and 3, whereas showing a similar level of accuracy in Experiment 1. While this pattern did not remain as clear in the exploratory analysis showing mixed patterns for the comparison with different aggregation measures, it is noteworthy that sequential collaboration and unweighted averaging overall performed similarly well in terms of accuracy.

The present work contributes to research on how judgments are influenced when providing information about the judgments of others. Several studies have already hinted toward dependent judgments being beneficial for individual judgments in certain situations (Becker et al., 2017; King et al., 2012; Koehler & Beauregard, 2006; Minson et al., 2017), with Miller and Steyvers (2011) implementing a design quite similar to sequential collaboration. We extend this line of research by showing that even a very high level of dependency of judgments can yield accurate estimates. The observed improvement in accuracy within chains of judgments indicates that sequential collaboration was neither obstructed

Figure 10

Accuracy of Estimates Obtained With Unweighted Averaging and Sequential Collaboration in Experiment 3



Note. Distances were converted from pixels to kilometers for this figure. Estimates for sequential collaboration pertain to the last judgment in a sequential chain. Points display the empirical means for each chain position, error bars show the corresponding 95% between-subjects confidence intervals. Violin and box plots illustrate the distribution of distances to the correct answers for the participants (aggregated across items).

by anchoring effects (Mussweiler et al., 2004; Tversky & Kahneman, 1974) nor by high rates of inaccurate changes, for instance, due to egocentric discounting (Bonaccio & Dalal, 2006). Our results show that sequential collaboration provides accurate estimates in controlled online studies involving quantitative knowledge and location judgments. This is in line with prior research showing that large-scale online collaboration projects that rely on this basic mechanism provide high-quality information (e.g., Leithner et al., 2010; Zielstra & Zipf, 2010).

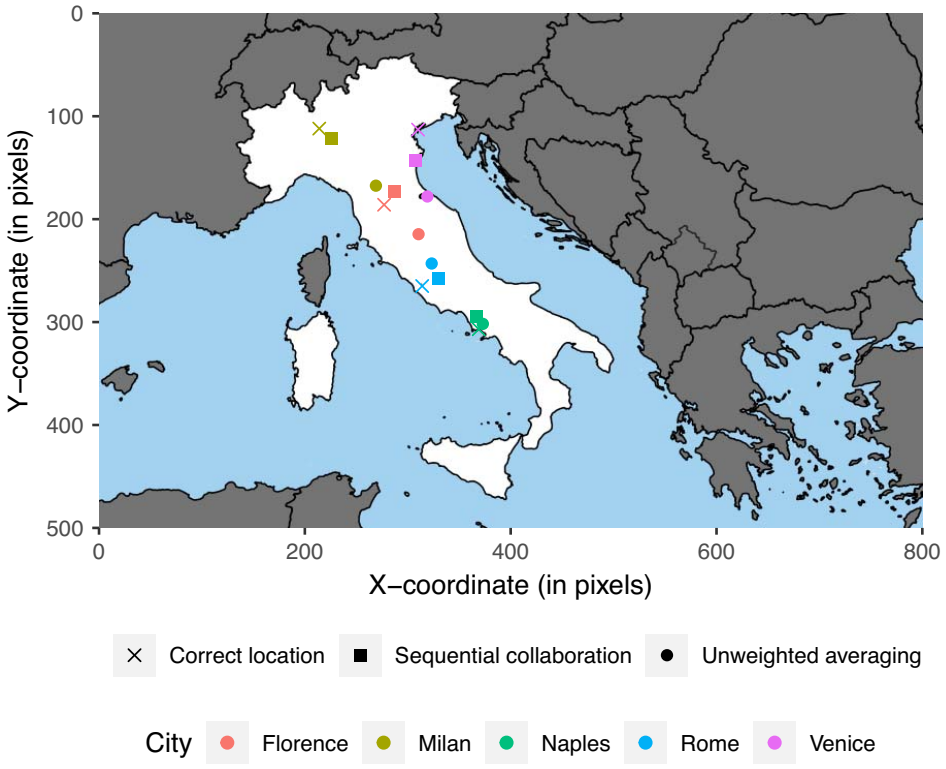
Moderators Across and Within Experiments

The mixed results across experiments may be due to several factors relevant for the performance of sequential collaboration and unweighted averaging. First, while Experiment 1 used a homogeneous sample of university students, Experiments 2 and 3 used more diverse samples (i.e., German adults with a wide range in age and educational background). Limited diversity can not only reduce the performance of unweighted averaging (Davis-Stober et al., 2014; de Oliveira & Nisbett, 2018). Sequential collaboration may also generally benefit from diverse samples since heterogeneity in knowledge increases the chances that a few experts provide accurate corrections to the judgments of the remaining, less knowledgeable contributors. Hence, participants may had fewer chances of correcting each other in Experiment 1 compared to the more diverse samples in Experiments 2 and 3.

Second, the inconsistent pattern of results may be due to the difficulty of the different tasks. Whereas the general knowledge questions in Experiments 1 and 2 were very difficult, locating cities on geographical maps (Experiment 3) was easier because participants likely had some basic geographic knowledge (e.g., from school, the media, or by visiting some of the cities). The difficulty of the tasks can explain the substantial difference in change probabilities (approximately 20% in Experiments 1 and 2 compared to 60% in Experiment 3). Higher change probabilities indicate that participants in the sequential-collaboration condition were more likely to improve judgments within a sequential chain.

Third, the experiments differed in the length of the sequential chains (i.e., the group size varied

Figure 11
Estimated and Actual Locations of Five Cities on the Map of Italy



Note. All maps presented in this experiment were generated using QGIS 3 (https://www.qgis.org/en/site/index.html) licensed under Creative Commons CCBY-SA. To illustrate the material, this map of Italy resembles the one presented in the experiment. See the online article for the color version of this figure.

between 4 and 6). Sequential collaboration may require longer chains, especially for more difficult items, because substantial improvements in judgments occur less frequently. In Experiment 1, the high item difficulty combined with the shorter sequential chains of only four contributors may have reduced the performance of sequential collaboration.

Fourth, certain tasks and materials may be better suited either for unweighted averaging or for sequential collaboration. Unweighted averaging has often been shown to be highly accurate when aggregating numerical point judgments as in Experiments 1 and 2 (Galton, 1907; Hueffer et al., 2013; Surowiecki, 2004) and is also quite robust against biases in such scenarios (Davis-Stober et al., 2014). In contrast, sequential collaboration may be more suitable for more complex tasks, such as positioning cities on maps as in Experiment 3 or generating rank orders (Miller & Steyvers, 2011; Steyvers et al., 2009). In more complex tasks, it is easier for contributors to integrate partial knowledge into the judgment. For instance, one may know that Berlin is located close to the Polish border, and in turn correct the latitude of a presented judgment to the East without modifying any other aspect of the judgment (e.g., longitude). The contribution of partial knowledge is also likely in online collaborative projects such as Wikipedia, where contributors usually edit only small parts of an article while rarely (re)writing complete articles. In complex tasks, mechanical aggregation methods may even lead to unreasonable estimates. For instance, unweighted averaging of two-dimensional location judgments can result in estimates for a city that is located in a lake or an ocean. In contrast, contributors in sequential collaboration usually do not provide such unreasonable judgments.

Fifth, we also compared different aggregation methods. For unweighted averaging, we computed the mean and the median. For sequential collaboration, we used the last value of a sequential chain as well as the mean, the median, and the weighted

mean across chain positions. We found that all aggregation methods yielded similar results for absolute errors in Experiment 1, and for absolute as well as squared errors in Experiment 2. However, in Experiment 1, unweighted averaging resulted in smaller squared errors than sequential collaboration. In Experiment 3, sequential collaboration generally yielded smaller absolute errors than unweighted averaging for all aggregation methods. However, with respect to squared errors, sequential collaboration showed a similar accuracy as unweighted averaging. Overall, different aggregation methods can affect the accuracy of estimates. This may be especially relevant in small groups of only four or six contributors, in which extreme judgments have a large influence on the resulting estimates.

Possible Mechanisms

Our results provide first insights into sequential collaboration, but the three experiments do not provide explanations *why* sequential collaboration yields accurate results. Prior research points to different mechanisms that could lead to improved judgments over the course of a sequential chain. Sequential collaboration may yield accurate results because individual judgments are implicitly weighted by expertise. Both the weighting of judgments by expertise (Budescu & Chen, 2014; Merkle et al., 2020) and the selection of experts based on prior performance (Mannes et al., 2014) improve accuracy when aggregating independent judgments. However, sequential collaboration does neither implement an explicit weighting mechanism nor a selection of individuals based on expertise. Instead, the task structure allows contributors to maintain a judgment if they do not feel they can substantially contribute to the presented judgment. In fact, subjective confidence can be a valid indicator of expertise (Mannes et al., 2014), leading participants to not contribute to group work when they feel their judgment is dispensable (Kerr & Bruun, 1983). Judgments in sequential collaboration are implicitly weighted since experts are more likely to make adjustments, thus adding a larger contribution to the final outcome. In contrast, contributors without expertise are more likely to maintain the presented judgment. In an ideal case, sequential collaboration may lead to improvements in judgments until a correct judgment is no longer changed.

The high accuracy of sequential collaboration may merely be due to the opportunity of contributors to opt out of answering and not due to the dependent nature of sequential judgments. In fact, allowing participants to self-select the subset of questions to be answered improves judgment accuracy and group estimates (Bennett et al., 2018; Galton, 1907). However, typical designs building on unweighted averaging do not allow individuals to select whether to answer a question or not, but rather requires them to answer all questions irrespective of their expertise or metacognitive knowledge (Larrick & Soll, 2006). In our three studies, we compared this typical paradigm to sequential collaboration, meaning that the two experimental conditions did not only differ in the level of dependency but also in the possibility of opting-out of providing a judgment. Future research should further examine the role of the opt-out mechanism in sequential collaboration.

In general, allowing participants to self-select whether to answer a question when applying unweighted averaging to the resulting judgments may lead to quite varying numbers of judgments per question. Bennett et al. (2018) observed that some of the easiest questions were answered by almost all participants, whereas other, more difficult questions were answered not even by a single participant or only very few participants. This can render independent individual judgments with an opt-out option rather uneconomical and inefficient. In contrast, the opt-out mechanism in sequential collaboration ensures a higher efficiency because the decision whether to opt out is made in context of the perceived quality of the presented judgment. Hence, opting-out of adjusting a judgment is informative with respect to the accuracy of the current estimate.

Limitations and Future Research

Our experiments on sequential collaboration have some limitations. First, we only studied a very simple type of sequential collaboration, where contributors could only change or maintain a previous judgment. However, Wikipedia and OpenStreetMap offer several additional functions, such as discussion sites, a board of moderators checking on the contributors' activities, and a history of all changes ever made to an entry. These additional functions are likely to influence contributors' behavior in online

collaborative projects, even though they are less prominent than the information itself, which is directly available in an article or on a map.

Last, our studies used small chains of only four or six contributors. While estimates obtained with unweighted averaging improve for larger crowds, estimates obtained with sequential collaboration may not profit from longer chains. On the contrary, sequential chains may not converge to the correct judgment because this requires that at least one contributor adjusts the presented judgment to be correct. Whereas correct judgments become more likely as chains become longer, they can still be modified by subsequent, less knowledgeable contributors. Moreover, anchoring due to the presented (possibly incorrect) judgments may hinder contributors to provide accurate judgments in the long run. Based on these considerations and our findings, we expect that sequential collaboration has advantages over unweighted averaging in short sequential chains. However, for longer chains of about 12–15 contributors, unweighted averaging would probably outperform sequential collaboration. To address this open question, future research should assess the performance of longer sequential chains.

Conclusion

Sequential collaboration is at the core of many large-scale online collaborative projects, such as Wikipedia or OpenStreetMap. Our studies show that contributors can successfully collaborate through adjusting and maintaining previous judgments of other contributors. More generally, sequential collaboration has a high practical and theoretical relevance and provides a fruitful paradigm for studying how individuals perceive and use information from others in order to decide whether and how to adjust previous judgments.

References

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*(1), 213–223. https://doi.org/10.1037/xge0001069

Arazy, O., Morgan, W., & Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in wikipedia. *16th Annual Workshop on Information Technologies & Systems (WITS) Paper*. https://doi.org/10.2139/ssrn.1025624

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*(26), E5070–E5076. https://doi.org/10.1073/pnas.1615978114

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes*, *103*(1), 121–133. https://doi.org/10.1016/j.obhdp.2006.05.001

Bray, R. M., Kerr, N. L., & Atkin, R. S. (1978). Effects of group size, problem difficulty, and sex on group performance and member reactions. *Journal of Personality and Social Psychology*, *36*(11), 1224–1240. https://doi.org/10.1037/0022-3514.36.11.1224

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. https://doi.org/10.1287/mnsc.2014.1909

Dalkey, N., & Helmer, O. (1963). An experimental application of the DELPHI method to the use of experts. *Management Science*, *9*(3), 458–467. https://doi.org/10.1287/mnsc.9.3.458

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. https://doi.org/10.1037/dec0000004

de Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, *115*(9), 2066–2071. https://doi.org/10.1073/pnas.1717632115

Dennis, A. R. (1996). Information exchange and use in small group decision making. *Small Group Research*, *27*(4), 532–550. https://doi.org/10.1177/1046496496274003

Dennis, A. R., Hilmer, K. M., & Taylor, N. J. (1998). Information exchange and use in GSS and verbal group decision making: Effects of minority influence. *Journal of Management Information Systems*, *14*(3), 61–88.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451. https://doi.org/10.1038/075450a0

Geist, M. R. (2010). Using the delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning*, *33*(2), 147–154. https://doi.org/10.1016/j.evalprogplan.2009.06.006

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

Girres, J.-F., & Touya, G. (2010). Quality assessment of the french OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40–46. https://doi.org/10.1016/0030-5073(78)90037-5

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*, Article 16. https://sjdm.org/~baron/journal/12/12924a/jdm12924a.pdf

Jeste, D. V., Ardelt, M., Blazer, D., Kraemer, H. C., Vaillant, G., & Meeks, T. W. (2010). Expert consensus on characteristics of wisdom: A delphi method study. *The Gerontologist*, *50*(5), 668–680. https://doi.org/10.1093/geront/gnq022

Keck, S., & Tang, W. (2020). Enhancing the wisdom of the crowd with cognitive-process diversity: The benefits of aggregating intuitive and analytical judgments. *Psychological Science*, *31*(10), 1272–1282. https://doi.org/10.1177/0956797620941840

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94. https://doi.org/10.1037/0022-3514.44.1.78

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

King, A. J., Cheng, L., Starke, S. D., & Myatt, J. P. (2012). Is the true "wisdom of the crowd" to copy successful individuals? *Biology Letters*, *8*(2), 197–200. https://doi.org/10.1098/rsbl.2011.0795

Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM conference on computer supported cooperative work* (pp. 37–46). https://doi.org/10.1145/1460563.1460572

Koehler, D. J., & Beauregard, T. A. (2006). Illusion of confirmation from exposure to another's hypothesis. *Journal of Behavioral Decision Making*, *19*(1), 61–78. https://doi.org/10.1002/bdm.513

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), Article e106930. https://doi.org/10.1371/journal.pone.0106930

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

Laughlin, P. R., Bonner, B. L., Miner, A. G., & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80*(2), 103–117. https://doi.org/10.1006/obhd.1999.2848

LeBeau, B., Song, Y. A., & Liu, W. C. (2018). Model misspecification and assumption violations with the linear mixed model: A meta-analysis. *SAGE Open*, *8*(4), 1–16. https://doi.org/10.1177/2158244018820380

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: JAMIA*, *17*(4), 373–374. https://doi.org/10.1136/jamia.2010.004507

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, *16*(1), 54–75. https://doi.org/10.1177/1088868311417243

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

Mayer, M., & Heck, D. W. (2021a). Sequential collaboration: The accuracy of dependent, incremental judgments. *PsyArXiv*. https://doi.org/10.31234/osf.io/w4xdk

Mayer, M., & Heck, D. W. (2021b). Sequential collaboration: The accuracy of dependent, incremental judgments. *OSF*. https://osf.io/96nsk/

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

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*, pp. 1292–1297). https://cogsci.mindmodeling.org/2011/index.html

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2017). The contingent wisdom of dyads: When discussion enhances vs. Undermines the accuracy of collaborative judgments. *Management Science*, *64*(9), 4177–4192. https://doi.org/10.1287/mnsc.2017.2823

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. In R. F. Pohl (Ed.), *Cognitive illusions* (1st ed., pp. 183–199). Psychology Press.

Niederer, S., & van Dijck, J. (2010). Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society*, *12*(8), 1368–1387. https://doi.org/10.1177/1461444810365297

OpenStreetMap Contributors. (2021). *OpenStreetMap*. https://www.openstreetmap.org/about

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

Pohl, R. F. (1998). The effects of feedback source and plausibility of hindsight bias. *European Journal of Cognitive Psychology*, *10*(2), 191–212. https://doi.org/10.1080/713752272

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, *38*(1), 1–15. https://doi.org/10.1086/658070

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. (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). Curran Associates. https://proceedings.neurips.cc/paper/2009/file/4c27cea8526af8cfee3be5e183ac9605-Paper.pdf

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wikipedia Contributors. (2021). *Wikipedia: About*. https://en.wikipedia.org/wiki/Wikipedia:About

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Zhang, X., Astivia, O. L. O., Kroc, E., & Zumbo, B. D. (2022). How to think clearly about the central limit theorem. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000448

Zielstra, D., & Zipf, A. (2010). *Quantitative studies on the data quality of OpenStreetMap in Germany* [Conference session]. AGILE 2010. The 13th AGILE international conference on geographic information science, Guimarães, Portugal.

(*Appendices follow*)

Appendix A

General Knowledge Questions

Table A1

Table of Items for Experiments 1 and 2 Using General Knowledge Questions

Item	Question	Correct answer
1.	How large is the Eiffel Tower?	300 m
2.	How many sovereign countries are located in Africa?	54 countries
3.	How long is the Nile?	6,650 km
4.	How old was Johann Wolfang von Goethe?	82 years old
5.	How many bones does a human have?	214 bones
6.	What is Earth's mean radius?	6,371 km
7.	How old was Martin Luther King Jr.?	39 years old
8.	How tall is the Brandenburg Gate?	26 m
9.	How high was the highest temperature ever measured on Earth?	57 °C
10.	At what temperature does lead melt?	328 °C
11.	In which year did the first manned space flight take place?	1961
12.	How high is Mount Everst?	8,848 m
13.	How much does a tennis ball weigh?	57 g
14.	How many keys does a typical piano have?	88 keys
15.	How fast can a cheetah run?	112 km/h
16.	How long can a blue whale become?	33 m
17.	How much do 10 l of oxygen weigh?	14 g
18.	When was UNICEF founded?	1946
19.	How many prime numbers are in the interval between 1 and 1,000?	168 prime numbers
20.	How many star constellations are officially recognized?	88 constellations
21.	How many kilocalories do 10 gummy bears have (i.e., 30 g)?	98 kilocalories
22.	How long is a soccer goal?	7 m
23.	When was the last capital punishment enforced in France?	1977
24.	How many plays of Shakespeare are preserved?	33 plays
25.	How long is the kidney of a full-grown person?	12 cm
26.	How many species of the hawaiian honeycreeper exist?	21 species
27.	When was the lightning rod invented?	1,752
28.	When did the first modern Oylmpic Games take place?	1896
29.	How fast can a raindrop fall?	9 meters per second
30.	When was Leonardo da Vinci born?	1,452
31.	What is the maximum time that a total solar eclipse can take?	7 min
32.	How many strings does a concert harp have?	47 strings
33.	What is mean life expectancy for women in Germany?	81 years
34.	How wide is Lake Constance at its widest point?	14 km
35.	How long is the distance between Earth and Sun in million kilometers?	150 million kilometers
36.	When was women's suffrage adapted in Swizerland?	1971
37.	How many chaptes does the Quran have?	114 chapters
38.	How many times larger is the diameter of Juptier compared to the diameter of Earth?	11 times
39.	How large is the island of Borkum?	31 square-kilometers
40.	How many singles did the Beatles officially release?	22 singles
41.	How old was Alexander the Great when he waged his first campaign?	18 years old
42.	How many species of insects live in Antarctica?	52 species
43.	How many federal states does Austria have?	9 federal states
44.	When was the first human heart transplant performed?	1967
45.	How many marriages were there in Germany in 2018?	449,466 marriages
46.	How many students were enrolled in German university in the winter term of 2019/2020?	2,897,336 students
47.	How many floors does Burj Khalifa have?	163 floors
48.	How far is Frankfurt (Main) from Berlin (linear distance)?	424 km
49.	When was the first color film available in Germany?	1936

(*table continues*)

(*Appendices continue*)

Table A1 (*continued*)

Item	Question	Correct answer
50.	When was the numerus clausus first applied in German universities?	1968
51.	How far is Paris from London (linear distance)?	343 km
52.	How far is Dortmund from Hamburg (linear distance)?	284 km
53.	How far is Munich from Athens (linear distance)?	1,496 km
54.	How tall is the Statue of Liberty including its pedestral?	93 m
55.	When was slavery officially ended in the United States?	1865
56.	When was the first Autobahn inaugurated?	1921
57.	When did Albert Schweitzer receive the Nobel Peace Price?	1952
58.	How long is the mean distance between Earth and Moon?	384,400 km
59.	In which year was Uranus discovered by William Herschel?	1,781
60.	How many letters does the Arabic script have?	28 letters
61.	How deep is the Pacific at the deepest point?	10,094 m
62.	When was Astrid Lindgren born?	1907
63.	How much does the heart of a full-grown person weigh?	300 g
64.	How long can a Green Anakonda become?	8 m
65.	After how many days has a person's top layer of skin completely renewed?	28 days

Appendix B

Cities Selected for Different Maps

Table B1
Table of Items for Experiment 3 Using Map Material

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

Received June 2, 2021
Revision received May 16, 2022
Accepted July 13, 2022 ■

Expertise Determines the Frequency and Accuracy of Contributions in Sequential Collaboration

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

[1] University of Mannheim

[2] Heidelberg Academy of Sciences and Humanities

[3] University of Marburg

Author Note

Maren Mayer, Department of Psychology, School of Social Science, University of Mannheim, Germany. https://orcid.org/0000-0002-6830-7768

Daniel W. Heck, Department of Psychology, University of Marburg, Germany. https://orcid.org/0000-0002-6302-9252

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

This work was presented at the 63rd Conference of Experimental Psychologists (TeaP 2022 in Cologne). The present version of the manuscript (June 20, 2022) has not yet been peer reviewed. A preprint was uploaded to PsyArXiv and ResearchGate for timely dissemination.

This publication is part of the research group 'Shared Data Sources' in the subprogram '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).

The authors made the following contributions. Maren Mayer: Conceptualization, Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing; Marcel Broß: Conceptualization, Investigation, Writing - Review & Editing; Daniel W. Heck: Conceptualization, Methodology, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Maren Mayer, B6, 30-32, 68169 Mannheim. E-mail: maren.mayer@students.uni-mannheim.de

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 consecutively encountered by other contributors who then decide whether to adjust or maintain the presented entry. 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 to 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 show that the high accuracy of sequential collaboration is due to an implict weighting of judgments by expertise.

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

Word count: 7,726

Expertise Determines the Frequency and Accuracy of Contributions in Sequential Collaboration

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). Moreover, OpenStreetMap provides similarly accurate geographic information as commercial map services and governmental data (Ciepłuch et al., 2010; Haklay, 2010; Zhang & Malczewski, 2017; Zielstra & Zipf, 2010). Both Wikipedia and OpenStreetMap build on a sequential process of information gathering referred to as sequential collaboration (Mayer & Heck, 2021). One contributor creates an entry whereas the following contributors decide whether to adjust or maintain the presented entries. Thereby, only the latest version of an entry is shown.

Mayer and Heck (2021) 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 to six contributors. Each chain started with one independent judgment. Following contributors then encountered the latest version of the judgment and could decide whether to adjust or maintain it. For instance, one individual may start with locating Rome on a map of Italy. 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 the presented judgment, and how often a judgment had already been adjusted.

While change probability and change magnitude were found to decrease over the course of a sequential chain, judgment accuracy improved (Mayer & Heck, 2021). These findings show that sequential collaboration is suitable for eliciting accurate judgments. Furthermore, 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. This is an important finding given that unweighted averaging is known to yield highly accurate estimates in various contexts and tasks (Hueffer et al., 2013; Larrick & Soll, 2006; Steyvers et al., 2009; Surowiecki, 2004).

However, the mechanisms contributing to the high accuracy of sequential collaboration are still unclear. In the present paper, we investigate whether individual differences in expertise are a relevant factor for the probability of changing presented judgments and for the accuracy of such changes. We hypothesize that individuals with higher expertise adjust presented judgments more frequently and more accurately since they better distinguish between presented judgments that they can improve and those they cannot improve (Mayer & Heck, 2021). Thereby, sequential collaboration would facilitate an implicit weighing 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. We then refer to the literature on the role of expertise for individual judgments to establish a theoretical framework of how expertise influences both the frequency and accuracy of changing presented judgments in sequential collaboration. In three experimental studies using a city-location task and a random-dots estimation task, we measured and manipulated expertise. Thereby, we examined whether expertise influences how frequently and how accurately presented judgments in sequential collaboration are changed. As expected, we found that contributors with higher expertise change presented judgments more frequently and more accurately. Furthermore, experts have a larger impact on sequential chains than novices, while this effect is more pronounced the later experts enter into the chain.

Expertise in Judgment and Decision Making

Expertise is a multifaceted concept (Baumann & Bonner, 2013). It comprises general abilities such as logical reasoning (Kruger & Dunning, 1999), semantic

knowledge (Schunn & Anderson, 1999) such as unique information received (Baumann & Bonner, 2013; Stewart & Stasser, 1995) or grammar rules learned (Kruger & Dunning, 1999), prior experience such as students making decisions on their curriculum (Dubrovsky et al., 1991) or forensic experts judging the frequency of handwriting features (Martire et al., 2018), and procedural skills such as techniques of designing experiments (Schunn & Anderson, 1999). All these aspects have in common that expertise is domain-specific. With respect to knowledge, experience, and skills, expertise can be acquired by formal training (Martire et al., 2018).

It has been shown that experts work on tasks in qualitatively different ways (Dubrovsky et al., 1991; Franz & Larson, 2002; Schunn & Anderson, 1999) and usually show better performance than novices (Budescu & Chen, 2014; Kruger & Dunning, 1999; Merkle et al., 2020; Wang et al., 2021). Various measures of expertise have been used in the past. Most obviously, an individuals' expertise can be measured based on previous performance on similar tasks (Lin & Cheng, 2009) or even the same task (Budescu & Chen, 2014; Mayer & Heck, 2022; Merkle et al., 2020; Merkle & Steyvers, 2011). Moreover, one can take advantage of experts' ability to accurately predict judgments of other individuals. For instance, experts are able to predict how likely others provide the same judgment (surprisingly popular method, Lee et al., 2018; Prelec et al., 2017) and which judgments are likely provided when asking others in general (social projection, Grüning & Krueger, 2021) or when asking their peers (peer prediction, Wang et al., 2021).

Expertise has a positive effect on task performance in various contexts. In group decision making, the more individuals are aware of the expertise of other group members, the more accurate group decisions become (Baumann & 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). When eliciting independent judgments by a group of individuals, weighting these judgments by expertise improves the accuracy of the aggregated estimates (Budescu & Chen, 2014; Lin & Cheng, 2009; Mayer & Heck, 2022; Merkle et al., 2020; Merkle & Steyvers, 2011).

In such cases, expertise needs to be estimated statistically.

Deviation of Presented Judgments from the Correct Answer

We focus on two factors that may affect whether contributors in sequential collaboration adjust presented judgments and how accurate such adjustments are, namely, the accuracy of presented judgments and the expertise of contributors. In sequential collaboration, two aspects determine how difficult it is to accurately change a presented judgment. First, from a standard test-theoretical perspective, items or stimuli generally differ in how easy or difficult the correct answer is available (Embretson & Reise, 2000; Lord et al., 1968). Second, in sequential collaboration, individuals encounter not only the item itself but also a judgment by a previous contributor which may deviate more or less from the correct answer. In our studies, we focus on the deviation of a presented judgment from the correct answer, referred to as presented deviation, as an important predictor of whether and how much presented judgments are adjusted.

We expect that as the presented deviation increases, judgments are adjusted more frequently since contributors can more easily detect whether they are able to provide a more accurate judgment. Furthermore, we hypothesize that with increasing deviation, contributors change the presented judgments to a larger degree as there is more opportunity for improvement:

> *Hypothesis 1:* With increasing deviation of the presented judgment from the correct answer, participants (1a) change judgments more frequently and (1b) provide larger improvements.

The Role of Expertise

When relying on the aggregation of independent judgments, weighing these judgments by expertise results in an increase in accuracy (Budescu & Chen, 2014; Lin & Cheng, 2009; Mayer & Heck, 2022; Merkle et al., 2020; Merkle & Steyvers, 2011). We hypothesize that sequential collaboration provides accurate outcomes because it results

in an implicit weighting of judgments by expertise. This follows when assuming that experts are able to distinguish between presented judgments they can improve and those they cannot improve. In sequential collaboration, an implicit weighing of judgments emerges due to the possibility to opt out of providing a judgment. When opting out and maintaining the presented judgment, participants assign more weight to the presented judgment. In contrast, when opting in and adjusting a presented judgment, participants give more weight to their own judgment compared to the presented judgment.

The fact that judgments become increasingly more accurate over the course of a sequential chain (Mayer & Heck, 2021) indicates that contributors decide whether to opt in or opt out of revising the presented judgments according to their expertise. Such a process requires individuals to rely on task-related metacognitive knowledge about their expertise. Metacognition describes individuals' "cognition about cognitive phenomena" (Flavell, 1979) which is divided into metacognitive knowledge and metacognitive control (Lai, 2011). In the context of sequential collaboration, metacognitive knowledge about one's own expertise allows contributors to evaluate the accuracy of presented judgments and one's own capacity to provide improvements (Kruger & Dunning, 1999). Given that contributors decide whether to opt in or opt out, sequential collaboration does not require the identification of experts. It is neither necessary to assign expert roles before group discussions, nor is it required to estimate expertise statistically as when eliciting independent judgments. Instead, contributors determine the weighting of judgments within sequential chains implicitly based on their metacognitive assessment of their expertise. Achieving high accuracy only requires a sample with sufficient diversity in task-related expertise (Davis-Stober et al., 2014).

However, individuals may not have well-calibrated metacognitive knowledge about their own expertise. Specifically, the Dunning-Kruger effect, according to which individuals with low expertise overestimate their performance, may negatively affect judgment accuracy in sequential collaboration (Jansen et al., 2021; Kruger & Dunning, 1999). Overestimating one's knowledge may be a consequence of low expertise itself, since expertise is necessary both to perform well and to evaluate the accuracy of

judgments (Kruger & Dunning, 1999). In sequential collaboration, expertise is especially relevant for assessing the performance of others. For instance, when seeing a location judgment of Rome on a map of Italy, contributors require geographic knowledge both to evaluate the presented judgment and, when deciding to make a change, to provide an improved judgment. If this evaluation fails due to a lack of expertise, contributors with lower expertise may decide to adjust already accurate judgments and provide worse judgments. If many individuals suffer from the Dunning-Kruger-Effect, biased judgments will negatively affect the accuracy of sequential collaboration.

Frequency of Adjustments to Presented Judgments. If contributors with higher expertise are better at detecting which judgments they can improve, the deviation of presented judgments from the correct answer will have a larger effect on change probability for contributors with higher expertise than for those with lower expertise. Imagine Rome being positioned either far away from the correct position (e.g., near Milan), closer to the correct position (e.g., near Naples), or at the correct position. Contributors with higher expertise should correctly adjust the incorrect judgments close to Milan and close to Naples, while realizing that they cannot improve the already correct judgment. In contrast, contributors with lower expertise may only know that Rome is roughly located in the center of Italy. Hence, they may recognize the incorrect position near Milan, but not the incorrect position close to Naples. Contributors with lower expertise may even adjust and worsen the already accurate position because they erroneously expect Rome to be closer to the eastern coast.

This line of argument directly implies an interaction of expertise and the deviation of the presented judgment from the correct answer. However, it is less clear whether to expect a main effect of expertise on change probability. On the one hand, contributors with higher expertise should detect incorrect judgments with higher probability, but on the other hand, they should also maintain already highly accurate judgments. In our studies, most of the presented judgments deviate from the correct answer to a considerable degree, and hence, we predict a main effect of expertise:

Hypothesis 2a: Participants with higher expertise change presented judgments more frequently compared to those with lower expertise.

Hypothesis 3a: Compared to participants with lower expertise, participants with higher expertise are better at distinguishing between judgments with larger than smaller deviations from the correct answer, in turn leading to a larger effect of the presented deviation on change probability.

Accuracy of Revised Judgments.

When deciding to adjust a presented judgment, we expect that contributors with higher expertise change judgments more accurately than contributors with lower expertise. Thus, we expect a main effect of contributors' expertise on the improvement of presented judgments:

Hypothesis 2b: Participants with higher expertise provide larger improvements to the presented judgments compared to participants with lower expertise.

According to Hypothesis 3a, contributors with higher expertise should be better at detecting those judgments they can improve. Similarly, contributors with higher expertise should make especially large improvements to highly inaccurate presented judgments, only minor improvements to moderately inaccurate judgments, and no adjustments to correct judgments. In contrast, contributors with lower expertise may not be able to make similarly large improvements to highly inaccurate presented judgments, because they may also suffer from anchoring on the presented judgments (Mussweiler et al., 2004; Tversky & Kahneman, 1974). While contributors with higher expertise reach a certain level of accuracy when adjusting presented judgments, contributors with lower expertise may be more strongly influenced by the presented judgments. Therefore, we predict an interaction of expertise and the deviation of presented judgments from the correct answer with respect to the improvement of presented judgments.

Hypothesis 3b: For participants with higher expertise, the effect of the deviation of presented judgments on the amount of improvement is larger compared to participants with lower expertise.

Experiment 1

Methods

In Experiment 1, we measured expertise in a city-location task before letting participants decide whether to change or maintain location judgments with varying distances to the correct answer. To this end, we draw on an established paradigm already used by Mayer and Heck (2021) 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 measure of expertise while the remaining items were used to examine how participants adjust judgments in terms of change probability and accuracy. Thereby, expertise was operationalized as knowledge acquired in the past (Schunn & Anderson, 1999). The study design, sample size, hypotheses, and planned analyses were preregistered at https://aspredicted.org/blind.php?x=JZ7_K2K. Materials, analysis scripts, and data are available at https://osf.io/z2cxv/.

Participants

We recruited 290 participants who were compensated with 0.75€ for participation via a German panel provider for this study. However, 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 8 participants who positioned more than 10% of the cities outside the area of 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 48.75% of participants being female. Concerning educational background, 15.71% had a college degree, 15% held a high school diploma, 31.07% had vocational education, and 38.21% had a lesser educational attainment.

Materials and Procedure

Participants had to locate 57 European cities on 7 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 800×500 pixels and were scaled to 1:5,000,000. Appendix A1 provides a list of all cities and the phase they were presented in.

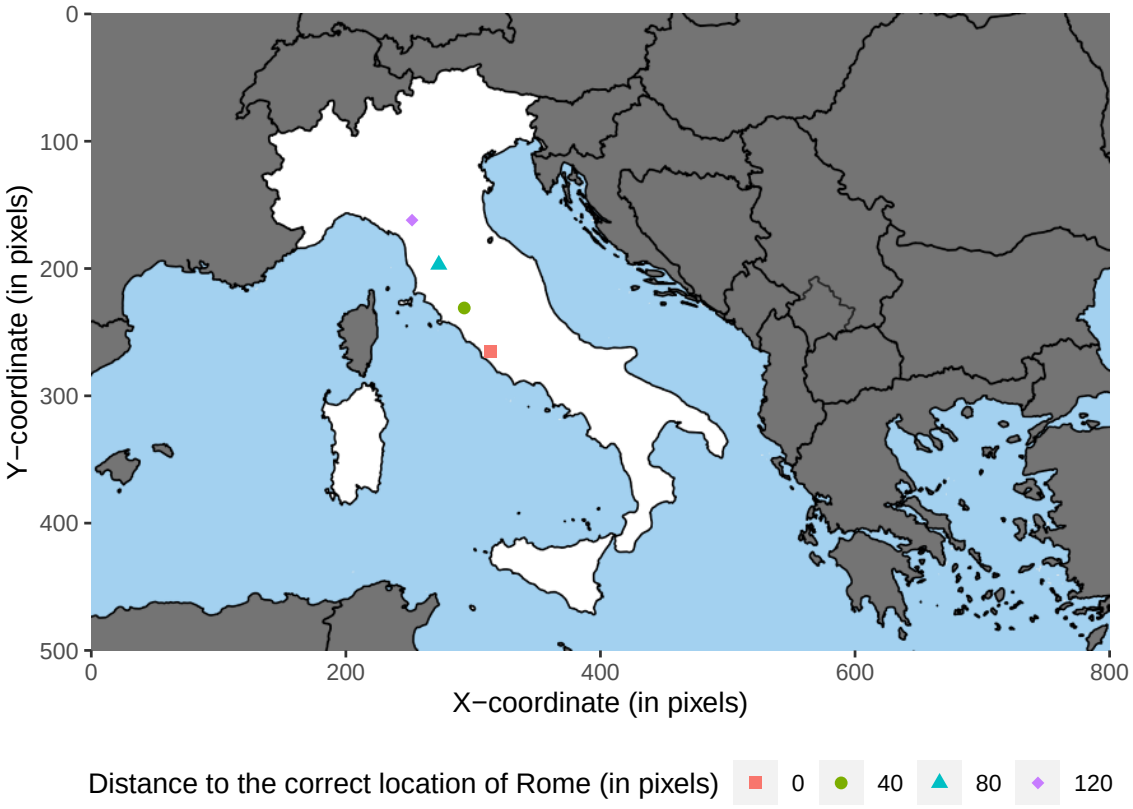
Participants first provided independent location judgments for 17 cities which served as a measure of expertise. We ensured a wide range in item difficulty by selecting cities based on the accuracy of independently provided judgments in a previous study (Mayer & Heck, 2021). Next, in the sequential phase, the remaining 40 cities were presented together with a preselected location judgment which was framed as a response of a previous participant. Figure 1 displays the map of Italy with four location judgments for Rome with different distances from the correct location. Separately for each city, participants decided whether to adjust or maintain the presented judgment before continuing to the next trial. All seven maps and the corresponding cities were presented in random order. Finally, participants provided demographic information, indicated their subjective expertise concerning the location of large European cities, and were debriefed and thanked for their participation.

Unknown to the participants, the locations presented in the sequential phase were not provided by other participants but preselected to manipulate the presented deviation with the Euclidean distance to the correct answer (0, 40, 80, or 120 pixels). For each of the 40 cities, one deviation was randomly selected such that each was presented 10 times. Furthermore, we ensured that all levels of deviation occurred within each of the seven maps. As participants were deceived about the presented locations being judgments of other participants, the study was reviewed and approved by the ethical committee of the University of Mannheim and participants were debriefed after participation.

To ensure that participants complied to the instructions, the online study was

Figure 1

Presented location judgments for Rome with different distances to the correct position.



Note. Each participant was only presented with one of the four preselected judgments.

accessible only for participants using a computer (as opposed to 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.

Results and Discussion

We estimated participants' expertise based on the independently provided judgments to the first 17 cities. For each participant, we computed the mean of the Euclidean distances between the location judgments and the correct positions. To ensure that larger values indicate higher expertise, we use the negatively inverted distance as a measure for expertise in the analyses below. We examined the validity of this expertise measure by correlating it with the self-reported expertise 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 Hypotheses 1a, 2a, and 3a concerning change probability using a generalized linear mixed model with a logistic link function. The model predicts the decision whether to adjust (= 1) or maintain (= 0) a presented judgment depending on expertise and presented deviation. We standardized our expertise measure for all analyses to address issues with model convergence. Moreover, we used a mean-centered linear contrast with values -1.5, -0.5, 0.5, 1.5 for the four levels of deviation of the presented locations from the correct location. The model accounts for the nested data structure by including random intercepts for items and participants (Pinheiro & Bates, 2000).[1]

Figure 2A displays the average change probability whereas Table 1 shows the estimated regression coefficients. In line with Hypothesis 1a, the linear contrast for the presented deviation was positive and significant ($\beta = 0.444$, $CI = [0.392, 0.495]$). The model also indicated a significant positive relationship between expertise and change probability, thus supporting Hypothesis 2a ($\beta = 0.622$, $CI = [0.202, 1.042]$). Furthermore, we found a significant interaction between expertise and the linear contrast for presented deviation ($\beta = 0.218$, $CI = [0.165, 0.272]$). However, contrary to our predictions, Figure 2A shows that individuals with higher expertise changed correct judgments more frequently than individuals with lower expertise which only partially supports Hypothesis 3a. The high change probability for accurate presented judgments can be explained by demand effects. Participants did not know that 25% of the presented judgments were already correct, and thus, they may not have expected that optimal behavior required to maintain a substantial proportion of the presented judgments.

Before assessing whether the improvement depends on the presented deviation (Hypothesis 1a), expertise (Hypothesis 2b), and their interaction (Hypothesis 3b), we

[1] It is often recommended to include random slopes for within-person factors. However, our models failed to converge when adding random slopes for the presented deviation.

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						
	Presented deviation	0.444	0.026	0.392	0.495	$< .001$
Experiment 1	Expertise	0.622	0.214	0.202	1.042	.004
	Presented deviation × expertise	0.218	0.027	0.165	0.272	$< .001$
	Presented deviation (V-shaped contrast)	0.208	0.024	0.160	0.256	$< .001$
	Presented deviation (linear contrast)	0.311	0.069	0.176	0.447	$< .001$
Experiment 2	Expertise	0.566	0.213	0.148	0.984	.008
	Presented deviation (V-shape) × expertise	0.067	0.031	0.006	0.128	.030
	Presented deviation (linear) × expertise	-0.342	0.092	-0.522	-0.163	$< .001$
	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	$< .001$
Experiment 3	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						
	Presented deviation	32.289	0.455	31.398	33.181	$< .001$
Experiment 1	Expertise	15.545	1.047	13.492	17.598	$< .001$
	Presented deviation × expertise	3.819	0.453	2.930	4.707	$< .001$
	Presented deviation (V-shaped contrast)	6.770	0.229	6.320	7.220	$< .001$
	Presented deviation (linear contrast)	-0.591	0.566	-1.700	0.518	.305
Experiment 2	Expertise	8.542	2.338	3.959	13.124	$< .001$
	Presented deviation (V-shape) × expertise	0.666	0.241	0.194	1.138	.006
	Presented deviation (linear) × expertise	-0.190	0.563	-1.293	0.912	.735
	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
Experiment 3	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. All models included crossed random effects for participants and items. The models for change probability (0 = no adjustment, 1 = adjustment) assumed a logit link function.

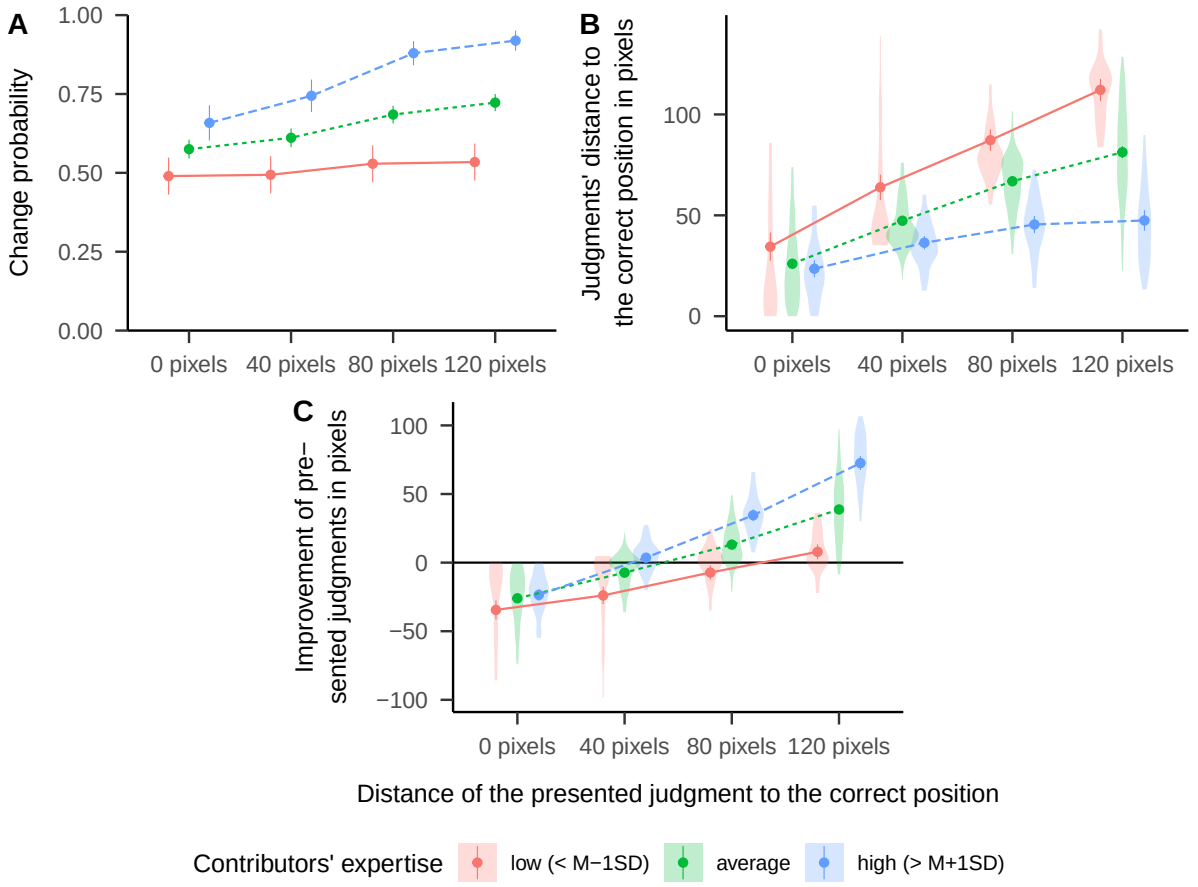
assess the accuracy of the provided judgments. To this end, we only included trials in which participants actually adjusted the presented judgments. Accuracy was operationalized as the Euclidean distance between the adjusted and the correct location.[2] Figure 2B displays the average distance to the correct position for the revised location judgments. Participants with higher expertise provided similarly accurate judgments for all levels of deviation. In contrast, for participants with lower expertise, inaccuracy of the revised judgments increased for larger presented deviations. This indicates that participants with lower expertise are prone to an anchoring effect.

Next, we examined the improvement of presented judgments by computing the difference in accuracy between the revised and the presented judgment. The accuracy of presented judgments corresponds to the presented deviation (i.e., 0, 40, 80, or 120 pixels distance to the correct position). Positive (negative) values of the improvement measure imply that a revised judgment is more (less) accurate than the presented judgment. We used improvement as dependent variable in a linear mixed model with (standardized) expertise and presented deviation (linear contrast) as independent variables. Figure 2C displays the average improvement in judgment accuracy, whereas Table 1 shows the regression coefficients. As expected, improvements increased for larger presented deviations (Hypothesis 1b: $\beta = 32.289$, $CI = [31.398, 33.181]$) and higher expertise (Hypothesis 2b: $\beta = 15.545$, $CI = [13.492, 17.598]$). In line with the plot, the model also showed a significant interaction such that more knowledgeable participants showed a steeper increase in improvement than less knowledgeable participants ($\beta = 3.819$, $CI = [2.930, 4.707]$).

[2] The statistical analysis yielded similar results when including non-adjusted judgments in the analysis with improvement scores of zero (main effect of expertise: $\beta = 12.200$, $CI = [10.523, 13.876]$, $t(275.82) = 14.263$, $p < .001$; main effect of deviation: $\beta = 21.932$, $CI = [21.291, 22.574]$, $t(10,861.76) = 66.979$, $p < .001$; interaction of expertise and deviation: $\beta = 5.726$, $CI = [5.085, 6.367]$, $t(10,860.76) = 17.500$, $p < .001$).

Figure 2

Change probability, distance to the correct position, and improvement of presented judgments in Experiment 1.



Note. Points and vertical lines show the empirical means with the corresponding 99% between-subjects confidence intervals, respectively. Violin plots indicate the distribution of the dependent variable aggregated across items within each person.

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 (Kruger & Dunning, 1999; Schunn & Anderson, 1999). While acquiring knowledge is usually a time-consuming process, acquiring skills or strategies can often be achieved much easier by learning and rehearsing (Anderson et al., 1997; Anderson & Fincham, 1994).

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. This 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 a disadvantage 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/DGV_R52. Materials, data and analysis scripts are available at https://osf.io/z2cxv/.

Methods

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. Of the 11 participants excluded from the analysis, one did not complete the study conscientiously, one vastly underestimated and one vastly overestimated the number of dots for most items, one almost always gave the correct number of dots, one did not answer attention-check questions about the instructions correctly, and six participants in the experimental condition indicated that they did not apply the learned strategy. The remaining 113 participants (69.03% female) had a mean age of 25.72 (SD = 10.14).

Procedure

Participants were randomly assigned either to the expertise-manipulation condition (referred to as “experts”) or the control condition (“novices”). Experts were introduced to raster scanning, a strategy for estimating the number of objects on a presented image more accurately 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. Participants simply had to multiply the result by nine to obtain an estimate for the total number of dots in the image. To make multiplication easier, we advised participants to multiply the number of dots by ten and then subtract the number. Participants in the control condition only read an essay about the importance of accurate judgments. Afterwards, both groups answered four attention-check questions concerning the instructions.

First, all participants had to 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 were presented with five more images, now without a raster. The judgments in this phase served as a manipulation check. Participants then saw 30 images, each with an (alleged) judgment of a previous participant and had to decide whether to adjust or maintain the presented number of dots. The images were shown in random order with a time limit of 60 seconds (including a warning after 40 seconds). As in Experiment 1, presented judgments were not actually provided by previous participants but rather preselected to manipulate the deviation from the correct answer. Lastly, participants provided demographic information and were asked whether they used raster scanning in the experimental condition, whether they used any special strategy to estimate the number of dots in the control condition, and whether they completed the study conscientiously.

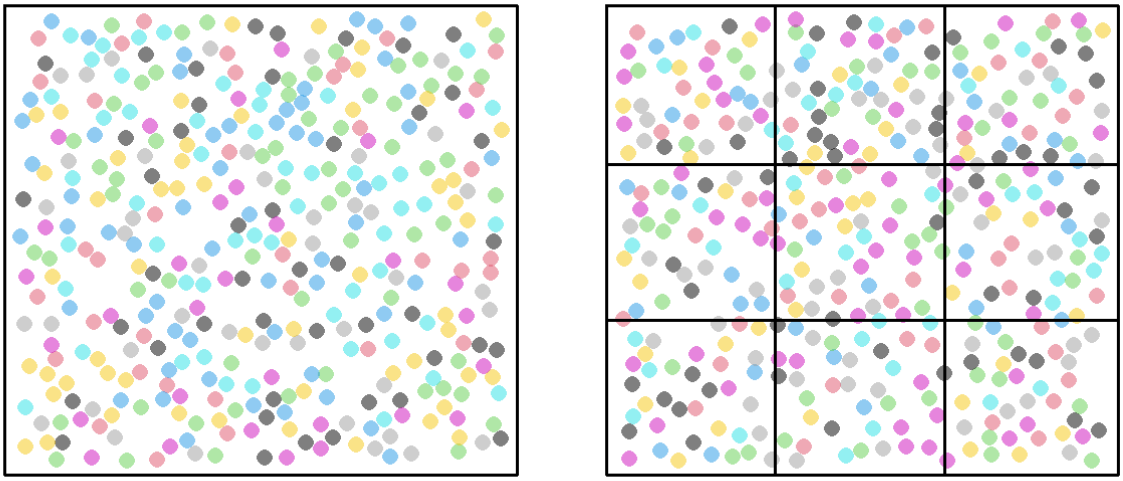
Materials

We generated 30 images (600×600 pixels, see Figure 3) with white background depicting between 100 and 599 randomly-positioned, non-overlapping, 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 values were preselected and either correct or deviated $\pm 35\%$ or $\pm 70\%$ from the correct answer.

Moreover, for motivational purposes, we also showed 10 additional images depicting only 10 to 59 dots which were displayed with a judgment that was either correct or deviated ±20% or ±35% from the correct answer. For these items, it was very easy for participants in both conditions to detect whether the presented judgment was correct since the time limit allowed to simply count the small number of dots.

Figure 3

Example images in the random-dots estimation task.



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.

Results and Discussion

To test whether the manipulation was successful, we examined whether experts provided more accurate independent judgments than novices. 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 with 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 = expertise, 0 = control). We found a significant negative effect of condition on the percentage error ($\beta = -15.805$, $CI = [-23.593, -8.017]$, $t(111.18) = -3.977$, $p < .001$). Hence, our manipulation of expertise was successful with novices showing a mean error of 35.81% in contrast to experts who showed a mean error of only 20.00%.

We first focus on Hypotheses 1a, 2a, and 3a using a generalized linear mixed model for change probability. While expertise was coded with a dummy contrast (1 = experts, 0 = novices), we used two orthogonal, centered contrasts for presented deviation. Since the presented deviation includes both over- and underestimation of the correct answer, we use a centered, V-shaped contrast (values: 4, -1, -6, -1, 4) to test whether change probability is lowest for correct presented judgments and increases the more the presented judgment deviates from the correct judgment. The regression coefficient of this contrast is positive for a V-shape, negative for an inverse V-shape, and zero for the absence of such an effect. Participants may not equally often adjust the presented judgments when these are over- or underestimating the correct judgment. Hence, we also include a linear contrast testing whether the slope of the V-shaped contrast differs for over- and underestimation of the presented judgment. A positive coefficient indicates a steeper slope for underestimation, a negative coefficient indicates a steeper slope for overestimation, and a value of zero indicates a symmetric V-shape.

Figure 4A 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. Table 1 shows the fixed-effects coefficients of the logistic model. Supporting Hypothesis 1a, the model revealed a significant, positive V-shaped contrast for item difficulty ($\beta = 0.208$, $CI = [0.160, 0.256]$). The positive linear contrast was also significant, indicating a smaller effect of the presented deviation (i.e., a smaller slope of the V-shape) for presented judgments underestimating the correct answer ($\beta = 0.311$, $CI = [0.176, 0.447]$). In line with Hypothesis 2a, we found a significant positive effect of

condition ($\beta = 0.566$, $CI = [0.148, 0.984]$). As expected in Hypothesis 3a, the interaction between condition and the V-shape contrast of the presented deviation was positive, meaning that experts better distinguished between accurate and inaccurate judgments ($\beta = 0.067$, $CI = [0.006, 0.128]$). However, experts adjusted already correct presented judgments more frequently than novices (Figure 4), and thus, our results only partially support Hypothesis 3a. Besides demand effects, this could also 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 already was correct. Lastly, we found a significant interaction between condition and the linear contrast of presented deviation, indicating that the V-shape is more symmetric (with respect to over- or underestimation of the correct answer) for experts than for novices ($\beta = -0.342$, $CI = [-0.522, -0.163]$).

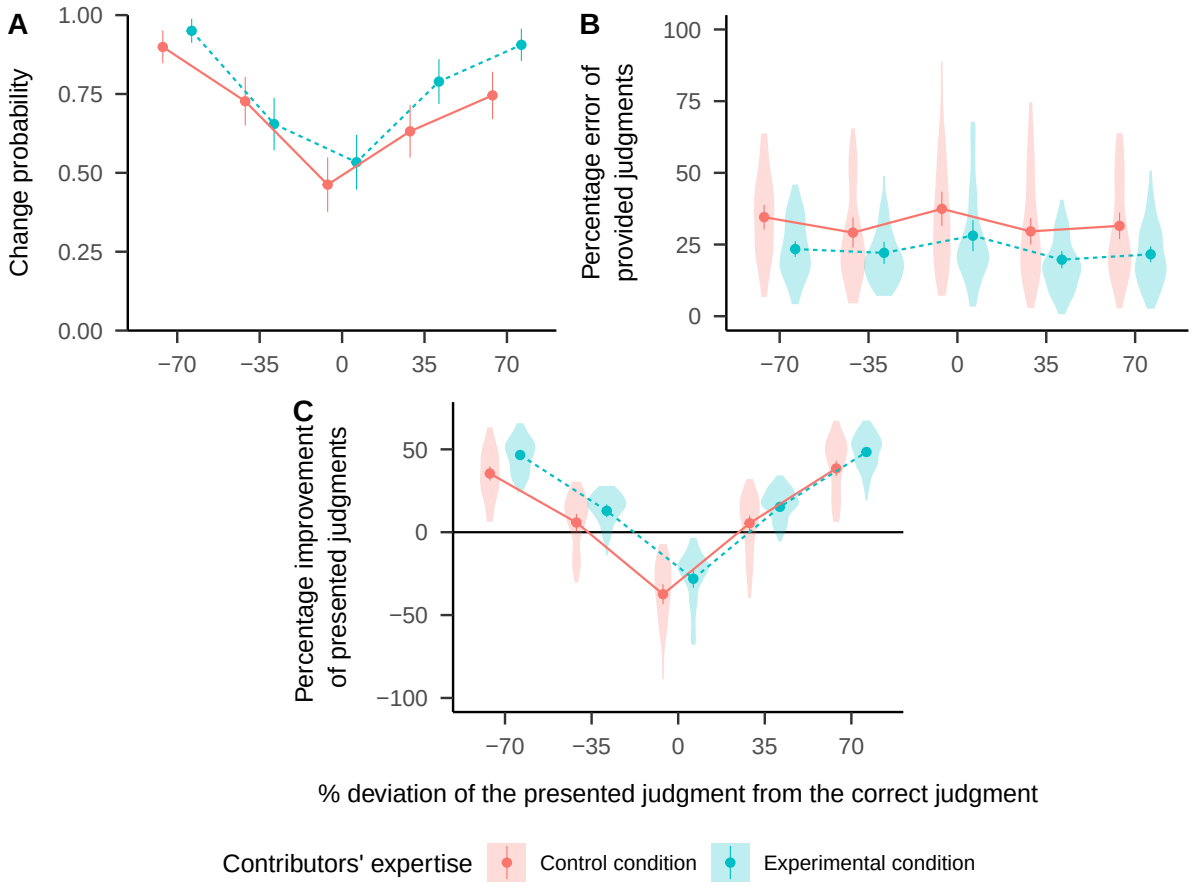
Next, we assess Hypotheses 1b, 2b, and 3b concerning the accuracy and amount of improvement of revised judgments. We thus consider only trials in which the presented judgment was adjusted. [3] The percentage error of the revised judgments is displayed in Figure 4B. Judgments of experts were generally more accurate than those of novices. In both conditions, accuracy appeared to be similar for all levels of presented deviation suggesting that there is no anchoring effect due to the presented judgments in both conditions.

We statistically test Hypotheses 1b, 2b, and 3b by focusing on the percentage improvement, defined as the difference between the percentage errors of the presented and the revised judgment. We used a linear mixed model to predict the improvement of presented judgments using the same contrasts for condition and presented deviation as

[3] Similar results were obtained when analyzing all trials while assigning an improvement of zero to maintained judgments (condition: $\beta = 8.575$, $CI = [4.543, 12.607]$, $t(111.02) = 4.168$, $p < .001$; V-shaped contrast for presented deviation: $\beta = 4.833$, $CI = [4.484, 5.182]$, $t(37.58) = 27.139$, $p < .001$; interaction of condition and V-shaped contrast: $\beta = 1.197$, $CI = [0.792, 1.601]$, $t(2, 117.18) = 5.800$, $p < .001$; all other effects were not significant).

Figure 4

Change probability, percentage error, and percentage 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 over items.

in the model for change probability. Figure 4C displays the mean percentage improvement of presented judgments including 99% confidence intervals and violin plots, while Table 1 shows the estimated regression coefficients. Supporting Hypothesis 1b, presented deviation had a V-shaped effect such that presented judgments were improved more the larger the deviation from the correct judgment was. This effect was significant in the model-based analysis ($\beta = 6.770$, $CI = [6.320, 7.220]$). Compared to novices, experts improved presented judgments more if there was room for improvement and worsened already correct judgments less ($\beta = -0.591$, $CI = [-1.700, 0.518]$).

Furthermore, the model showed a positive interaction between condition and the V-shaped contrast for presented deviation ($\beta = 0.666$, $CI = [0.194, 1.138]$). This speaks for a larger anchoring effect for novices compared to experts, which provides evidence for Hypothesis 3b.

Experiment 3

While experiment 1 and 2 showed that change probability and improvement of presented judgments depend on expertise, they implemented only a single incremental step in sequential collaboration using preselected values for the presented judgments. Importantly, the effects should still hold if individuals encounter actual judgments of previous individuals rather than preselected judgments. The benefits of expertise on the accuracy of sequential chains of judgments should especially manifest for the final estimates.

In the following, we derive additional hypotheses for sequential judgments made by groups of contributors. These hypotheses focus at the sequential-chain level rather than the individual level (as Hypotheses 1 to 3). Individuals with higher expertise should better distinguish between presented judgments provided by other experts and those by novices, which should in turn affect change probability and improvement of judgments. In contrast, novices are predicted to be worse at making this distinction, meaning that change probability is affected less by the status of the previous contributor, and that only judgments of other novices can be improved. Moreover, the more experts are assigned to a sequential chain, the more accurate the final estimates are expected to be. The improvements made by experts are less likely to be changed by others (and possibly worsened) if experts enter into the sequential chain later than sooner.

Hypothesis 4: In sequential chains, experts change presented judgments of novices more frequently than those of other experts. In contrast, novices have similar change probabilities regardless of the expertise of the previous participant.

Hypothesis 5: In sequential chains, accuracy improves most when experts adjust judgments of novices. Smaller improvements occur when experts correct experts or when novices correct novices. In contrast, novices worsen judgments of experts.

Hypothesis 6: The more experts are in a sequential chain, the better the final estimates. For sequential chains with the same number of experts and novices, final estimates are more accurate if experts are at the end of the chain than at the beginning.

To test Hypothesis 1 to 6, 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. The design allowed us to manipulate the number and position 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/.

Methods

Materials and Procedure

We used the same experimental paradigm as in Experiment 2 with 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. Thus, it was not necessary to exclude data from other persons in the same sequential chain later during the analysis. We also generated five new images 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 status and order of the contributors (i.e., novice-novice, expert-novice,

novice-expert, and expert-expert). As in Experiment 2, the first participant in each chain saw preselected judgments which were either correct, ±35% or ±70% below or above the correct number of dots. The second participant in each chain then saw the revised judgments provided by the first participant. If the first participant maintained a presented judgment, the second participant encountered the same value.

Participants

We recruited 464 participants via a German panel provider who were compensated with 1€. One participant was excluded because they answered "1" to all items, which in turn required to remove another participant assigned to the same chain. Moreover, five participants were excluded due to duplicate assignments to sequential chains. The final sample included 457 participants (46.83% female) with mean age 46.16 ($SD = 14.36$) and various educational background (college degree: 34.79%; high-school diploma: 26.04%; vocational education: 24.07%; lesser educational attainment: 15.10%).

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 lead to a decrease of the percentage error ($\beta = -28.898$, $CI = [-36.319, -21.477]$, $t(111.18) = -3.977$, $p < .001$) indicating that judgments of experts were twice as accurate as those of novices (mean error = 27.46% vs. mean error = 56.36%, respectively).

Change Probability and Improvement of Judgment Accuracy at the Individual Level

We first test Hypotheses 1 to 3 which refer to individual-level decisions and judgments. To analyze change probabilities in the sequential phase, we included only participants who saw the preselected judgments but not those who saw the judgments of other participants. Similar as in Experiment 2, we used a generalized linear mixed

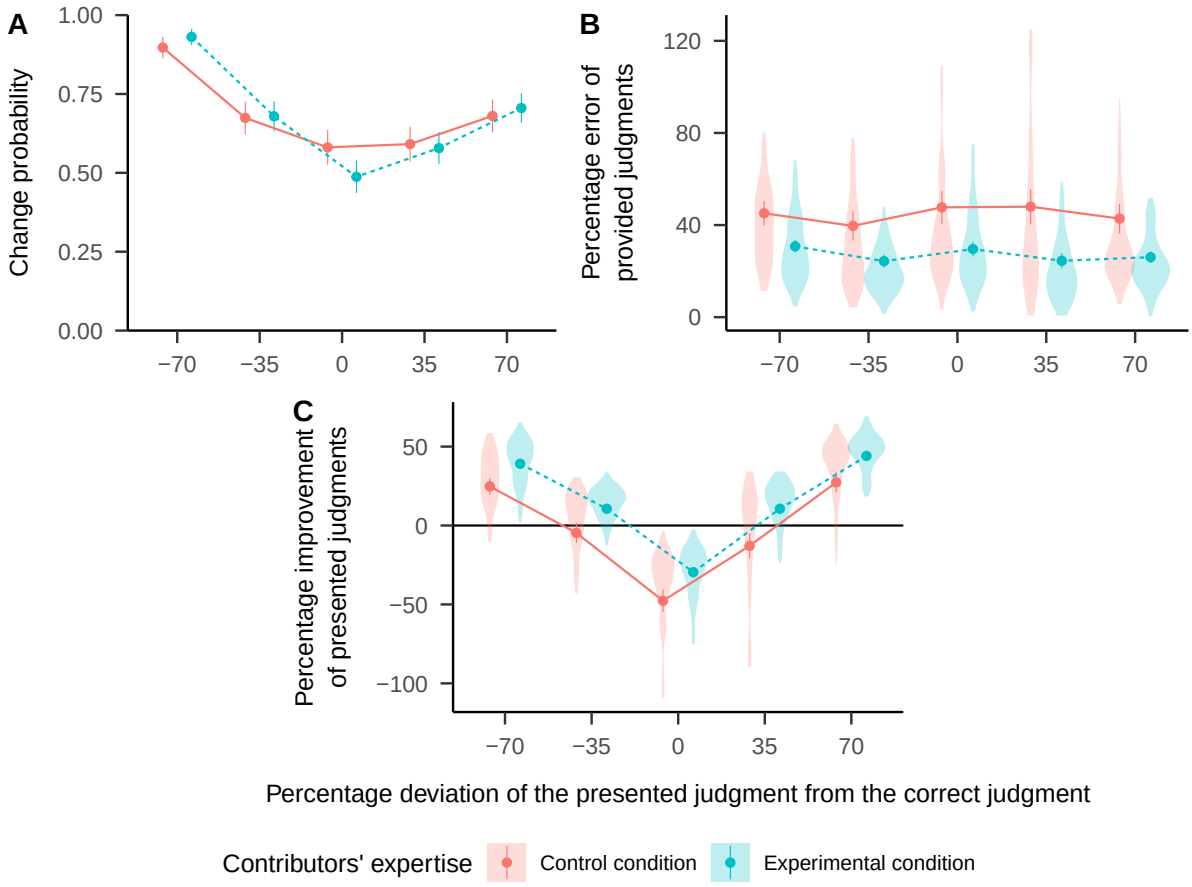
model to predict whether a presented judgment was changed, using the same contrasts for presented deviation and condition. Figure 5A displays the average change probabilities in Experiment 3. As expected, the V-shaped effect of presented deviation emerged, while it was steeper for presented judgments that underestimated rather than overestimated the correct answer. Moreover, the plot does not indicate an effect of condition. This impression was supported by the model-based analysis (see Table 1). In line with Hypothesis 1a, the V-shaped contrast of presented deviation on change probability was significant ($\beta = 0.141$, $CI = [0.116, 0.167]$). The linear contrast of deviation was also significant, indicating a steeper slope for the left than the right limb of the V-shaped effect ($\beta = 0.367$, $CI = [0.292, 0.441]$). Contrary to Hypothesis 2a, the effect of experimental condition was not significant ($\beta = 0.052$, $CI = [-0.297, 0.401]$). The interaction between condition and the V-shaped contrast was significant ($\beta = 0.075$, $CI = [0.040, 0.111]$) indicating that the effect of presented deviation on change probability was slightly stronger for experts than for novices. As shown in 5A, 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.

We tested Hypotheses 1b, 2b, and 3b concerning the improvement of presented judgments including only participants at the first chain position.[4] Figure 5B displays the percentage error of the revised judgments. As in Experiment 2, participants achieved a certain level of accuracy in both conditions independent of the presented deviation, while accuracy was generally higher for experts than for novices. Figure 5C displays the improvement of presented judgments which followed a V-shaped pattern, with already correct presented judgments being slightly worsened. Fitting a linear mixed model for the percentage improvement provided similar results. We used the same contrasts for condition and presented deviation as above. In line with Hypothesis

[4] Similar results were obtained when including maintained judgments as providing an improvement 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).

Figure 5

Change probability, percentage error, 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 over items.

1b, the model showed a V-shape effect of presented deviation ($\beta = 6.653$, $CI = [6.290, 7.016]$). Supporting Hypothesis 2b, the main effect of condition was significant, indicating more improvement of judgments for experts than novices ($\beta = 16.518$, $CI = [10.701, 22.336]$). In contrast to Hypothesis 3b, the interaction of condition and presented deviation was not significant ($\beta = -0.024$, $CI = [-0.527, 0.479]$). 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 to novices ($\beta = -1.516$,

$CI = [-2.745, -0.287]$).

As robustness check, we also tested Hypotheses 1, 2, and 3 using judgments of all participants. 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 included the standardized deviation and the corresponding, quadratic trend as predictors. For this analysis, we excluded 44 judgments provided by participants at the first chain position which had an percentage error of more than 200% since these judgments serve as presented judgments for participants at the second chain position and could obstruct the analysis.

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 ($\beta = 0.273$, $CI = [0.208, 0.338]$, $z = 8.223$, $p < .001$) and a significant interaction with condition ($\beta = 0.362$, $CI = [0.262, 0.461]$, $z = 7.098$, $p < .001$), whereas the effect of condition on change probability was not significant ($\beta = -0.232$, $CI = [-0.490, 0.025]$, $z = -1.768$, $p = .077$). 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 ($\beta = 19.345$, $CI = [18.245, 20.445]$, $t(7, 197.89) = 34.466$, $p < .001$) and a positive effect of condition ($\beta = 15.575$, $CI = [10.649, 20.501]$, $t(555.30) = 6.197$, $p < .001$) while the interaction was not significant ($\beta = 0.389$, $CI = [-1.102, 1.879]$, $t(7, 182.28) = 0.511$, $p = .609$).

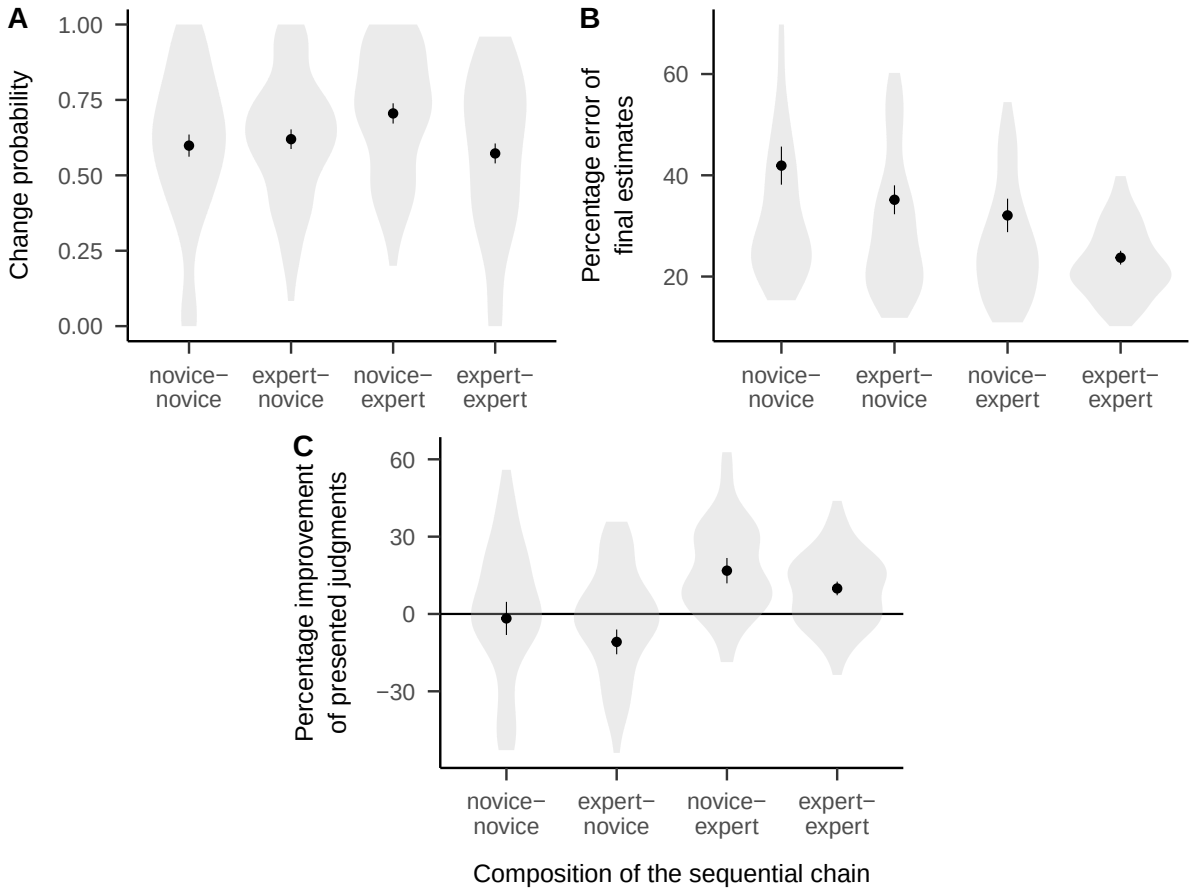
Change Probability, Judgment Accuracy, and Improvement of Presented Judgments at the Chain Level

We tested the hypotheses at the chain level based on the data of participants at the second chain position. Concerning Hypothesis 4, we fitted a generalized linear mixed model to predict whether change probability differs 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: one comparing

novice-novice chains against expert-novice chains, and another comparing novice-expert chains against expert-expert chains. In line with Hypothesis 4, change probability was larger for novice-expert than for expert-expert chains ($\beta = 0.326$, $CI = [0.063, 0.588]$, $z = 2.432$, $p = .015$) while novices changed the entries of experts and novices similarly frequently ($\beta = 0.136$, $CI = [-0.098, 0.370]$, $z = 1.140$, $p = .254$). As illustrated in Figure 6A, novices showed similar change probabilities when encountering judgments of novices and experts, while experts were more likely to change judgments of novices compared to those of experts.

Figure 6

Change probability, accuracy, and amount of improvement 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 jugdments aggregated for each participant across items.

To test Hypothesis 5, we only considered judgments that were adjusted by participants at the second chain position[5] and implemented a linear mixed model with percentage improvement as dependent variable and type of sequential chain as predictor. We additionally used Helmert contrasts to test our hypothesis 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 6C displays the empirical means for percentage improvement for all compositions of sequential chains. In line with this pattern and Hypothesis 5, we found a significant 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, as Figure 6 displays, novices worsen judgments of experts. Lastly, we did not find a significant difference in improvement between expert-expert and novice-novice groups ($\beta = -5.965$, $CI = [-12.137, 0.208]$, $t(222.70) = -1.894$, $p = .060$). These findings are in line with Hypothesis 5.

To test Hypothesis 6, 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 participants adjusted the presented judgment, the final judgment could either be the presented judgment, the judgments entered by the first participant, or the judgment entered by the second participant. We used a linear contrast to test for a decreasing percentage error and thus increasing accuracy across chain compositions.

In line with Hypothesis 6, we found a significant linear trend between chain composition and accuracy of the final estimates ($\beta = 5.779$, $CI = [2.199, 9.359]$,

[5] Similar results are obtained when maintained judgments are considered as not improved with a value 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).

$t(216.79) = 3.164$, $p = .002$). This pattern is displayed in Figure 6B showing that the percentage error was 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.

General Discussion

Our three experiments show that the probability of changing a presented judgments depends on its deviation to the correct answer (Hypothesis 1a), on participants' expertise (Hypothesis 2a), and on the corresponding interaction (Hypothesis 3a). However, Experiment 1 and 2 did not provide evidence for the interaction, whereas Experiment 3 did not indicate an effect of expertise. Presented deviation, expertise, and their interaction also affected the amount of improvement made to presented judgments (Hypothesis 1b, 2b, and 3b), while the interaction was not supported in Experiment 3. Experiment 3 investigated sequential chains of contributors, showing that experts adjust judgments of novices more frequently than those of other experts (Hypothesis 4), that experts improve judgments of novices most (Hypothesis 5), and that final estimates become more accurate the more experts are in a sequential chain and the later they enter (Hypothesis 6).

Overall, expertise is an important predictor of change probability and the amount of improvement of judgments in sequential collaboration. This supports the theoretical assumption that contributors adjust and maintain judgments based on their expertise which in turn results in 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 experts can be observed at the chain level. Still, 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 also add to the literature on the wisdom of crowds, supporting the notion that weighing judgments by expertise increases accuracy (Budescu & Chen, 2014; Mayer & Heck, 2022; 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, respectively. Instead, sequential collaboration results in an implicit weighting of judgments which is determined by the contributors' meta-cognitive assessment of whether they can improve the present judgments.

Future Research Directions

Our three studies are limited in that they only examined the effect of expertise for short sequential chains with only two contributors. We expect that the effects on change probability and improvement of judgments should similarly hold for longer sequential chains, given that participants were not aware about the number of contributors. However, this assumptions needs to be tested using experiments with longer chains.

Other variables besides expertise may also affect the frequency and improvement of judgments in sequential collaboration. Specifically, individuals' confidence will likely determine the decision whether to adjust a judgment. Domain expertise can be a source of high confidence if individuals do not hold erroneous belief (Koriat, 2008, 2011). However, confidence can also stem from miscalibrated meta-cognition (Kruger & Dunning, 1999) or from item-specific knowledge. Especially for general knowledge questions, contributors' knowledge for specific facts becomes relevant. For instance, contributors may know the location of certain cities since they lived there or recently visited them. Future research should thus examine the role of confidence in sequential collaboration.

While our studies show that expertise predicts change probability and the amount of improvement in sequential chains of judgments, it remains unclear whether the increased accuracy is due to the sequential judgment process itself or due to the

possibility to opt out of answering. Providing the opportunity to opt out increases the accuracy of independent individual judgments, since individuals can use their metacognitive knowledge to select those tasks that fit their individual expertise best (Bennett et al., 2018). Future research should thus disentangle the effects of the sequential judgment-elicitation process and of the opportunity to opt out of providing a judgment.

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 evidence for the implicit-weighting of expertise in sequential collaboration, which may contribute to the high accuracy of online collaborative projects.

References

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1322–1340. https://doi.org/10.1037/0278-7393.20.6.1322

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*, 932–945. https://doi.org/10.1037//0278-7393.23.4.932

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

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. *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, 337–340.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When

772 is a crowd wise? *Decision*, *1*, 79–101. https://doi.org/10.1037/dec0000004
773 Dubrovsky, V. J., Kiesler, S., & Sethna, B. N. (1991). The equalization
774 phenomenon: Status effects in computer-mediated and face-to-face
775 decision-making groups. *Human–Computer Interaction*, *6*, 119–146.
776 https://doi.org/10.1207/s15327051hci0602_2
777 Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.
778 https://doi.org/10.4324/9781410605269
779 Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of
780 cognitive–developmental inquiry. *American Psychologist*, *34*, 906–911.
781 https://doi.org/10.1037/0003-066X.34.10.906
782 Franz, T. M., & Larson, J. R. (2002). The impact of experts on information
783 sharing during group discussion. *Small Group Research*, *33*, 383–411.
784 https://doi.org/10.1177/104649640203300401
785 Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901.
786 https://doi.org/10.1038/438900a
787 Grüning, D. J., & Krueger, J. (2021). *Vox peritorum: Capitalizing on confidence*
788 *and projection to characterize expertise*.
789 https://doi.org/10.31234/osf.io/6vndh
790 Haklay, M. (2010). How good is volunteered geographical information? A
791 comparative study of OpenStreetMap and ordnance survey datasets.
792 *Environment and Planning B: Planning and Design*, *37*, 682–703.
793 https://doi.org/10.1068/b35097
794 Honda, H., Kagawa, R., & Shirasuna, M. (2022). On the round number bias and
795 wisdom of crowds in different response formats for numerical estimation.
796 *Scientific Reports*, *12*, 8167. https://doi.org/10.1038/s41598-022-11900-7
797 Hueffer, K., Fonseca, M. A., Leiserowitz, A., & Taylor, K. M. (2013). The
798 wisdom of crowds: Predicting a weather and climate-related event. *Judgment*
799 *and Decision Making*, *8*, 91–105.
800 http://journal.sjdm.org/12/12924a/jdm12924a.html

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*, 756–763. https://doi.org/10.1038/s41562-021-01057-0

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*, 117–139. https://doi.org/10.1037/a0022171

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). 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

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*, 322–333. http://sjdm.org/~baron/journal/18/18331/jdm18331.pdf

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 : JAMIA*, *17*, 373–374. https://doi.org/10.1136/jamia.2010.004507

Lin, S., & Cheng, C. (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

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

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. (2021). *Sequential collaboration: About the accuracy of dependent, incremental judgments*. https://doi.org/10.31234/osf.io/w4xdk

Mayer, M., & Heck, D. W. (2022). *Cultural consensus theory for two-dimensional data: Expertise-weighted aggregation of location judgments*. https://doi.org/10.31234/osf.io/unhvc

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). Springer. https://doi.org/10.1007/978-3-642-19656-0_34

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. In R. F. Pohl (Ed.), *Cognitive illusions* (1st ed., pp. 183–199). Psychology Press.

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). Springer. https://doi.org/10.1007/978-1-4419-0318-1_1

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question

crowd wisdom problem. *Nature*, *541*, 532–535. https://doi.org/10.1038/nature21054

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

Stewart, D. D., & Stasser, G. (1995). Expert role assignment and information sampling during collective recall and decision making. *Journal of Personality and Social Psychology*, *69*, 619–628. https://doi.org/10.1037/0022-3514.69.4.619

Steyvers, M., Miller, B., Hemmer, P., & Lee, M. (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). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/4c27cea8526af8cfee3be5e183ac9605-Paper.pdf

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

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*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Zhang, H., & Malczewski, J. (2017). Accuracy evaluation of the canadian OpenStreetMap road networks. *International Journal of Geospatial and Environmental Research*, *5*(2). https://ir.lib.uwo.ca/geographypub/347

Zielstra, D., & Zipf, A. (2010). *Quantitative studies on the data quality of*

OpenStreetMap in Germany. AGILE 2010. The 13th AGILE international conference on geographic information science.