The benefit of collaboration: Disentangling the sources of synergy in group judgments

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

–

im Promotionsprogramm Biologie

der Georg-August University School of Science (GAUSS)

vorgelegt von

Matthias Albert Lippold

aus Jever

Göttingen, 2021

Betreuungsausschuss

Prof. Dr. Stefan Schulz-Hardt, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. York Hagmayer, Abteilung Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Dr. Thomas Schultze, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Mitglieder der Prüfungskommission

Referent: Prof. Dr. Stefan Schulz-Hardt, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Korreferent: Prof. Dr. York Hagmayer, Abteilung Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

2. Korreferent: Dr. Thomas Schultze, Abteilung Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Margarete Boos, Abteilung Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-Institut für Psychologie

Prof. Dr. Sascha Schroeder, Abteilung Pädagogische Psychologie, Georg-Elias-Müller-Institut für Psychologie

Prof. Dr. Michael Waldmann, Abteilung Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie

Tag der mündlichen Prüfung: 19.11.2021

Danksagung

Gruppenforschung ist ein aufwendiges Unterfangen, welches nur durch die Zusammenarbeit von unterschiedlichen Personen ermöglicht wird. Daher möchte ich an dieser Stelle noch kurz einigen Menschen meinen herzlichen Dank aussprechen, welche über die Jahre hinweg an meinem Promotionsprojekt direkt oder indirekt mitgewirkt haben.

Als erstes möchte ich mich bei meinem Doktorvater Stefan Schulz-Hardt bedanken, welcher mir sein Vertrauen geschenkt hat, dieses Promotionsprojekt in Göttingen anzugehen. Seine methodische Herangehensweise an psychologische Forschung hat mich schon während meines Masterstudiums geprägt. Seiner Unterstützung konnte ich mir während meiner Promotionszeit immer sicher sein. Seine Anregungen und Ideen sind in die Arbeit miteingeflossen.

Mein besonderer Dank gilt Thomas Schultze-Gerlach, der mich während der gesamten Promotionszeit direkt begleitet und betreut hat. Seine Tür stand für mich und meine spontanen Fragen immer offen. Durch seinen fachlichen Input und die vielen gemeinsamen Gespräche hat er mir immer wieder direkt geholfen. Ohne seine wissenschaftliche Vorarbeit wäre mein Dissertationsprojekt nicht zustande gekommen. Daher bin ich ihm sehr dankbar.

Des Weiteren möchte ich mich bei York Hagmayer für seine Bereitschaft bedanken, mein Promotionsprojekt als Zweitgutachter zu betreuen und für die Möglichkeiten, die er mir als Tutor während meines Masterstudiums eingeräumt hat.

Bei Margarete Boos, Sascha Schroeder und Michael Waldmann möchte ich mich für die Bereitschaft bedanken, im Prüfungskomitee meiner Dissertation mitzuwirken.

Ich möchte mich ebenfalls bei meinen beiden Büropartnern Jana und Jacob bedanken. Mit beiden bin ich den Weg zur Promotion gegangen. Das hat mir Sicherheit gegeben und ich bin froh viele Erfahrungen gemeinsam gemacht zu haben.

Zudem möchte ich mich bei Christian Treffenstädt bedanken. Durch seine Programmierungen konnten die Experimente dieser Dissertation erst durchgeführt werden. Auch hatte er für mich immer ein offenes Ohr und stand mir mit Rat zur Seite.

Durch die Arbeit der vielen Hilfskräfte, welche die aufwendigen Gruppendaten für die einzelnen Experimente der Dissertation erhoben haben, ist mein Promotionsvorhaben erst möglich gewesen. Daher möchte ich mich an dieser Stelle bei allen Hilfskräften bedanken, welche bei diesem Projekt involviert waren.

Schlussendlich möchte ich mich bei Melli bedanken, die mich immer wieder motiviert hat, diese Arbeit zu Ende zu bringen und eine große Stütze für mich während des gesamten Prozesses war.

Leipzig, den 30.09.2021 Matthias Lippold

Table of Content

Preface

This dissertation follows the form of a cumulative dissertation. The first of the included manuscripts has been published. The following two manuscripts are included in my dissertation:

Lippold, M., Schulz-Hardt, S., & Schultze, T. (2021). G-I transfer in multicue judgment tasks: Discussion improves group members' knowledge about target relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(3), 532–545. https://doi.org/10.1037/xlm0000947

Lippold, M., Schultze, T., & Schulz-Hardt, S. (2021). *The benefit of collaboration: Disentangling the sources of synergy in group judgments.* Unpublished Manuscript.

1 Introduction

Human beings often form groups to conduct all kinds of tasks. Therefore, group structures can be found in almost any human organization. Governing bodies of industrial companies, supervisory boards in universities, panels of sports associations, or juries in court systems, there are countless examples for permanent implementations of groups. The presence of so many different groups might indicate that there is, in general, a strong faith in the quality of group performance (Larson, 2010). In other words, groups are often considered to have a superior performance in comparison to individuals (e.g., Gigone & Hastie, 1997), presumably because groups can rely on the experience and expertise of all of their members, and they have the potential to integrate their unique contributions. Accordingly, group researchers have focused intensively on the evaluation of group performances in various tasks. Some work on group performance even belongs to the earliest inquiries in the field of psychology (e.g., Ringelmann, 1913). While group performance seems to be highly regarded, research on this matter shows, in general, a rather poor record of groups. In many tasks, groups are unable to reach their full potential (for a review see Kerr & Tindale, 2004) and there are even loud voices calling for the avoidance of interacting groups (Armstrong, 2006). Nevertheless, there is hope for proponents of groups. Recent research points to one particular task domain, where group performance might justify the implementation of groups: quantitative judgments (e.g., Minson, Mueller, & Larrick, 2018; Schultze, Mojzisch, & Schulz-Hardt, 2012; Sniezek & Henry, 1989).

Quantitative judgment tasks are not only omnipresent in our daily lives, but our daily lives also depend heavily on the accuracy of these judgments. For example, the newest climate prognoses by scientists, the daily weather forecasts on the radio, and the newest stock market predictions can all have a strong influence on individual, group, and governmental decisions. Groups rather than individuals make many important quantitative judgments in our

society. Therefore, the importance of group performance in quantitative judgment tasks cannot be rated highly enough. Nevertheless, how can group performance in quantitative judgments be evaluated? How big is the performance benefit that groups presumably have over individuals in quantitative judgments? What are the group specific mechanisms leading to this effect? The current dissertation tries to answer these questions from a group researcher's point of view by relying on the empirical work from the two included manuscripts. First, I will provide a short overview about how group researchers assess the performances of small groups in general. Second, I will explain how group performance can be evaluated in the specific case of quantitative judgment tasks. Third, I will describe the group mechanisms affecting group judgments, how they can be explored, and what is already known about these mechanisms. Then, I will lead to the summary of the two manuscript and close with a discussion of their empirical findings.

1.1 Evaluating group performance

Group performance does not stand on its own but is subject to the same task restrictions as the performance of individuals. The nature of a specific task and human performance boundaries are important factors to consider for performance evaluation. When Steiner (1966), one of the most influential group researcher, conceptualized group performance, he had this in mind and considered group performance as a function or fit of the *resources* of the group members, the specific *task demands*, and the *process*. In this framework, the resources of the group members are their capabilities. They are comprised of the relevant skills, abilities, and knowledge for the specific task possessed by the individual group members. The task demands include the required resources and the corresponding utilization pattern needed to reach the optimal performance for a specific task. Note that if a group lacks the required resources demanded by a task, they will perform poorly on this task or even fail to complete it. For example, you would not expect a group of toddlers to solve a

complex mathematical problem as the group members lack the skills to conduct this type of task. Finally, the process consists of the actual behavior of the group members working on the task. This includes all interpersonal and individual actions that form the group outcome by relying on the resources of the group members. In this view, not the absolute group product on its own is the decisive factor for the evaluation of the group performance, but the focus is on the group process in the light of the resources of the group members and the task demands. Based on the resources of the group members and the tasks demands, the potential productivity of the group can be inferred and compared to the actual group performance. This allows for the evaluation of the group process. In order to do that, a meaningful baseline needs to be determined.

It is not enough, to compare the performance of a group to the performance of one single individual person given that in many cases, the superior number of individuals working in the groups almost guarantees a higher performance on the side of the groups. Following the terminology of Steiner, a group of three people has simply more resources than a single individual does. For example, the fact that a group of three people is able to move a household faster than a single person would not be surprising at all. Instead, the performances of an *N*-person-group needs to be compared to the performances of *N* independent individuals working alone whose individual contributions are then combined in a meaningful way to form the analog of a group performance. These groups consisting of non-interacting individuals are often named *nominal groups*. Their combined performance marks the group's potential productivity and is referred to as *group potential* (Steiner, 1972). Steiner defines the group potential as the case when the group members perform at their full potential and these individual contributions are combined in a meaningful way. When groups perform below the level of their group potential, that is making not the best use of their own available resources, they suffer from *process losses*.

The way the individual performances should be combined – e.g., the group potential is formed – depends on the specific task (Steiner, 1972)[1]. Steiner differentiates between *additive*, *conjunctive*, *disjunctive*, and *discretionary* tasks. As the name suggests, the group potential in additive tasks can be calculated by simple summing up the individual performances of the nominal group members. For example, in a brainstorming task, the number of ideas generated by each individual can be added to form the group potential. In conjunctive tasks, the group potential is determined by the performance of the weakest group member. The common example for this type of task is hill climbing. The task is only successfully completed whenever the entire team, including the slowest group member, has reached the peak of the mountain. On the other side of the spectrum are disjunctive tasks, in which the group potential is determined by the performance of the best - most capable - group member. For example, in a problem solving task, the most capable group member might be able to solve the problem on his own and can demonstrate the right solution to the other group members. The last task type described by Steiner is a discretionary task. In these tasks, there is no fixed rule for the group potential. The groups can decide freely on how the individual contributions of the group members are combined for the group product. Note that quantitative judgment tasks fall under this category as the group might decide to follow the judgment of one particular member, to weight the articulated opinions of their members equally or in any other constellation. How to form a meaningful group potential in quantitative judgments will be discussed in detail in the next chapter.

In the eyes of Steiner (1972), these different group potentials mark the optimal group performance. In his view, when groups perform at its absolute best, the group reaches only its

[1] Steiner (1972) classified tasks based on three dimensions. In the following, only the Interdependence dimension, which describes the relation between the individual contributions of the group members and the group potential, is explained. Another dimension differentiates between tasks that can be subdivided and unitary tasks that cannot. Either way, the group members must work on the same main task all the time, or different group members are able to work on subtasks in parallel. The presented research focuses only on unitary tasks. Steiner also differentiates between tasks, where the criterion of success is focused on the quantity (maximization tasks) outcome or the quality (optimization tasks) of the outcome.

expected potential. When the group outcome is below the optimal performance, the group is considered to suffer from process loss. In other words, Steiner describes the group process as detrimental to the group's performance. Framing the issue of group performance in this way might have contributed to the negative view on groups and might have led group researchers to focus mainly on the weaknesses of groups rather than on their strengths [2](Wang & Thompson, 2006).

Contrary to the relatively negative view on groups, some researchers also considered the possibility that groups outperform their potential, which has also been coined *process gains* (Hackman & Morris, 1975) or *synergy* (Larson, 2010). I will use both terms synonymously in this dissertation. Whereas Steiner (1972) defined the group potential as the ideal combination of the group members' resources, Larson (2010) set direct criteria for synergy. In particular, he defined synergy as a gain in performance that is attributable in some way to group interaction. When a group is able to do something collectively that could not have been reasonably achieved by any simple combination of their members' individual effort (Larson, 2010, p. 4).[3]

When evaluating group performance in quantitative judgments, I will consider the possibility of process losses and process gains. Therefore, it is important to define a group potential that takes into account which combination of individual efforts can be classified as *reasonable*. How the specific group potential should be defined in quantitative judgments, will be explained in the next section.

---

[2] Note that the general way of thinking about psychological research at that time might have also contributed to the focus on weaknesses of groups due to the fact that researcher from other areas of psychology also tended to concentrate their efforts primarily on weaknesses. For example, the focus on cognitive biases and irrational behavior (Tversky & Kahneman, 1974) led to many research programs pointing out the weaknesses of individual decision makers.

[3] In his framework, Larson (2010) differentiates between weak and strong synergy. Weak synergy is reached if the group outperforms the expectations based on the performance of the average group member. Strong synergy is reached if the group outperforms the expectations based on the performance of the group's best member.

1.2 Group potential in quantitative judgments

In a quantitative judgment task, participants have to guess a numerical value of interest that corresponds to present or future facts. A quantitative judgment task is an optimization task, where the quality of the judgment is the criterion for the performance. Specifically, no or just a small deviation between a judgment and the underlining true value would be an indicator of high performance or high judgment accuracy. It is not straightforward how a group should combine their members' resources to come to the best possible group judgment. Consider for example forecasting tasks; in these tasks, participants have to estimate a future value of an event that has not happened yet. How can they know, which group member has made the best individual judgment? What would be a reasonable baseline for the individual performance in this case?

There are several potential baselines that can and have been considered for the group performance in quantitative judgment tasks (Einhorn, Hogarth, & Klempner, 1977; Gigone & Hastie, 1997) but not all of them are theoretically or practically meaningful or represent a reasonable baseline from the group researcher's perspective. For example, one could select the best individual judgment made by any nominal group member as the relevant benchmark on any given task. In other words, the most accurate individual judgment made by any member of the nominal group would mark the criterion for the group. Accordingly, this model has been named accuracy model (Bonner, Sillito, & Baumann, 2007). However, this level of accuracy is not something that a group can reasonably accomplish, because the accuracy model highly capitalizes on chance (Gigone & Hastie, 1997). Specifically, beating this model, groups would have to recognize – with certainty – if a relatively weak member makes the best estimate by chance. Therefore, this model needs to be disregarded as a realistic benchmark for a group.

One alternative idea is to evaluate the performance of groups in quantitative judgment tasks by relying on the members' relative individual performance. For example, it is theoretically possible to rank order the overall individual performances of the members of the nominal group and then compare the group performance to the performances of these individuals. The important benchmark, in this regard, is the best performing member, which was introduced earlier as the relevant group potential in disjunctive tasks such as problem-solving tasks. The best member model represents an important benchmark that can easily be compared to other task types and has been used to evaluate group performance in quantitative judgments (e.g., Bonner & Baumann, 2008; Einhorn et al., 1977; Laughlin, Gonzalez, & Sommer, 2003; Sniezek & Henry, 1989). There are, however, some arguments against the use of the best member model as the criterion for synergy in quantitative judgment tasks. First, the process of identifying the best individual might already require a group process. Specifically, to determine the relative expertise of the group members and to find the best member, group interaction of some sort does need to take place. Second, similar to the accuracy model, the best member model does also, however to a lesser degree, benefit from chance. By selecting the person based on the best performance, it is likely that the performance of this person will deteriorate in the future due to regression to the mean. While the degree of this effect depends on the number of observations, it still overstates the accuracy of the group's potential.

It is also not advisable to go back to the comparison between the group performance and the performance of a single individual. Einhorn and colleagues (1977) showed that group judgments are more accurate than randomly selected individual judgments. Considering the ideas of Steiner (1966), this is not surprising. As it is the case for other unitary tasks, the pure size advantage of a group in comparison to an individual might already lead to an advantage of the group. In the particular case of quantitative judgment tasks, a statistical effect plays into the hand of the group, which does not require any group process. When aggregating individual judgments, for example, by building the average, this aggregation can be highly

accurate due to statistical error cancellation (e.g., Soll & Larrick, 2009). This effect, first observed over more than 100 years ago by Galton (1907)[4], is also commonly known as *Wisdom of the Crowd* (Surowiecki, 2004). There is a vast literature showing the benefit of averaging (e.g., see for an overview Mannes, Larrick, & Soll, 2012). Averaging is quite often the most accurate solution in many forecasting tasks (Merkle, Saw, & Davis-Stober, 2020). However, averaging does not represent a real group effect (Stroop, 1932), because it does not require a group to form the average of independent individual judgments. Therefore, the average of a comparable number of independent individual judgments is often considered the relevant individual criterion for synergy (Minson et al., 2018; Schultze, et al., 2012).

I will consider the average as the main criterion or synergy because the average model represents a theoretical and reasonable alternative to group work. When the group members are unable to agree on a common solution through group discussion, they might just meet in the middle and form the average. In this sense, forming the average can be seen as a substitute for a group judgment process, without any interaction or exchange of arguments. In fact, there is accumulating evidence that groups are able to surpass the average of a comparable number of independent individual judgments and thereby achieving synergy (e.g., Keck & Tang, 2019; Schultze et al., 2012; Sniezek & Henry, 1989).

While I have discussed the group potential for the evaluation of group performances in quantitative judgments in the current section, I will introduce mechanisms affecting the performance of groups in quantitative judgment tasks in the next section.

1.3 Group mechanisms affecting group performance

The performance of a group depends on the individual contributions of the group members as well as group specific components resulting from group work. The mechanisms,

[4] Galton (1907) reported the median.

leading to process gains or process losses, can take place on these different dimensions. Hackman and Morris (1975) identified three levels where group mechanisms can affect the group performance: group members' individual effort/motivation, group members' capability/resources, and the groups' tasks performance strategy (coordination).

While participating in a group can potentially have an influence on the group members' motivation[5], there is no evidence that motivational factors play a major role for quantitative group judgments. The outcome of quantitative judgment tasks is evaluated by its quality as quantitative judgment tasks can be categorized as optimization tasks. These tasks do not require much effort to be performed and increased motivation might not necessarily lead directly to better outcomes as in additive tasks[6]. The empirical results from Stern, Schultze, and Schulz-Hardt (2017) provide no empirical evidence that changes in the individual motivation of the group members affect their individual performance leading to synergy. In particular, group members did not show individual performance changes due to group members' awareness that they are part of a group. Stern and colleagues compared the individual performance between trials, in which the group members worked alone, and their individual performance in their first group trial prior to the first group interaction. While the group members were aware that they were part of a group, the individual performance did not decrease or increase. Therefore, lower or higher individual motivation due to group membership seem to have no measurable effect on the individual performances. Against this background, changes in motivation of group members are not considered in this dissertation.

---

[5] Group members might be more motivated to perform the tasks and therefore might have a higher individual performance leading to a better group performance. For example, participants in a swim relay, who have late starting positions in their team, do show improved individual performance resulting from motivation gains due to their participation in a team (Hüffmeier & Hertel, 2011). Processes losses due to motivation can also occur. For example, group members are less motivated to do a task if their individual contribution on the final group product is not identifiable (Latané et al., 1979).

[6] This is especially true for general knowledge questions, which are primarily used as quantitative judgment tasks. The group performance in more complex quantitative judgments tasks, such as multicue judgments tasks, might have a higher potential to be affected by motivational factors.

The second level concerns the individual capabilities of the group members. Participation in a group can potentially change the individual capabilities of its members, which in turn can affect the group performance. Examples are cognitive stimulation or cognitive restriction. In a brainstorming task, group members might provide new information during the group interaction. This might either help other group members to generate more ideas individually or hinder them in their thought process, which in turn might lead to fewer individually generated ideas (e.g., Nijstad, Stroebe, & Lodewijkx, 2002). The mechanism I will focus on is *group-to-individual transfer* (hereafter, G-I transfer). G-I transfer is concerned with the individual improvement of the group members' task performance as a result of the group members' participation in prior group work (Brodbeck & Greitemeyer, 2000). G-I transfer plays an important role in quantitative judgment tasks. In particular, group members are able to reduce their individual judgment errors because of their participation in the group (Schultze et al., 2012; Stern et al., 2017). In these group interactions, more knowledgeable group members can share information unknown to the other group members or explain their effective judgment policy to the other group members. Consequently, the improved individual knowledge of the group members can then lead to an improved group performance.

Finally, the last level concerns the group performance strategy or coordination. In other words, how group members combine their resources effectively. Coordination does necessarily take place on the group level. In many tasks, only process losses and not process gains due to coordination are possible, because often the group potential already implies a perfect coordination of the group members' contributions (Drewes, Schultze, & Schulz-Hardt 2021). One example for process loss may be within an additive task such as physical rope pulling (Latané et al., 1979). Within this task, the group potential is defined as the sum of the group members' individual strength to pull a rope. Pulling a rope together as a group can only lead to process loss due to imperfect synchronization of the group members' efforts.

Unlike in such additive tasks, coordination gains can potentially be detected in discretionary tasks like quantitative judgments. In fact, synergy in quantitative judgments has long been primarily attributed to effective coordination. The corresponding mechanism is named *differential weighting* (e.g., Bonner et al., 2007; Sniezek & Henry, 1989). Differential weighting is concerned with how group members combine their individual members' inputs to come up with a consensus group judgment. Groups can benefit from differential weighting in quantitative judgments if the group members are able to weight the more accurate individual judgments of their group members stronger when forming the groups' consensus judgment. On the one hand, groups could identify the most capable group member and weight their input stronger (expertise-based weighting). One the other hand, groups might also be able to assign greater weight to the most accurate judgment – irrespective of which member made it – on a trial-by-trial basis (accuracy-based weighting; Bonner et al., 2007). Both weighting strategies can potentially lead to an enhanced group judgment accuracy. In fact, differential weighting can surpass the accuracy of an unweighted average of the group members' judgments (see Bednarik & Schultze, 2015).

1.4 Disentangling G-I transfer and differential weighting

On a theoretical level, G-I transfer and differential weighting can easily be distinguished. However, to do this empirically, an experimental design needs to be used which is able to capture the individual contributions of the individual group members and which allows for a disentanglement of different group mechanisms (Schultze et al., 2012; Lippold, Schultze, & Schulz-Hardt, 2021). The so-called *alternating-individual-group design* (aI-G design; Schultze et al., 2012) fulfills this criterion.

The aI-G design is commonly conducted with an interacting group condition/real group condition, where the group members are able to discuss the tasks, as well as a nominal condition, where the same number of individuals perform the tasks alone. To assess the

individual capabilities of the group members before the group interaction, the design is subdivided into two phases, a *Practice Phase* and a *Group Phase* (see Figure 1). The practice phase is the same for the real group condition and the nominal group condition. In this phase, all participants work on a series of judgment tasks alone without any interaction. In the following group phase, the procedure differs for the real und nominal group conditions. For each trial, all group members of the real group first make an initial individual judgment alone without any group interaction. After the initial individual judgment, the group members talk to each other to come up with a group consensus judgment. This allows to capture changes in the individual performance of the group members over all trials in the group phase. In addition, the individual group members' contributions for every trial can be determined. Therefore, G-I transfer and differential weighting can be accounted for. As the nominal condition functions as the individual baseline condition, the individuals in the nominal condition work on the same tasks but without any interaction. In particular, participants in the nominal group condition also make two judgments in the group phase; however, both judgments are made alone. All experiments included in this dissertation were conducted with the aI-G design.

Previous experiments conducted with the aI-G design showed strong support for G-I transfer (Schultze et al., 2012, 2021; Stern et al., 2017) while only one experiment - under favoring conditions - showed support for differential weighting. In particular, differential weighting did only occur in a task with a population bias, where the judges overestimated the weight of small items presented in front of them (Stern et al., 2017). Therefore, the current state of the literature implies that the reduction of the group members' individual judgment errors through G-I transfer might be the main driver for synergy in group judgments.

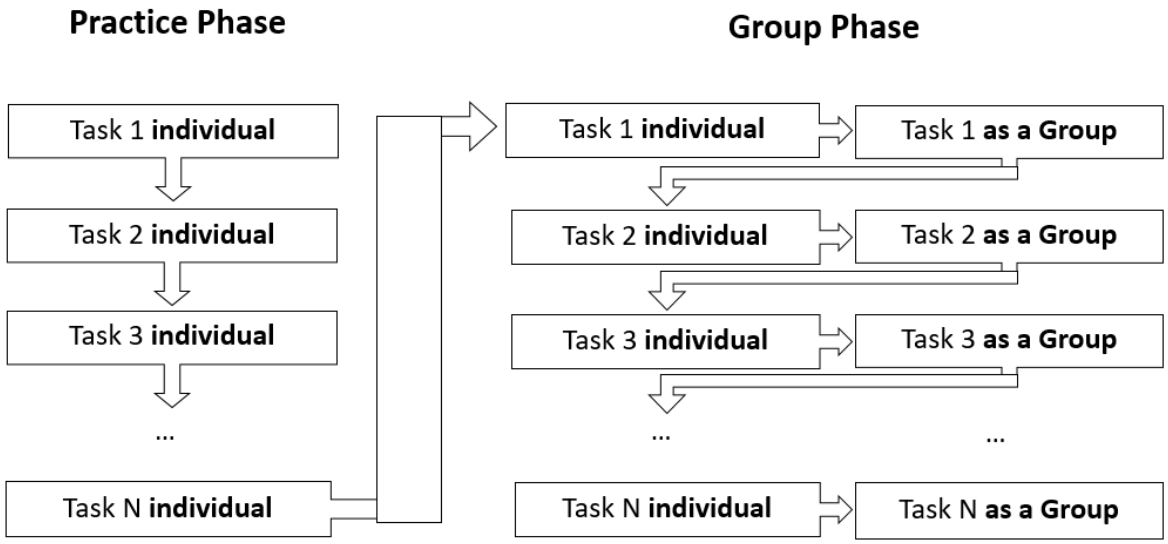

Figure 1. Display of the al-G design for the interacting group condition, adapted from Lippold et al. (2021).

1.5 Understanding G-I transfer

Follow up research on G-I transfer in quantitative judgment tasks investigated what is learned during the group discussions. To do that, Stern and colleagues (2017) relied on the taxonomy by Brown and Siegler (1993). Brown and Siegler decompose judgments into two separate knowledge dimensions: *metric* and *mapping*. The metric dimension focusses on basic distributional properties of the underlining criterion, for example the mean, the median, or the range of judgments. When the judge is wrongfully calibrated on the criterion, she commits metric errors by systematically over- or underestimating all entities from one category. For example, in one of the studies by Stern and colleagues, participants had to estimate the weight of small items (e.g., hammer or umbrella) by only looking at the items and not being allowed to touch or lift them. Most judges overestimated the weights of all of these items because they might have relied on a false frame of reference. In particular, they might have wrongfully considered the weight of a pencil to be 100 grams (true value approximately below 10 grams) and adjusted the weight of the other small items in accordance to this reference. Metric errors can only be measured when many judgments are considered, thereby allowing random errors

to cancel each other out. Metric errors can be measured by calculating the mean error over all judgments.

The mapping component is, in general, concerned with the relation between entities in one domain. In other words, not the overall accuracy of the single judgment is important for this component but rather the correct ranking of the individual items within a domain. Mapping errors are committed when the relative size of an item is not taken into account. In the example of weight judgments, a mapping error is committed when a judge assume that a pencil has a higher weight than a hammer. Mapping errors are measured indirectly. Not the error itself is measured but the mapping can be determined by taking the inverse of the Spearman's rank correlation over the items into question. This correlation index refers directly to the correspondence of the order of the judgments to the order of the actual underlying true values. Mapping and metric are both important error terms. Theoretically, they can be independent from each other. While one person might have knowledge or a correct intuition about the order of items in a domain (i.e., low mapping errors), the same person might conduct high metric errors as she overestimates the values of the items in general.

The findings from Stern and colleagues (2017) indicate that group members are able to reduce their individual metric errors via G-I transfer by recalibrating their individual metric, leading to an improved individual judgment accuracy. In other words, group members improved their individual metric knowledge by adjusting their belief about what constitutes a plausible range. Nevertheless, an improvement of the individual mapping knowledge was not detected in this study. Therefore, it is unknown if G-I transfer might also lead to an enhanced individual mapping knowledge.

1.6 Current research

Quantitative judgments seem to be one of the few tasks, where groups are able to reach synergy. In particular, there is recent evidence suggesting that group judgments are more accurate than the average of a comparable number of independent individual judgments (e.g., Keck & Tang, 2019; Minson et al., 2018; Schultze et al., 2012; Sniezek & Henry, 1989; Stern et al., 2017). Two mechanisms, G-I transfer, the improvement of the individual group members capabilities due to their participation in the group, and differential weighting, the group's ability to weight their members' input effectively, were identified as the main mechanism for this synergetic effect. While differential weighting (e.g., Bonner et al., 2007) was the first proposed mechanism, recent experiments indicate that the synergetic process can be mainly explained by G-I transfer (Schultze et al., 2012). The majority of experiments designed to disentangle both mechanisms show in most cases no evidence for differential weighting, when G-I transfer is controlled for (Schultze et al., 2012; Stern et al., 2017). Further analyses by Stern and colleagues on G-I transfer suggest that group members seem to learn only about the distributional properties of the quantitative judgment tasks (*metric knowledge*) but are unable to learn about the underlying relational properties in the domain of interest (*mapping knowledge*). This indicates that G-I transfer in quantitative judgments might only be based on the transfer of information, such as the general range, which is relatively easy to learn. Bonner and colleagues (2007) demonstrated that individuals could improve their metric knowledge, when they were provided with little extra information. Learning about the relational properties in a domain is in some cases crucial for the judgment accuracy. Therefore, it is important to know, whether G-I transfer generalizes to mapping knowledge. Against this background, the empirical studies in the first manuscript investigate whether group members are able to improve their mapping knowledge via G-I transfer.

While the first manuscript focuses on a specific mechanism, G-I transfer, the second manuscript takes a meta-analytic approach to investigate the performance of groups in quantitative judgments. As outlined above, there is recent evidence for groups' judgments to be more accurate than the average of a comparable number of independent individual judgments. However, the extent of the group advantage over the statistical aggregate of individual judgments is still up for debate. Previous systematic reviews were not completely convinced of such synergy in group judgments. Hastie (1986), for example, assumed that if there would be an advantage of group judgments over the average of individual judgments, this advantage would only be small. The latest review on group judgments by Gigone and Hastie (1997) stated that in previous research, the accuracy of the group judgments was on the same level as the average of individual judgments, but *not* surpassing it. However, due to the methodological differences of the group judgment studies, the computation of meta-analyses was impossible. Accordingly, several meta-analyses were conducted in the second manuscript. In addition, to gain a better understanding of synergy in quantitative judgments, the second manuscripts also takes an in-depth analysis of the different components behind this synergetic effect.

2 Summary of Manuscript 1: G-I transfer in multicue judgment tasks: Discussion improves group members' knowledge about target relations

The main goal of the first manuscript was to investigate whether mapping knowledge can be transferred in quantitative judgments by examining what group members are able to learn during group interactions. To do that, the current work relied on the judgment error decomposition by Brown and Siegler (1993), which is based on metric and mapping errors. Up to this dissertation, G-I transfer in quantitative judgments tasks was only associated with the reduction of the group members' individual metric errors (Stern et al., 2017). However, metric knowledge is relatively easy to learn as it does rely only on a few pieces of

information, such as the range or the median. In fact, group discussions might not even be necessary for the transfer of metric knowledge to occur. The improvement of one's mapping knowledge, on the other hand, might be more difficult and might take more time due to the fact that the exact relation between cues and target values needs to be learned. Therefore, it is important to evaluate if G-I transfer can also improve the transfer of mapping knowledge.

While the only two previous experiments conducted on this subject by Stern and colleagues (2017) did not reveal any evidence for the transfer of mapping knowledge, we cannot conclude that G-I transfer of mapping knowledge does not exist. Although the two experiments by Stern and colleagues were conducted with the previously introduced al-G design (Schultze et al., 2012), some procedural choices made them less suited for the detection of a potential transfer of mapping knowledge. In particular, the relatively low number of group trials in these experiments (Stern and colleagues reported 10 group trials in their two experiments) might not have been sufficient for the group members to improve their mapping knowledge given that learning about cue target relations takes more time than learning about the metric properties of a specific task. This leads to a second problem affecting the measurement of mapping knowledge. When measuring the individual mapping knowledge of participants, it is important that knowledge transfer has been completed as mapping knowledge is measured via a correlation coefficient over consecutive trials. Calculating this mapping knowledge based on mixed trials, where in some of them the group members might have still relied on their previous inferior mapping knowledge, whereas in other (later) trials, they might already have used their updated and improved mapping knowledge, might lead to an underestimation of the group members' actual mapping knowledge. To tackle these problems and to conclusively test whether participating in a group discussion could also reduce individual mapping errors, we used a modified version of the aI-G design as well as a new multicue judgment tasks with an increased number of trials. In particular, the modified design consisted of three phases with 30 multicue judgment tasks in

each phase. In addition to a practice and a group phase, which followed the above described procedure, the modified version of the al-G design contained an additional individual phase at the end of the experiment (post phase). The post phase mirrored the practice phase, thereby allowing for a calculation of the individual mapping knowledge of the group members for the practice and the final phase without bias.

The results of the first experiment showed G-I transfer in the form of an improved overall individual accuracy in the real group condition in comparison to the nominal condition. In particular, members of the interacting group condition reduced their error while members of the nominal condition did not. However, with regard to mapping, the results were not as clear. In the real group condition, participants' individual mapping knowledge was greater in the post phase than in the practice phase while in the nominal group condition, participants' individual mapping knowledge did not differ between post phase and practice phase. Nevertheless, the predicted interaction of these two factors could not be confirmed. Specifically, we could not confirm that participants' individual mapping knowledge increased more from the practice phase to the post phase in the real group condition as compared to the nominal group condition. An additional analysis revealed a baseline problem: In the practice phase, the mapping knowledge in the nominal groups was already higher than the mapping knowledge of the real groups. In other words, our randomization procedure might have not worked properly. Therefore, the differences in the improvements between nominal and real group condition could be due to differences in the baseline in Phase 1 (i.e., real group members' mapping knowledge improved because they had more potential to learn in the first place). Accordingly, in the second experiment, we were able to replicate the findings of the first experiment and experimentally controlled for baseline differences in individual mapping knowledge by using parallelization instead of randomization.

The results of this second experiment could confirm G-I transfer on all error types. Participants who worked in the real groups outperformed participants who worked

individually in the nominal groups with regard to the overall accuracy and the metric. With regard to mapping, we could confirm that participants' individual mapping knowledge increased more in the real group condition as compared to the nominal group condition. Furthermore, an additional analysis showed that the real groups' judgments in the group phase were more accurate than the average of the individual judgments in the nominal groups. In other words, the groups achieved synergy by outperforming the nominal groups. Interestingly, the groups' judgment showed also less mapping errors than the average models of the nominal groups. In sum, these two experiments showed that G-I transfer does generalize to multicue judgment tasks. Furthermore, it suggests that G-I transfer is not just restricted to the improvement of individual metric knowledge, but can also lead to an increase in individual mapping knowledge. These results emphasize G-I transfer as an important benefit of tasking groups with quantitative judgments.

3 Summary Manuscript 2: The benefit of collaboration: Disentangling the sources of synergy in group judgments

The main goal of the second manuscript was to investigate systematically to what degree groups are able to achieve synergy in quantitative judgments and to disentangle the underlining sources of synergy. For this purpose, multiple meta-analyses based on altogether eleven experiments conducted with the al-G design (Schultze et al., 2012) were computed. Only experiments from our lab in Göttingen were included in these analyses due to the fact that we were unable to find any other studies with the particular experimental design required to differentiate between the two relevant mechanisms, namely G-I transfer and differential weighting. The tasks in these experiments included a wide range of quantitative judgments such as distance estimates, forecasting tasks, and multicue judgment tasks. In a first step, several meta-analyses were conducted to compare the accuracy of groups with meaningful individual benchmarks.

Following the rationale outlined in section 1.2, group performance exceeding the average of judgments made by a comparable number of individuals working alone indicates synergy in group judgments (e.g., Minson et al., 2018; Schultze et al., 2012; Stern et al., 2017). The corresponding meta-analysis on eight experiments showed that group judgments were more accurate than this baseline. The meta-analysis revealed a moderate effect size (d = 0.62), which provides substantial evidence that groups are able to achieve synergy in quantitative judgment tasks. Notably, in two additional meta-analyses, we were able to show that the groups also outperformed the average model with regard to the mapping and metric error components.

In addition to the average model, meta-analyses on the best member model (e.g., Bonner & Baumann, 2008) were conducted to gain more insights into the extend of the group performance. As the name suggests, the best member model refers to the best individual of a nominal group. To calculate the best member model, the member with the best performance during the group phase needs to be identified, and her performance is then selected as the best member model. The corresponding meta-analysis showed that the groups' judgments were as accurate as the judgments of the best members of the nominal group. Note that in many tasks, groups do not even reach the level of the best group member (Larson, 2010). Therefore, this result highlights the high standing of group performance in quantitative judgments in comparison to the group performance in other tasks. Based on this finding, it might seem reasonable not to rely on groups and use the best member of the nominal group instead. However, as indicated in the introduction, the best member model is solely of theoretical value. In practice, there is no group phase to identify the best member. Either the practitioner must be able to fortune tell the future or she has to rely on previous performances to estimate the expertise of each individual group member.

There is an alternative to the best member model that does, indeed, have the potential to be practically applied, and that was also considered in the second manuscript: the initial

best member model. When group members complete tasks individually prior to the group work, their performance can be evaluated. Based on these performances, the initial best member can be identified. All experiments considered employed a practice phase, which allowed for the identification of the initial best member. Therefore, another meta-analysis was conducted, which revealed that the group judgments were more accurate than the judgments of the nominal groups' initial best members. Hence, this finding suggests that performance wise, groups may not be substituted by the best individual without loss of judgment accuracy.

While the previous meta-analyses clearly showed a performance advantage of group judgment over the statistical aggregation of individual judgments, we also examined the mechanisms leading to these synergetic gains. Further analyses could clearly provide meta-analytic evidence for the two proposed mechanisms: G-I transfer and differential weighting. Previous findings of G-I transfer could be reproduced (Lippold, Schulz-Hardt, Schultze, 2021; Schultze et al., 2012; Stern et al., 2017), that is individuals working in groups increased their individual accuracy as a consequence of this collaboration. Again, we relied on the taxonomy by Brown and Siegler (1993) and decomposed the judgment errors into their mapping and metric components. As indicated by previous research, G-I transfer was mainly the result of an improved metric component. While a meta-analysis showed that group members improved their individual metric knowledge from the practice to the group phase, group members did, on average, not seem to improve their individual mapping knowledge during the group phase. However, as we have described in the first manuscript, the transfer of mapping knowledge might take longer than the transfer of metric knowledge. Therefore, we also conducted meta-analyses on those experiments, which were specifically designed to capture a transfer of mapping knowledge. These four experiments had an additional post phase, which was added to obtain an unbiased measure of the individual mapping knowledge (as explained in the first manuscript). The corresponding meta-analysis showed that group members significantly improved their individual mapping knowledge because of working together.

So far, the focus was on the improvement of the individual capabilities of the group members, but we also looked into group coordination. In particular, across several meta-analyses, we investigated differential weighting, i.e., groups' ability to weight their members' input effectively. As has been outlined before, when controlling for G-I transfer, most previous studies did not find evidence for differential weighting (Lippold et al., 2021; Stern et al., 2017). Contrary to these previous findings, our meta-analysis revealed a small but significant overall effect for differential weighting. In other words, there was evidence for coordination gains in groups. Given that the effect size was relatively small, we assumed that the power within most individual experiments was too small to detect an effect of differential weighting. Based on our meta-analysis, however, we can conclude that groups are able to weight their members' input based on expertise. That is, groups benefited from coordination in the group. We further explored the resulting question of how groups weight their members' input effectively.

There are two possible weighting strategies, which are in line with our analyses. We did find that the group judgments were as accurate as the judgments of the best members of the groups. Therefore, the groups might have been able to identify their best member and then only rely on the input of this particular member. While this strategy is in line with the data, we argue that the assumption that groups actually use this strategy might not be very realistic. In the absence of any feedback, groups might struggle to identify their best member (Kurverts et al., 2019). Additionally, other motives such as fairness concerns (Mahmoodi et al., 2015) might play a role and may prevent a group from solely relying on one group member's input. Alternatively, we proposed another potential adjustment strategy based on the detailed analysis of error decomposition. In fact, the benefit of the group judgments over the average model of their members' individual estimates was mostly due to improvements in the mapping error component. Based on this finding, we propose a *mapping error monitoring strategy*, assuming that the group members weight their members' input based on centrality

by weighting individual judgments closer to the average stronger (as postulated earlier by Bonner and Baumann, 2012). At the same time, the group members monitor if a group judgment would be consistent with the rank order of the previous judgments. When forming the average would lead to an incorrect order compared to the previous judgments, the group would adjust their judgment accordingly.

In sum, the meta-analyses conducted in the second manuscript demonstrate that groups are able to outperform a comparable number of individuals in quantitative judgments. This effect is the result of both G-I transfer and differential weighting, and it affects both the metric and the mapping component of their judgments.

4 General Discussion

In the two manuscripts presented in this dissertation, group judgments were systematically analyzed by examining the underlying group judgment error structures and disentangling different group mechanisms. While the first manuscript investigated G-I transfer of mapping knowledge in depth, the second manuscript provided a more comprehensive view on the comparison between group judgment accuracy and individual accuracy by conducting meta-analyses based on multiple experiments from our lab. Both manuscripts suggest that groups can benefit from information exchange in quantitative judgments.

The first manuscript demonstrated that, in addition to group members improving their metric knowledge (Stern et al., 2017), they are also able to boost their mapping knowledge through group discussion. In other words, group members do not only learn simple metrical information about a specific domain during the group discussions, but are also able to learn more complex cue target relations, which indicates a substantial knowledge transfer. This finding contributes to the research on group learning by demonstrating the extent of G-I transfer in quantitative judgments. Note that G-I transfer is not restricted to quantitative

judgment tasks as it has also been observed in a variety of other tasks. In particular, G-I transfer has been previously identified in groups working on collective induction tasks (Brodbeck & Greitemeyer, 2000), in groups solving reasoning tasks (Laughlin, Carey, & Kerr, 2008), and most recently in groups controlling dynamic systems (Schultze, Drewes, & Schulz-Hardt, 2021). Therefore, relying on interacting groups in organizations is an effective way to enhance the individual capabilities of the workforce.

The second manuscript emphasizes the performance advantage of interacting groups over nominal groups in quantitative judgments by demonstrating unequivocal meta-analytic evidence that groups are able to and systematically do reach synergy in this type of task. In particular, groups reached synergy as their judgments were more accurate than the average based on a comparable number of independent judgments from individuals working alone. While in many areas of group research, interacting groups usually do not outperform nominal groups and group performance falls even short of their groups' potential (e.g., Kerr & Tindale, 2004; Steiner, 1972), our results are in line with more recent findings of synergy in quantitative judgments (e.g., Keck & Tang, 2019; Minson et al., 2018; Schultze et al., 2012). Based on this, we can infer that in practice professionals should use interacting groups over simply averaging independent individual judgments if their aim is to ensure high judgment accuracy. While forming the average of independent individual judgments might not be the common choice for tackling quantitative judgment tasks in practice, we also evaluated another alternative. In particular, professionals might also consider tasking the person with the best-known capabilities to perform judgment tasks. However, when we compared the groups' accuracy with the accuracy of the initial best member from the nominal groups, the groups were again more accurate. In other words, the initial best individual cannot substitute the group without risking accuracy losses.

In addition to evaluating the overall extent of synergy in quantitative group judgments, the second manuscript also explored the mechanisms leading to synergy. As suggested by

previous research (Schultze et al., 2012; Stern et al., 2017), synergy in quantitative group judgments was the result of G-I transfer. The achieved synergy can be attributed to an improvement of the individual capabilities of the group members. Noteworthy, we also found meta-analytic evidence that groups were able to engage in beneficial coordination by weighting their members' inputs effectively. Specifically, we can now assume that synergy in group judgments might not only be the result of group members' improvement due to their participation in group discussion, but also due to effective coordination within the group.

Overall, the findings of this dissertation highlight the benefits of collaboration in quantitative judgment. I will briefly outline implications, limitations and future research in the next sections.

4.1 Implications

Based on the presented findings, some general implication for group research can be derived. It needs to be emphasized that changes in the group members' individual capabilities might play an important role for the group performance and potential synergy. The presented empirical results of this dissertation and previous findings (Schultze et al., 2012; Stern et al., 2017) indicate that improvements of the individual capabilities of the group members can lead to synergy in quantitative judgments. Apart from these findings, only research on brainstorming tasks has emphasized the relation between changes in individual capabilities due to group work and group performance[7]. Despite this crucial role of individual capabilities for group outcomes in these two task types, the link between changes in the individual capabilities of the group members and the group performance seem to be understudied. One reason might be that it is in general difficult to track the individual capabilities during experiments. Only complex research designs like the al-G design (Schultze et al., 2012) make

[7] In particular, group discussion can affect the group member's ability to generate ideas through cognitive stimulation or cognitive restriction (Nijstad, Stroebe, & Lodewijkx, 2002).

it possible to closely monitor the capabilities of the group members in more sophisticated tasks. Such complex designs, where participants switch constantly between working individually and working in a group, however, are not often used in group performance research.

In addition, numerous studies on the improvement of the individual capabilities of the group members, G-I transfer, were conducted on disjunctive tasks such as problem solving (e.g., Laughlin & Ellis, 1986). This line of research focused on G-I transfer and created ideal conditions for knowledge transfer between stronger and weaker groups. In these tasks, more knowledgeable group members are able to demonstrate the correct solution to the other group members, which constitutes G-I transfer. However, the performance of the best member defines the group potential in these tasks. The best member herself has the slightest chance of performance improvement. Therefore, it is quite difficult to find synergy due to G-I transfer in this type of task, masking the potential relation between the changes in individual capabilities and the group performance. In general, changes in the individual capabilities might play a more important role in understanding group performance and should be warranted a stronger consideration in the future.

Further thoughts should be given to the potential strengths of groups. In the past, group researchers have shown a rather pessimistic view on group performance and focused their research primarily on the weaknesses of groups (Wang & Thompson, 2006). As it has been outlined in section 1.1, the highly influential work by Steiner (1972) defined the group potential in a way that only process losses due to group interaction were possible. This way of argumentation had a tremendous influence on many group researchers. Accordingly, group research has systematically investigated the weaknesses of groups since the 70s. There are still many critiques of group work and group performances (e.g., Armstrong, 2006), however, the findings of this dissertation join recent research in identifying strengths of group work. For example, participating in a group can lead to motivation gains benefitting the overall

group performance (Hüffmeier & Hertel, 2011). These motivation gains do often even offset potential motivation losses in field research (Hüffmeier et al., 2021). Groups can even outperform their best members in some problem solving tasks (e.g., Laughlin et al., 2003). The strengths of groups in quantitative judgment tasks as presented in this dissertation should also be clear. All of these findings might help to improve the bad reputation of group work by highlighting the benefits of collaboration.

In order to identify further benefits of groups, small group research on interacting groups is necessary and must still play an important role in social psychological. However, group research on interacting groups seems to be on the decline. Due to the replication crisis in Psychology (Open Science Collaboration, 2015), there is a general trend in social psychological to focus on research designs which allow for higher sample sizes and as a result higher statistical power. Accordingly, research in social psychology has shifted its focus away from interacting groups to online studies, vignette studies, and studies, which primarily use self-report measures (Sassenberg & Ditrich, 2019). These research approaches do not require as much resources per participant as studies with actual groups do. While it is reasonable and necessary to rely on higher sample sizes, the shift away from actual human interaction and away from actual behavior might hinder the progress of group research. Most online studies in social psychology tend to focus on the individual perception of specific interdependent situations and not on the collaboration of individuals. Although, these vignette studies can be useful for some areas, they might only be able to approximate a group member's actual behavior. Note that the correlation between self-measures and actual behavior tends to be rather weak (Baumeister, Vohs, & Funder, 2007), reducing the external validity of the conducted research. Field studies and observations might represent an intriguing alternative for group researchers, nevertheless, they often come with the high burden of insufficient experimental control. In contrast, group experiments with the focus on real interacting groups and actual group performance, which can be measured, evaluated, and compared to an

individual's performance, provide valuable, reliable data for the research community. Therefore, it is important to preserve group research in experimental settings.

4.2 Limitations and future research

The research conducted for this dissertation was set out to systematically analyze group judgment accuracy and to investigate two important mechanisms that can contribute to the improvement of this accuracy, namely G-I transfer and differential weighting. Exploring other potential moderators affecting the group performance and the mechanism was outside of the scope of this dissertation. Nevertheless, in practice, several other factors may have an impact on group judgment accuracy and the underlining group mechanisms and should be further investigated in future research.

The first factors concerns group size. All experiments reported in the two manuscripts where exclusively conducted with groups consisting of three individuals. Hence, it remains unclear whether our results may also generalize to groups of different sizes. A study by Minson and colleagues (2018) revealed that interacting dyads can be more accurate than the average based on two independent individual judgments, which by our definition does indicate synergy. Nevertheless, there is limited knowledge on group performance of interacting groups with more than three group members. On the one hand, having more people in a group might increase the potential for synergy, as the additional group members might have extra individual knowledge, which may help improve group performance. On the other hand, the criterion for synergy - the average model - may be harder to pass given that the statistical benefit of averaging in the form of error cancelation does increase with the number of included independent judgments (e.g., Einhorn et al., 1977). In addition, with a higher group size the chance of a particular group member to contribute actively to the group might decrease. Therefore, it will be interesting for future research to investigate whether

interacting groups consisting of more than three individuals are also able to outperform the corresponding nominal groups with the same number of individuals.

Another potential overlooked factor arises from the fact that different organizations rely on different group decision-making structures. For example, in some organizations, the members of the governing board make all decisions based on the principle *one person one vote* while in others, the votes of the shareholders count in relation to their company shares. There are also many organizations, which rely on a decision-making structure with a sole individual having the power to make all decisions on her own. Empirically, there is some evidence suggesting that groups with one member being given the full decisional power (with the other members only advising this member) might have a higher judgment accuracy than groups who are asked to find a democratic consensus judgment (Keck & Tang, 2019). However, the used experimental design within this research does not allow for the disentanglement between G-I transfer and differential weighting. Therefore, it remains unclear whether a change in decision-making structure leads to an extended G-I transfer or to more effective coordination in the group. In order to fill this gap, future research should replicate this finding using the al-G design.

The last limitation concerns the external validity of the findings presented in this dissertation. In the tasks used in our experiments, we mostly relied on general world knowledge judgments with known true values. However, the most important quantitative judgment task groups face outside of the laboratory are forecasting tasks. The defining feature of realistic forecasting tasks is uncertainty. For example, the economic forecasts on the world economy made in 2019 turned out to be all wrong as the repercussions of the ongoing Corona crisis of 2020/21 affected the worldwide economic outlook in an unexpected way. While it is literally impossible even for groups to make accurate forecasts under unpredictable conditions of an uncertain world, future research can compare the adaptability of groups and nominal groups to changing environments in quantitative judgment tasks. For example, groups and

nominal groups could be tasked with predicting the value of an artificial stock index. In such an experiment, all participants would be provided with some predictor variables affecting the stock index and constant time-delayed feedback about the market rate of this stock. Based on the feedback, the participants can learn the relationship between the predictors and the stock index. However, in the course of the experiment, the initial relationship between the predictors and the stock index could change, forcing the participants to adapt their learned belief about the relation between the predictors and the stock index. Within such a design, it would be interesting to see whether groups are able to alter their judgment policy faster than nominal groups and to explore additional benefits of collaboration.

5 Conclusion

There is still a lot of skepticism towards the use of groups with regard to their performance. Nevertheless, calls for the avoidance of interacting groups may have been premature. The findings of the current dissertation highlight the benefits of group work in quantitative judgments. Tasking groups with quantitative judgments leads to high performance as groups do outperform the same number of individuals. The benefits of collaboration are not only restricted to the group performance. Being part of a group can also have positive effects on the individual group members with their individual capabilities increasing due to their participation in a group. Therefore, this work might help in convincing critics that the use of groups can still be worthwhile.

6 References

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting*, (5), 3-15.

Baumeister, R. F., Vohs, K. D., & Funder, D. C.(2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396-403.

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making*, *10*(3), 265-276.

Bonner, B. L., & Baumann, M. R. (2008). Informational intra-group influence: the effects of time pressure and group size. *European Journal of Social Psychology*, *38*(1), 46-66.

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology*, *102*(2), 337.

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.

Brodbeck, F. C., & Greitemeyer, T. (2000). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes and Intergroup Relations, 3,* 159-182.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*(3), 511.

Drewes, S., & Schultze, T. (2021). *An integrative review of research on process gains and losses in groups.* Unpublished Manuscript.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In *Advances in Experimental Social Psychology* (Vol. 8, pp. 45-99). New York, NY: Academic Press.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.). *Decision Research* (Vol. 2). Greenwich, CT: JAI Press.

Hüffmeier, J., & Hertel, G. (2011). When the whole is more than the sum of its parts: Group motivation gains in the wild. *Journal of Experimental Social Psychology*, *47*(2), 455-459.

Hüffmeier, J., Hertel, G., Torka, A. K., Nohe, C., & Krumm, S. (2021). In field settings group members (often) show effort gains instead of social loafing. *European Review of Social Psychology*, 1-40.

Galton, F. (1907). Vox populi. *Nature, 75*, 450-451.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*(1), 149.

Keck, S., & Tang, W. (2019). Elaborating or aggregating? The joint effects of group decision-making structure and systematic errors on the value of group interactions. Unpublished Manuscript.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-655.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., ..., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), eaaw9011.

Larson, J. R. (2010). *In search of synergy in small group performance*. New York, NY: Psychology Press.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822.

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11*, 319-330.

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

Laughlin, P. R., Gonzalez, C. M., & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice, 7*, 55-63.

Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. K. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective strategies. *Journal of Personality and Social Psychology, 85*, 684 – 694. http://dx.doi.org/10.1037/0022-3514.85.4.684

Lippold, M., Schulz-Hardt, S., & Schultze, T. (2021). G-I transfer in multicue judgment tasks: Discussion improves group members' knowledge about target relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(3), 532–545. https://doi.org/10.1037/xlm0000947

Lippold, M., Schultze, T., & Schulz-Hardt, S. (2021). *The benefit of collaboration: Disentangling the sources of synergy in group judgments.* Unpublished Manuscript.

Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Social Judgment and Decision Making* (pp. 227-242). New York, NY: Psychology Press.

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

Merkle, E. C., Saw, G., & Davis-Stober, C. (2020). Beating the average forecast: Regularization based on forecaster attributes. *Journal of Mathematical Psychology*, *98*, 102419.

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*(9), 4177-4192.

Nijstad, B. A., Stroebe, W., & Lodewijkx, H. F. (2002). Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of Experimental Social Psychology*, *38*(6), 535-544.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Ringelmann, M. (1913). Recherches sur les moteurs animés: Travail de l'homme. *Annales de l'Institut National Agronomique, 12*, 1-40.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107-114

Schultze, T., Drewes, S., & Schulz-Hardt, S. (2021). A test of synergy in dynamic system control tasks. *Journal of Experimental Psychology: General, 150*(5), 890–914. https://doi.org/10.1037/xge0000975

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes, 118*, 24-36.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*, 1-28.

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

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science, 11*(4), 273-283.

Steiner, I. D. (1972). *Group Process and Productivity*. New York, NY: Academic Press.

Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). How Much Group is Necessary? Group-To-Individual Transfer in Estimation Tasks. *Collabra: Psychology, 3*(1).

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15*(5), 550.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations, 296.*

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Wang, C. S., & Thompson, L. L. (2006). The negative and positive psychology of leadership and group research. *Advances in Group Processes, 23,* 31-61.

Appendix A.

Manuscript 1

Lippold, M., Schulz-Hardt, S., & Schultze, T. (2021). G-I transfer in multicue judgment tasks: Discussion improves group members' knowledge about target relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 532–545. https://doi.org/10.1037/xlm0000947

G-I transfer in multi-cue judgment tasks:

Discussion improves group members' knowledge about target relations

Matthias Lippold, Stefan Schulz-Hardt, & Thomas Schultze

University of Göttingen

Leibniz ScienceCampus Primate Cognition, Göttingen

Author Note

Correspondence regarding this article should be addressed to Matthias Lippold, Georg-Elias-Müller Institute for Psychology, Economic and Social Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. E-mail: matthias.lippold@uni-goettingen.de. This research was supported by grants of the Deutsche Forschungsgemeinschaft (DFG) to the second and last authors (SCHU 1279/11-2; SCHU 2813/2-2). The data and analysis scripts for this manuscript are available at the Open Science Framework (https://osf.io/925wh/). Both studies in this manuscript were also pre-registered at the Open Science Framework (Experiment 1: https://osf.io/mpfbu, Experiment 2: https://osf.io/zkbpg).

Abstract

One benefit of working in groups is that group members can learn from each other how to perform the task, a phenomenon called group-to-individual transfer (G-I transfer). In the context of quantitative judgments, G-I transfer means that group members improve their individual accuracy as a consequence of exchanging task-relevant information. This improved individual accuracy allows groups to outperform the average of a comparable number of individuals, that is, G-I transfer leads to synergy. While there is mounting evidence that group members benefit from G-I transfer in quantitative judgment tasks, we still know rather little about what exactly group members learn from each other during this transfer. Here, we build on the distinction between metric knowledge (knowing what constitutes a plausible range of values) and mapping knowledge (knowing the relative magnitude of the targets) to gain further insights into the nature of G-I transfer. Whereas previous research found evidence that G-I transfer improves group members' metric knowledge, there is, so far, no evidence that group discussion also improves mapping knowledge. Using a multi-cue judgment task, we tested whether group members would benefit from G-I-transfer and, if so, whether this G-I transfer would manifest in the form of improved mapping knowledge. The results of two experiments suggest that this is the case. Participants who worked in real interacting groups outperformed participants who worked individually, and this increase in accuracy was accompanied not only by improved metric, but also by increased mapping knowledge.

Keywords: group judgment, group performance, group-to-individual transfer, quantitative estimates, group learning

G-I transfer in multi-cue judgment tasks: Discussion improves group members' knowledge about target relations

Quantitative judgments play a fundamental role in most people's lives. These judgments include seemingly trivial instances such as estimating the amount of food needed for the next family dinner. On the other hand, judgments can also have far-reaching consequences because they form the basis for many important political, entrepreneurial, or medical decisions. Groups are often consulted with judgment tasks because they produce – on average – more accurate judgments than individuals (Gigone & Hastie, 1997). This is not surprising because groups benefit from statistical aggregation of their members individual judgments, which leads to error cancellation. This increase in accuracy due to statistical aggregation is commonly known as the wisdom-of-the-crowds-effect (Surowiecki, 2005). Importantly, groups benefit from discussion beyond the aggregation of their members' judgments. There is mounting evidence that group judgments are more accurate than the simple average of a comparable number of individual judgments, that is, group interaction leads to synergy in quantitative group judgment (e.g. Minson, Mueller, & Larrick, 2017; Schultze, Mojzisch, & Schulz-Hardt, 2012; Sniezek & Henry, 1989; Stern, Schultze, & Schulz-Hardt, 2017). Recent research suggests that the main process driving the superiority of group judgments is group-to-individual transfer (G-I transfer), a group-specific learning effect that denotes increases in group members' individual accuracy due to exchanging task relevant information during discussion (Schultze et al., 2012; Stern et al., 2017).

While it is confirmed that group members *do* learn from each other, we still know rather little about *what* they learn. In general, we can decompose the relevant knowledge in quantitative judgment tasks into two components: *metric* knowledge and *mapping* knowledge (Brown & Siegler, 1993). Metric knowledge refers to knowledge about the scale a target is measured on and, in particular, what constitutes a plausible range of values. For example, a person with good metric knowledge of car prices in Germany might know that used compact

cars have a plausible price range of €4,000 to €30,000. Mapping knowledge, in contrast, refers to knowledge about the position of specific targets on a scale. For example, a person with good mapping knowledge about car prices will know that a used car from a premium car manufacturer is more expensive than a used car from a high-volume manufacturer. Metric knowledge is based on relatively simple information that is easy to learn, such as the range of the target values, or plausible frames of reference. G-I transfer of metric knowledge can also be substituted easily by providing individual decision-makers with frames of reference (Bonner, Baumann, & Sillito, 2007). In contrast, improving one's mapping knowledge is more difficult, and likely takes longer, as it requires an improved understanding of the relationship between cue characteristics and target values (Brown, 2002). Thus, if group interaction also improved group members' mapping knowledge, one could make an even stronger case for employing interacting groups rather than relying on the wisdom of non-interacting crowds when aiming to maximize judgmental accuracy.

To the best of our knowledge, only one previous study investigated G-I transfer with regard to mapping and metric knowledge (Stern et al., 2017). The results of this study showed that group members improved their metric knowledge during discussion, but found no evidence of improvements in mapping knowledge. However, as we will point out below, the tasks employed in this particular study as well as the study design may have been not particularly well suited to detect G-I transfer of mapping knowledge. Accordingly, the main goal of the present research is to provide a conclusive test of group members' ability to increase their mapping knowledge as a function of group discussion.

The sources of synergy in quantitative group judgment

Group judgments are often more accurate than the average of a comparable number of individual judgments (e.g., Bonner & Baumann, 2012; Schultze et al., 2012; Stern et al, 2017). There are currently two explanations for this synergy in group judgments: *differential weighting* (e.g., Bonner et al. 2007; Sniezek & Henry, 1989) and the above-mentioned *G-I*

transfer (Schultze et al., 2012; Stern et al., 2017). Differential weighting connotes that the group members assign more weight to judgments that are more accurate. For example, groups could identify the member with the best overall performance and weigh this member's input more (expertise-based weighting), or they might assign more weight to the most accurate individual judgment for a specific trial, irrespective of which member suggested it (accuracy-based weighting; Bonner et al., 2007). If groups succeed at identifying their experts and/or the most accurate judgments, differential weighting can outperform an unweighted average of group members' input (see Bednarik & Schultze, 2015, for a formal analysis).

G-I transfer, on the other hand, refers to the improvement of group members' individual accuracy due to group interaction (Brodbeck & Greitemeyer, 2000). In the context of group judgments, this means that, after a group discussion, group members' individual accuracy increases beyond mere practice effects due to the exchange of task-relevant information (Schultze et al. 2012; Stern et al., 2017). For example, the most capable group member could explain important underlying principles that the weaker group members then use to enhance their judgment accuracy. Note that differential weighting and G-I transfer are not mutually exclusive phenomena; they can rather have additive effects on group judgment accuracy. That is, even if the weaker group members` individual capability to perform the task increases due to G-I transfer, group members' individual pre-discussion estimates might still differ to some extent, leaving the group with the opportunity to weight these individual judgments by (perceived) accuracy or expertise.

Experiments conducted with the so-called *aI-G* design (alternating-individual group design) can differentiate between differential weighting and G-I transfer (Schultze et al., 2012). In the group condition of an *aI-G* design, individual and group judgments alternate, that is, for each task, the group judgment is formed directly after the group members' individual judgments. Therefore, this design can measure an improvement of participants' individual accuracy in the group condition by focusing on the individual estimates. Schultze

and colleagues (2012) showed that, while members of real and nominal groups initially performed equally well, individual accuracy increased in the real groups but not the nominal groups, that is, real group members benefited from G-I-transfer whereas nominal ones did not. At the same time, Schultze and colleagues (2012) did not find evidence for differential weighting when controlling for G-I transfer. In particular, they compared the accuracy of real group judgments to the average of the real group members' corresponding individual pre-discussion judgments. In general, experiments conducted with this *aI-G* design showed continuous support for G-I transfer (Schultze et al., 2012; Stern et al., 2017). In contrast, differential weighting only occurred under specific circumstances, and only with a rather low effect size (Stern et al., 2017). Therefore, G-I transfer plays an important role in explaining the effectiveness of groups in group judgments. However, it remains unclear how G-I transfer works and what kind of individual knowledge is transferred in groups.

Understanding G-I transfer

To investigate what is learned via G-I transfer in quantitative judgment tasks, we follow the approach set forth by Stern et al. (2017), namely to decompose the overall judgment error. This approach draws from the taxonomy of Brown and Siegler (1993) we introduced earlier. However, rather than measuring the extent of metric and mapping knowledge, the idea is to focus on the *lack* of this knowledge, which manifests as specific kinds of judgment errors, namely metric and mapping errors. Since metric knowledge comprises distributional attributes of the criterion, such as the mean, median, range, or the underlying form of the distributional function, metric errors are made when a judge is incorrectly calibrated on the criterion, leading to a systematic over- or underestimation of entities from the same category. Metric errors can be measured via the mean error based on all judgments (Stern et. al., 2017). This measure indicates the degree of systematic error. Positive [negative] values indicate a tendency to overestimate [underestimate] the targets, on average. When overestimations and underestimations of the true values fully cancel each other out

across multiple judgments, the metric error is zero. In contrast, mapping knowledge pertains to the relation between targets in a domain. Accordingly, a mapping error is committed when the relative magnitude of the targets is determined incorrectly. Mapping can be measured using the Spearman`s rank correlation between the estimated and the true values, which indicates the correctness of the order of the judgments. A correlation of 1 indicates that a person is able to rank order all targets by magnitude without error. A value of zero means that a person has virtually no idea about which targets are relatively large and which are small.

Stern and colleagues (2017) did not find evidence for transfer of mapping knowledge; G-I transfer was mostly the result of a reduction of metric errors in their study. This pattern was observed with two different experimental tasks. In their first experiment, participants had to estimate the distances between European cities, whereas in their second experiment, participants estimated the weights of objects. In both experiments, the observed G-I transfer of metric knowledge required only one single group interaction, with the individual improvement in accuracy being stable over the subsequent trials that did not involve any further social interaction. Hence, brief social interaction seems to be sufficient for the transfer of metric knowledge.

However, the absence of evidence for any transfer of mapping knowledge in the study by Stern et al. (2017) need not necessarily indicate a general inability of group members to learn mapping knowledge during group interaction. Instead, it is possible that the Stern et al. (2017) experiments were not ideally suited to test for transfer of mapping knowledge. Arguably, mapping knowledge is more difficult to learn than metric knowledge, and doing so may require more time. Specifically, the exact relation between cues and targets might only be learned during multiple subsequent trials. It is difficult to say, a priori, how many trials of group interaction are necessary to produce a noticeable increase in group members' mapping knowledge, and how long it takes for this type of G-I transfer to be complete. Given that the

experiments reported by Stern et al. (2017) contained relatively few group trials[1], it is conceivable that there might have been somewhat insufficient opportunities for group members to improve their mapping knowledge.

The relatively low number of trials entails a second problem that may have masked possible increases in participants' mapping knowledge: As already outlined, the measure of mapping knowledge is the rank correlation of a participant's estimates, on the one hand, and the true values, on the other hand. Ideally, one would compute one correlation coefficient prior to any opportunity to learn from the other group members (i.e., a baseline measure of mapping knowledge), and a second one once group members have exchanged all relevant knowledge (i.e., after possible G-I transfer of mapping knowledge is complete). However, computing the second correlation may be difficult when there are not enough data points, for example, because mapping knowledge increased relatively late (e.g., after the 8th out of 10 trials). Computing mapping knowledge based on only a few trials is ill advised, because it would produce highly unreliable estimates. Computing mapping knowledge across all 10 trials is not a solution either, because it will produce biased estimates. When some of the judgments are based on group members' original (and inferior) mapping while the others already benefit from improved mapping knowledge, the inconsistencies between these two sets of judgments will artificially lead to lower rank order correlations.

Detecting Transfer of Mapping Knowledge

A conclusive test of G-I transfer with regard to mapping knowledge requires some changes to the experimental *aI-G* design described above. Specifically, we argue that such a modified design must meet two criteria. First, in addition to the individual practice phase and the group phase, it should entail a third phase where participants work individually again (i.e.,

[1] 20 group trials in Experiment 1, and 10 group trials in Experiment 2.

a post-group phase). Adding this post-group phase allows measuring individual mapping knowledge after the group phase without bias.

Second, the design needs to include a sufficient number of trials per phase for two reasons. On the one hand, the reliability of the measurement for mapping knowledge increases the more trials are included (Brown, 1910; Spearman, 1910). On the other hand, since we do not know a priori how many trials are necessary for the transfer of mapping knowledge to take place, having more trials increases the chances for detecting such transfer. However, we also need to take into account that the more trials participants need to work on, the more they might become bored with the task. Thus, in an attempt to balance these costs and benefits, we decided for 30 trials per phase.

An informative test of G-I transfer with regard to mapping knowledge also requires a judgment task that meets two criteria. First, G-I transfer requires stable differences in expertise. The idea of weaker members learning from stronger members makes sense only if there are differences in group members' ability to perform the task (Bonner & Baumann, 2004). By the same logic, improvements in group members' mapping knowledge should be contingent on stable differences between group members' mapping knowledge. Second, in the chosen task, mapping knowledge must be learnable in principle. That is, at least if participants work on the task under optimal conditions for learning, namely with immediate feedback after each trial, their mapping should improve.[2]

The aim of the present study was to test for G-I transfer of mapping knowledge using such an informative design and a task that meets these criteria. To this end, we conducted two pre-registered experiments, in which participants worked on a set of multi-cue judgment tasks, either in real groups or in nominal groups of three members.

Experiment 1

[2] Brown and Siegler (1993) show that such improvements in individuals' mapping knowledge under ideal conditions are possible, in principle.

Experiment 1 aimed to provide a first conclusive test of G-I transfer of mapping knowledge. To this end, participants worked on a multi-cue judgment task in three phases of thirty trials each. All participants worked on the task individually in Phases 1 and 3, while in Phase 2 they worked on the task in either real or nominal groups. Since previous studies on G-I transfer in quantitative group judgment exclusively used general knowledge tasks (Schultze et al., 2012; Stern et al., 2017), the first question of interest is whether G-I transfer generalizes to multi-cue judgments tasks. On the one hand, finding evidence of G-I-transfer is the prerequisite for our main goal, namely testing whether it (also) manifests on the level of mapping knowledge. In addition, testing for G-I transfer allows making some statements about the generalizability of the effect that has already been found in previous studies. Therefore, we proposed the following hypothesis:

Hypothesis 1: Participants' individual accuracy increases more from Phase 1 to Phase 3 in the real group condition as compared to the nominal group condition.

While we treat Hypothesis 1 as the comparison of a difference measure (the increase in individual accuracy) between two groups, it is, conceptually, the test of an interaction of time and group type. Thus, we further specified the exact nature of this interaction by postulating two additional hypotheses. Specifically, we predicted that this interaction is due to an increase in participants' individual accuracy in the real groups (Hypothesis 1a) and no increase in the nominal groups (Hypothesis 1b).

In case that the data support Hypothesis 1 as well as Hypotheses 1a and 1b, the crucial question is whether G-I transfer manifests as improved mapping knowledge, leading us to postulate the following hypothesis.

Hypothesis 2: Participants' individual mapping knowledge increases more in the real groups from Phase 1 to Phase 3 as compared to the nominal groups.

As with Hypothesis 1, Hypothesis 2 conceptually represents an interaction effect. Therefore, we again specified the nature of this interaction, predicting an increase in

participants' individual mapping knowledge in the real groups (Hypothesis 2a) and no increase in the nominal groups (Hypothesis 2b).

We did not specify hypotheses about G-I transfer with regard to metric knowledge because the task we used placed some restrictions on the range of possible values, which somewhat limits the occurrence of metric errors and, thus, the potential for G-I transfer of metric knowledge. However, we report exploratory analyses testing whether participants' individual metric knowledge increased more in the real groups from Phase 1 to Phase 3 as compared to the nominal groups. Experiment 1 was pre-registered at the Open Science Framework (https://osf.io/mpfbu).

Method

Procedure. We invited up to 12 participants to each session and randomly assigned them to three-person groups. The groups, in turn, were randomly assigned to one of two conditions: the real group and the nominal group condition. In cases where the number of participants who showed up could not be divided by three, the remaining participants were allocated to the nominal group condition. Participants were placed in separate rooms (i.e., one room per real group/nominal group) and seated at individual computers at different ends of the same table. The experiment was programmed in the open source software *Alfred* (Treffenstaedt & Wiemann, 2018). Prior to the beginning of the experiment, participants were asked to provide informed consent on the computer screen. In the consent form, the participants were informed that they might be filmed.

For our multi-cue judgment task, we asked participants to estimate the prices of used cars from a specific German car manufacturer as accurately as possible based on several cues. To this end, we retrieved publicly available data on 90 used cars from a German car search-engine website, treating the displayed car prices as the true values, and information on a number of car features as cues. The following cues were available to the participants: the car's type (exact model), information on the year of the car's initial registration, the driven

kilometers, the horse power, the gear type (manual vs. automatic), fuel type (gasoline vs. diesel), and fuel consumption (represented in the German standard: liters per 100 kilometers). We then formed three relatively similar subsets of 30 used cars each. All participants worked on all three sets, but the order of the sets and the order of stimuli within each set was randomized for each group (i.e., the same order of stimuli was applied to all persons taking part in the same group). On each trial, participants were asked to provide their best estimate of the price of the automobile. Note that all numeric estimates in all phases were restricted to values between 100 Euro and 200,000 Euro to prevent extreme outliers. Participants were also asked to rate their confidence in the accuracy of their estimate on a 7-point Likert scale ranging from 1 (‘not at all confident) to 7 (‘very confident’).

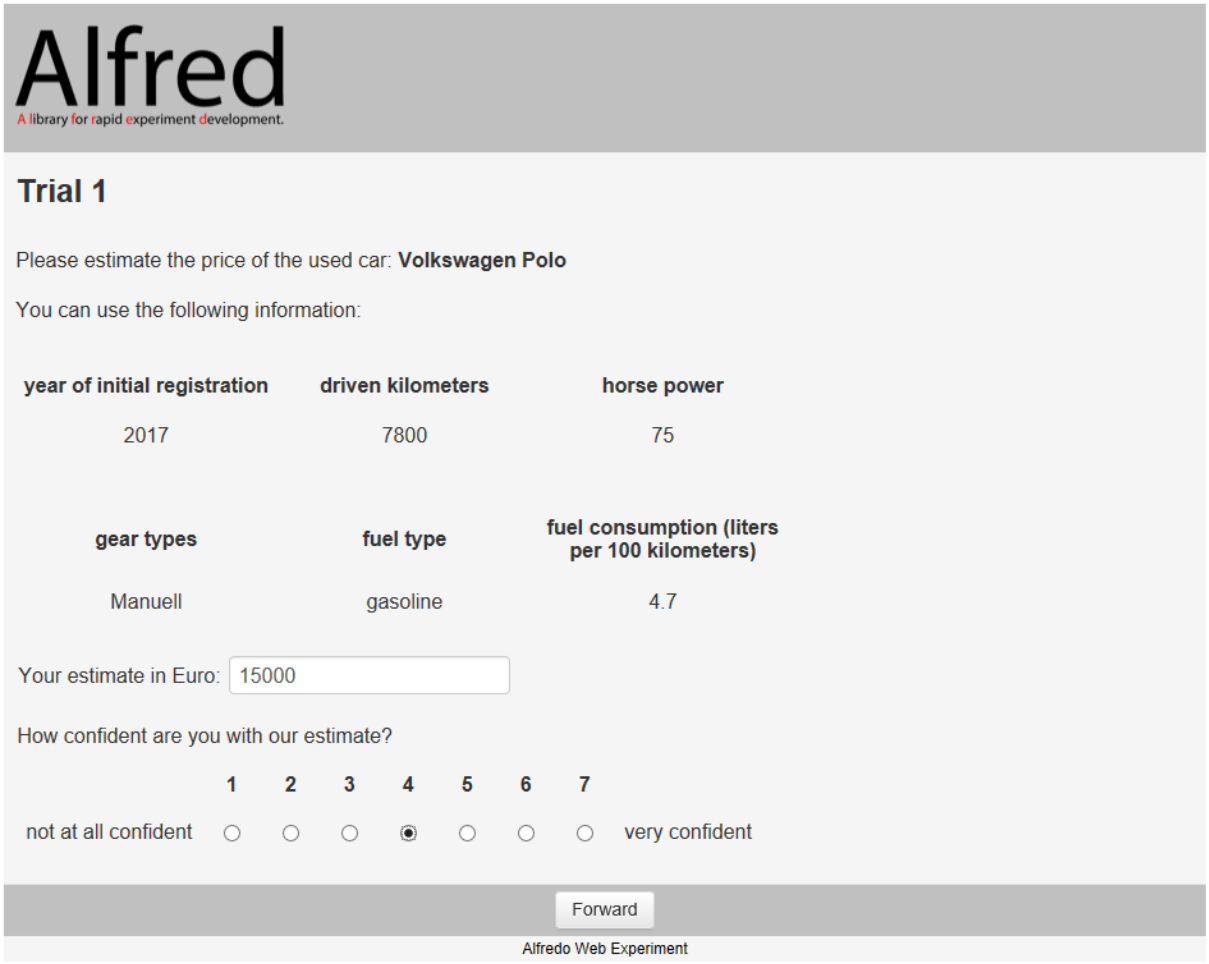

Figure 1. Display of an example trial (translated to English). This example shows an individual-level trial as presented to participants in Phases 1 and 3 of the experiments.

Across both conditions, participants started by working on 30 judgment tasks on their own in an individual practice phase (Phase 1). In the group phase (Phase 2), the procedure differed according to the assigned condition. In the real group condition, participants worked on the 30 tasks in the *aI-G* design. On each trial, all group members first submitted an individual initial estimate of the price of the used car, again accompanied by a confidence rating. Once all group members had provided their individual judgments, the group discussion started. The initial judgments of all group members were displayed on each group member's computer, and group members were asked to consult with their fellow group members to come up with a consensus estimate and a joint confidence judgment of their group estimate. Once all group members had entered the consensus judgment and confidence rating on their computers, the next trial started. In the nominal group condition, participants worked on all 30 estimates individually. In order to avoid a confound between the experimental condition and the number of judgments per trial, nominal group participants had to provide two estimates per trial. After their first estimate, participants were asked to estimate the price of the same car once more, with their initial judgment being displayed on screen. The second estimate was also accompanied by a confidence rating. Note that we videotaped each group in both conditions in Phase 2. Because we were only interested in the group discussions, we only kept the recordings for the groups in the real group condition.[3]

Regardless of the experimental condition, all participants worked on another 30 trials on their own in Phase 3. Following Phase 3, all participants filled in a final questionnaire. This questionnaire contained questions regarding participants' demographics (age, gender), a suspicion check ("What do you think was the purpose of this study?", open answer format), and a question asking participants whether they had worked on the task in a serious manner

[3] We conducted qualitative and quantitative analyses of the videos from Experiment 2. The aim of these analyses was to test for possible links between aspects of the group discussions and the amount of G-I-transfer. The analyses did, however, not reveal sufficiently reliable insights. Therefore, we do not report them here, but instead added them as an online supplement (https://osf.io/7e8w5/).

('yes' vs. 'no'). For exploratory purposes only, and depending on the condition, participants were asked open-ended questions about their strategies in Phase 2. In particular, members of the real group condition were asked how they came up with their group judgments. Members of the nominal group condition were asked to provide details on how they came up with their individual judgments in general terms.

Participants' payment for taking part in the study consisted of a fixed reimbursement of €10 for completing the study and a performance-based bonus payment of up to €9. The bonus was determined as follows: Participants made 30 individual estimates in Phase 1, 30 individual/initial estimates and 30 group/revised estimates in Phase 2, and 30 individual estimates in Phase 3. The computer program randomly drew one estimate from each phase. For each of the three estimates that deviated less than 10% from the true value, the participant received €3 (for a maximum bonus of 3 × €3 = €9). This payoff scheme was explained to participants in the initial instructions.

Dependent Variables. The dependent variables of interest are variables at the level of individual (nominal) group members. However, due to interaction and interdependence in the real group, data of real group members are not statistically independent. To account for this interdependence, we aggregated all indices on the (nominal) group level.

***Accuracy*.** We measured the overall accuracy using the mean absolute percentage error (MAPE), which is defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|estimate_i - true\ value_i|}{true\ value_i} * 100 \qquad (1)$$

Individual Accuracy Gain. We computed the differences between participants' individual MAPE scores in Phase 1 and their individual MAPE scores in Phases 3. Positive values indicate an increase in individual accuracy.

***Mapping Knowledge*.** As the indicator of mapping knowledge, we calculated Fisher-*z*-transformed Spearman's rank correlations between the true values and the respective estimates within each phase of the experiment.

***Individual Mapping Knowledge Gain*.** We computed the differences between participants' mapping knowledge in Phase 3 and their mapping knowledge in Phase 1. Positive values indicate an increase in individual mapping knowledge.

***Metric knowledge*.** For exploratory reasons, we also calculated participants' metric error as the mean percentage error (MPE) as well as the individual metric knowledge gain, defined as the difference between participants' metric error in Phase 1 and that of Phase 3 (again, positive values indicate an increase in metric knowledge). [4]

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{estimate_i - true\ value_i}{true\ value_i} * 100 \qquad (2)$$

Pretest. For our experimental task, a pretest revealed that it fulfilled the above-mentioned requirements. Similar to the main experiment, the pretest consisted of three phases, each comprising 30 trials of the multi-cue judgment task. We grouped the 90 trials of judgment task into three sets of 30 trials each, so that items were roughly similar between sets in terms of difficulty as well as in terms of variety of used cars. The item sets were presented in randomized order. The order of the 30 items within each set was also randomized. In Phases 1 and 3, participants worked on the 30 trials without any form of feedback. In Phase 2, feedback in the form of the true values was provided to the participants. A comparison between Phase 3 and Phase 1 allowed us to evaluate the ability to learn when performing the task under optimal feedback. The pretest showed that there was sufficient variability with regard to participants' mapping knowledge (median: 0.72; range: 0.45-0.93), overall accuracy (median: 57.88; range: 21.35-281.68) and metric knowledge (median: 29.76; range: 0.38-

[4] Note that metric error can also be measured by using the median overall deviation (MOD), which is the average difference between the median of all judgments and the median of the corresponding true values (Brown & Siegler, 1993).

280.93) in Phase 1 of the pretest. Second, a one-tailed *t*-test revealed that participants were able to improve their mapping knowledge from Phase 1 (M = 0.98, SD = 0.29, ρ = 0.75) to Phase 3 (M = 1.10, SD = 0.36, ρ = 0.80), $t(49) = 2.18$, $p = .017$, $d = 0.31$. Another one-tailed *t*-test revealed a reduction of the overall error from Phase 1 (M = 74.24, SD = 55.46) to Phase 3 (M = 57.61, SD = 35.85), $t(49) = 1.83$, $p = .037$, $d = 0.26$). Additionally, a one-tailed *t*-test revealed that participants were able to improve their metric knowledge from Phase 1 (M = 50.61, SD = 63.65) to Phase 3 (M = 29.97, SD = 39.07), $t(49) = 1.93$, $p = .030$, $d = 0.27$. That is, it is generally possible for participants to acquire better mapping knowledge in this task.[5]

Sample size and stopping rule. We applied Bayesian sequential testing (Schönbrodt et al., 2017). Accordingly, we relied on the Bayes factor of the alternative hypothesis (BF_{10}) as our inference criterion. The BF_{10} is an odds ratio indicating how likely the alternative hypothesis is relative to the null hypothesis given the observed data. We considered Bayes factors greater than 3 or smaller than 0.33 to be sufficient evidence for our alternative or the corresponding null hypothesis. As we predicted directional effects between the group and nominal condition, we defined the null hypothesis for the Bayes Factor as an interval from $d = -\infty$ to $d = 0.20$ (Morey & Rouder, 2011) to conduct an appropriate one-tailed test. We began by initially testing 30 groups per condition (180 participants). In accordance with our pre-registration, data collection would continue past this initial sample until a) both of the two BF_{10} for Hypotheses 1 and 2 showed evidence for the null hypothesis ($BF_{10} < 0.33$) or evidence in favor of the alternative hypothesis ($BF_{10} > 3$) or b) we hit a sample size of 51 groups per condition (upper boundary reflecting our budgetary constraints).

We deviated from the original plan and stopped collecting data after testing 30 groups per condition because the first inspection of the data revealed that the randomization of

[5] One reviewer pointed out that an improvement in individual mapping knowledge as observed in our pre-test might be an interesting finding in its own right, because previous evidence for individuals improving in mapping knowledge is still largely lacking. Therefore, we have made the full data of the pretest as well as our analysis code publicly available at the Open Science Framework (https://osf.io/hujr6/).

experimental conditions was not sufficiently effective. Specifically, we observed differences in mapping between real group and nominal group members in Phase 1 that make the interpretation of the results concerning Hypotheses 1 and 2 difficult (see results). Despite these baseline differences, the data are not entirely uninformative, and some tests pertaining to our hypotheses are still feasible. Thus, we proceeded with the planned data analysis, but we also decided to run a second study to provide a clear test of our hypotheses.

Data exclusion. We excluded participants from the analysis based on the following pre-registered criteria: First, if they reported in the final questionnaire that they did not work on the task seriously; second, if they failed to comply with the instructions concerning the discussion of tasks (i.e., nominal group members discussed the target values, real group members discussed their individual pre-discussion estimates, or real group members discussed their individual estimates in Phases 1 or 3). Exclusion of one participant in the real group condition necessitated excluding the whole group from the data analysis. Participants/groups excluded along these lines were replaced to ensure the aspired sample size. Furthermore, we excluded outlier judgments on a trial-by-trial basis. All individual judgments which deviated by more than five standard deviations from the average individual judgment for that specific task were excluded.

Since there were some participants with a substantial number of excluded trials (in some cases we had to exclude more than twenty trials in one phase), we had to implement an additional exclusion criterion. Thus, other than specified in the pre-registration, we removed participants from the final analyses if more than seven of their estimates (i.e., more than 25%) had to be excluded in one of the three phases. We had to implement this criterion to guarantee the precision of our mapping measure. Note that the results would not change if we used fewer trials as a criterion.

Participants. Our final sample comprised 180 German-speaking undergraduate and graduate students (117 women, 62 male and 1 other) of different faculties from the University

of Goettingen, with a median age of 23 years (range: 18-56). They were recruited via the participant database ORSEE (Greiner, 2015).

Results

We conducted all analyses with the statistical software R (Version 3.5.1; R Core Team, 2018). The reported Bayesian *t*-tests (Rouder et al., 2009) were calculated using the BayesFactor package (Morey & Rouder, 2015).

Phase 1 individual performance. As mentioned above, we observed some unexpected baseline differences between real and nominal groups. We investigated these differences using two Bayesian *t*-tests for independent samples. We used the then default prior for the effect size, namely the JZS prior (Rouder, 2009) with the default scaling factor of $r = 1$.[6] Since we were interested in testing whether the baseline performances were about equal, we used two-tailed Bayesian *t*-tests with a point null hypothesis. With regard to overall accuracy, there were no baseline differences between real ($M = 59.51$, $SD = 11.29$) and nominal groups ($M = 59.87$, $SD = 23.75$), $t(41.48) = 0.07$, $p = .942$, $BF_{10} = 0.19$, $d = 0.02$, allowing us to conclude that initial accuracy did not differ systematically between conditions. However, an analysis of participants' Phase 1 mapping knowledge indicated a substantial problem: We detected descriptive baseline differences between the real ($M = 0.96$, $SD = 0.25$, $\rho = 0.74$) and the nominal groups ($M = 1.09$, $SD = 0.22$, $\rho = 0.80$). In particular, the respective Bayesian *t*-test could not confirm that the individual mapping knowledge was similar in the nominal and real groups, $t(58) = 1.97$, $p = .053$, $BF_{10} = 1.09$, $d = 0.51$. Since the experimental procedure in the two conditions did not differ in Phase 1 (the experimental manipulation of group type took place after Phase 1 and prior to Phase 2), a possible baseline difference would indicate an ineffective randomization. That is, unfortunately, the random assignment of

[6] Note that the default scaling factor in the BayesFactor package was recently changed to $r = 0.707$. Thus, we refer to the default at the time we pre-registered Experiment 1.

participants to the two group conditions led to an imbalance regarding participants' mapping knowledge.

Individual accuracy gains. We compared participants' individual accuracy gains between the two conditions (real groups, nominal groups) using Bayesian *t*-test for independent samples, with the default JZS prior and a scaling factor of r = 1. We applied a one-tailed version of the test by defining the null interval between $d = -\infty$ and $d = 0.20$. The individual accuracy gains in the real group condition ($M = 6.15$, $SD = 14.29$) were higher than the individual accuracy gains in the nominal group condition ($M = 0.30$, $SD = 12.90$), $t(58) = 1.63$, $p = .051$, $BF_{10} = 4.45$, $d = 0.43$. Having found substantial evidence in favor of Hypothesis 1, as indicated by the BF_{10} exceeding 3, we next tested Hypotheses 1a and 1b to determine the exact nature of the observed difference. Separate paired-sample *t*-tests revealed that group members' individual accuracy in the real groups improved from Phase 1 to Phase 3, $t(29) = 2.36$, $p = .013$, $BF_{10} = 8.60$, $d = 0.46$, while the accuracy in the nominal groups remained unchanged, $t(29) = 0.13$, $p = .449$, $BF_{10} = 0.24$, $d = 0.01$. These results confirm both Hypothesis 1a and Hypothesis 1b. A detailed overview of participants' individual accuracy is displayed in the upper panel of Figure 1.

Individual mapping knowledge gains. We compared participants' individual mapping knowledge gains between the two conditions (real groups, nominal groups), again using the same one-tailed Bayesian *t*-test for independent samples (JZS prior with a scaling factor of $r = 1$, null interval between $d = -\infty$ and $d = 0.20$). The individual mapping gains in the real group condition ($M = 0.08$, $SD = 0.25$, $\rho_{Phase\ 1} = 0.74$, $\rho_{Phase\ 3} = 0.78$) were not substantially higher than the individual mapping gains in the nominal group condition ($M = 0.01$, $SD = 0.26$, $\rho_{Phase\ 1} = 0.80$, $\rho_{Phase\ 3} = 0.80$), $t(58) = 1.08$; $p = .142$; $BF_{10} = 1.80$, $d = 0.28$. Since the BF_{10} neither exceeded 3 nor fell below 0.33, the results of this test remain inconclusive, that is, we can neither accept nor reject Hypothesis 2. Yet, separate paired-

sample *t*-tests revealed that the individual mapping knowledge in the real groups improved from Phase 1 to Phase 3, $t(29) = 1.78$, $p = .043$, $BF_{10} = 3.40$, $d = 0.33$, while mapping knowledge in the nominal groups remained more or less unchanged, $t(29) = 0.24$, $p = .408$, $BF_{10} = 0.29$, $d = 0.04$, supporting Hypotheses 2a and 2b. A detailed overview of participants' individual mapping knowledge is displayed in the lower panel of Figure 2.

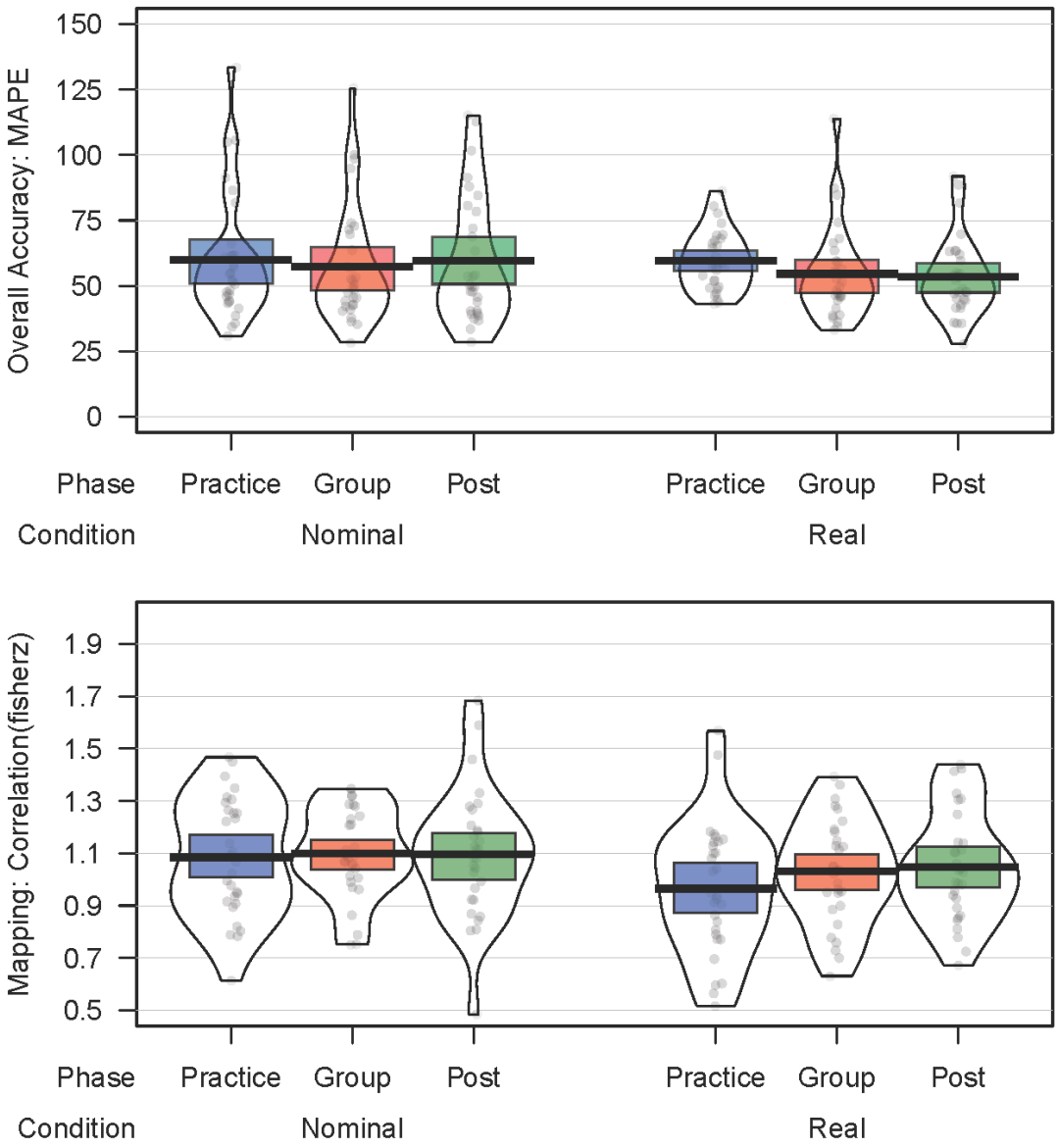

Figure 2: Individual Accuracy (upper panel) and individual mapping knowledge (lower panel) by phase and condition. In the upper panel, the overall accuracy in the form of the mean absolute percentage error is displayed. In the lower graph, mapping knowledge measured via fisher z transformed Spearman`s rank correlation is shown. Note that the points represent the

data points, the bold horizontal black lines represent the means, and the colored rectangles represent the 95% highest density intervals. The beans indicate the distribution of the data, with their width corresponding to the estimated density at the respective level of the y-axis.

Discussion

Interpreting the results of Experiment 1 is somewhat difficult because of the suboptimal randomization which led to baseline differences in participants' mapping knowledge. While this baseline difference limits the interpretability of some results, we can still make a number of valid inferences. First, the results indicate that G-I transfer generalizes to multi-cue judgment tasks. In particular, individuals who had worked in real groups increased their overall accuracy, whereas members of nominal groups did not. Second, we were able to confirm that participants' individual mapping knowledge in the real group condition was greater in Phase 3 than in Phase 1, whereas participants' individual mapping knowledge in the nominal group condition did not differ between these phases. Although this pattern fits our postulated hypotheses regarding the increase in mapping knowledge, we could not confirm that the increase in mapping knowledge actually differed between conditions. Yet, the differences in individual mapping that we observed between group types in Phase 1 were no longer evident in Phase 3. However, there are two possible interpretations. One is that, due to G-I transfer of mapping knowledge, real group members could catch up with the initially more knowledgeable nominal group members. Nevertheless, it is also possible that greater increases in real group members' individual accuracy reflected their greater potential to improve rather than genuine group processes. Put differently, a ceiling effect may have prevented nominal but not real group members from improving their mapping knowledge, thus artificially creating the observed results.

In sum, based on the results of the first experiment, it remains unclear to which extent group interaction is beneficial with regard to reducing individual mapping errors. Against this background, we replicated Experiment 1. However, this time we experimentally controlled for

baseline differences in individual mapping knowledge in Phase 1 by using parallelization instead of randomization.

Experiment 2

The aim of Experiment 2 was to test the same hypotheses as in Experiment 1, but without the risk of baseline differences between conditions. In addition to using parallelization in Experiment 2, we also made an additional prediction regarding participants' Phase 3 mapping knowledge. Assuming that there would be no baseline differences in mapping knowledge between conditions, and given that we would find G-I transfer of mapping knowledge, the logical conclusion would be that former real group members should possess better mapping knowledge than former nominal group members in Phase 3. We added this prediction as a third hypothesis. As with Experiment 1, Experiment 2 was also pre-registered at the Open Science Framework (https://osf.io/zkbpg).

Hypothesis 3: Participants' individual Phase 3 mapping knowledge is greater among participants who were members of real groups in Phase 2 as compared to nominal group members.

Procedure. The procedure was similar to that of Experiment 1 with the following exception: to prevent baseline differences between the real and the nominal groups, we allocated the participants to the two conditions based on their mapping errors in Phase 1. In particular, the experimenter tracked participants' Phase 1 mapping errors on the fly via a computer stream programmed with the open source experimental software *Alfred* (Treffenstaedt & Wiemann, 2018). Based on this information, the experimenter balanced the individual mapping of the groups between the two conditions. In addition, we used the knowledge about participants' Phase 1 mapping errors to create heterogeneous groups with regard to mapping knowledge. Weaker group members were paired with stronger group members, which should facilitate learning effects in the real group condition (Bonner & Baumann, 2004). Note that to allow for parallelization, this time we also had to change the

randomization of the stimuli. All participants in one session (as opposed to all participants in one real/nominal group) received the same order of items. From Phase 2 onwards, Experiment 2 was identical to Experiment 1.

Sample size and stopping rule. As in Experiment 1, we used Bayesian sequential testing. However, we included the Bayes Factor of Hypothesis 3 in the stopping rule for Experiment 2. That is, we started by collecting 30 groups per condition and planned to gather data until all three of the focal BF_{10} showed either evidence in favor of the null hypothesis ($BF_{10} < 0.33$) or evidence in favor of the alternative hypothesis ($BF_{10} > 3$).

Data exclusion. The criteria for excluding participants were identical to Experiment 1 with one addition: Participants were excluded if we had to exclude more than five of their estimates in at least one of the three phases.

Participants. In Experiment 2, 180 German-speaking undergraduate and graduate students (108 women, 71 male, and 1 other) from different faculties of the University of Goettingen with a median age of 25 years (range: 18-55) made up the final sample of the study.

Results

Phase 1 individual performance. We compared participants' Phase 1 mapping knowledge using a Bayesian *t*-test for independent samples with the default JZS prior and a scaling factor of $r = 1$. We used a two-tailed Bayesian *t*-test with a point null hypothesis. This test showed that the change from randomization to parallelization was successful in preventing baseline differences. Specifically, with regard to mapping knowledge, there were no baseline differences between real ($M = 0.99$, $SD = 0.16$, $\rho = 0.76$) and nominal groups ($M = 0.98$, $SD = 0.14$, $\rho = 0.75$), $t(58) = 0.32$, $p = .748$, $BF_{10} = 0.20$, $d = 0.08$. Furthermore, with regard to the overall accuracy, there were also no baseline differences between real ($M = 63.62$, $SD = 22.97$) and nominal groups ($M = 66.72$, $SD = 19.63$), $t(58) = 0.56$, $p = .576$, $BF_{10} = 0.22$, $d = 0.15$.

Individual accuracy gains. As in Experiment 1, we compared participants' individual accuracy gains between real groups and nominal groups. We conducted the same Bayesian *t*-test for independent samples as in Experiment 1 (JZS prior with a scaling factor of $r = 1$, null interval between $d = -\infty$ and $d = 0.20$). Individual accuracy gains in the real group condition ($M = 15.47$, $SD = 17.78$) were substantially greater than in the nominal group condition ($M = -1.85$, $SD = 15.56$), $t(58) = 4.02$, $p < .001$, $BF_{10} > 100$, $d = 1.04$. This finding supports Hypothesis 1. Paired sample *t*-tests comparing Phase 1 accuracy to Phase 3 accuracy revealed that individual accuracy improved from Phase 1 to Phase 3 in the real groups, $t(29) = 4.77$, $p < .001$, $BF_{10} > 100$, $d = 0.81$, but not in the nominal groups, $t(29) = -0.65$, $p = .521$, $BF_{10} = 0.05$, $d = -0.01$, confirming Hypotheses 1a and 1b, respectively. A detailed overview of participants' individual accuracy is displayed in the upper panel of Figure 3.

Individual mapping knowledge gains. We compared participants' individual mapping knowledge gains between the two conditions (real groups, nominal groups), again using the same one-tailed Bayesian *t*-test for independent samples. Individual mapping gains in the real group condition ($M = 0.16$, $SD = 0.22$, $\rho_{Phase\ 1} = 0.75$, $\rho_{Phase\ 3} = 0.82$) were higher than in the nominal group condition ($M = 0.02$, $SD = 0.14$, $\rho_{Phase\ 1} = 0.75$, $\rho_{Phase\ 3} = 0.76$), $t(58) = 3.00$, $p = .002$, $BF_{10} = 53.16$, $d = 0.78$, supporting Hypothesis 2. A paired sample *t*-test comparing Phase 1 and Phase 3 mapping knowledge showed a strong increase in real group members' mapping knowledge, $t(58) = 4.05$, $p < .001$, $BF_{10} > 100$, $d = 0.74$, confirming Hypothesis 2a. A similar test in the nominal group condition yielded some evidence that the mapping knowledge did not differ between Phases 1 and 3, $t(29) = 0.65$, $p = .260$; $BF_{10} = 0.58$, $d = 0.12$. However, since the BF_{10} did not fall below 0.33, this test remains inconclusive, and we can neither accept nor reject Hypothesis 2b. Again, a detailed overview of participants' individual mapping knowledge is displayed in the lower panel of Figure 2.

Individual metric knowledge gains. Our first exploratory analysis of Experiment 2 concerns participants' metric knowledge. We computed participants' improvement in metric

knowledge from Phase 1 to 3. The results of a two-tailed Bayesian *t*-test with a point null hypothesis revealed that the improvement of metric knowledge was greater in the real groups (M = 14.49, SD = 22.39) than in the nominal groups (M = -1.21, SD = 17.84), $t(58) = 3.00$, $p = .004$, $BF_{10} = 9.25$, $d = 0.78$. Furthermore, separate two-tailed *t*-tests for paired samples confirmed that individuals in the real groups improved their metric knowledge, $t(29) = 3.54$, $p = .001$, $BF_{10} = 22.64$, $d = 0.68$, while those in the nominal groups did not, $t(29) = -0.37$, $p = .713$, $BF_{10} = 0.15$, $d = -0.05$.

Phase 3 individual performance. The one-tailed Bayesian *t*-test showed that participants' individual mapping knowledge in the real groups (M = 1.15, SD = 0.22, ρ = 0.82) was higher than the mapping knowledge of participants in the nominal groups (M = 1.00, SD = 0.20, ρ = 0.76), $t(58) = 2.81$, $p = .003$, $BF_{10} = 35.27$, $d = 0.73$, confirming Hypothesis 3. We also ran exploratory analyses of participants' Phase 3 accuracy and metric knowledge using two-tailed Bayesian *t*-tests with a point null hypothesis. The individual accuracy of participants in the real groups (M = 48.15, SD = 14.24) was higher than in the nominal groups (M = 68.57, SD = 17.58), $t(58) = 4.94$, $p < .001$, $BF_{10} > 100$, $d = 1.28$. Additionally, the individual metric knowledge of participants in the real groups (M = 30.76, SD = 17.2) was higher than in the nominal groups (M = 48.02, SD = 21.42), $t(58) = 3.44$, $p = .001$, $BF_{10} = 28$, $d = 0.89$.

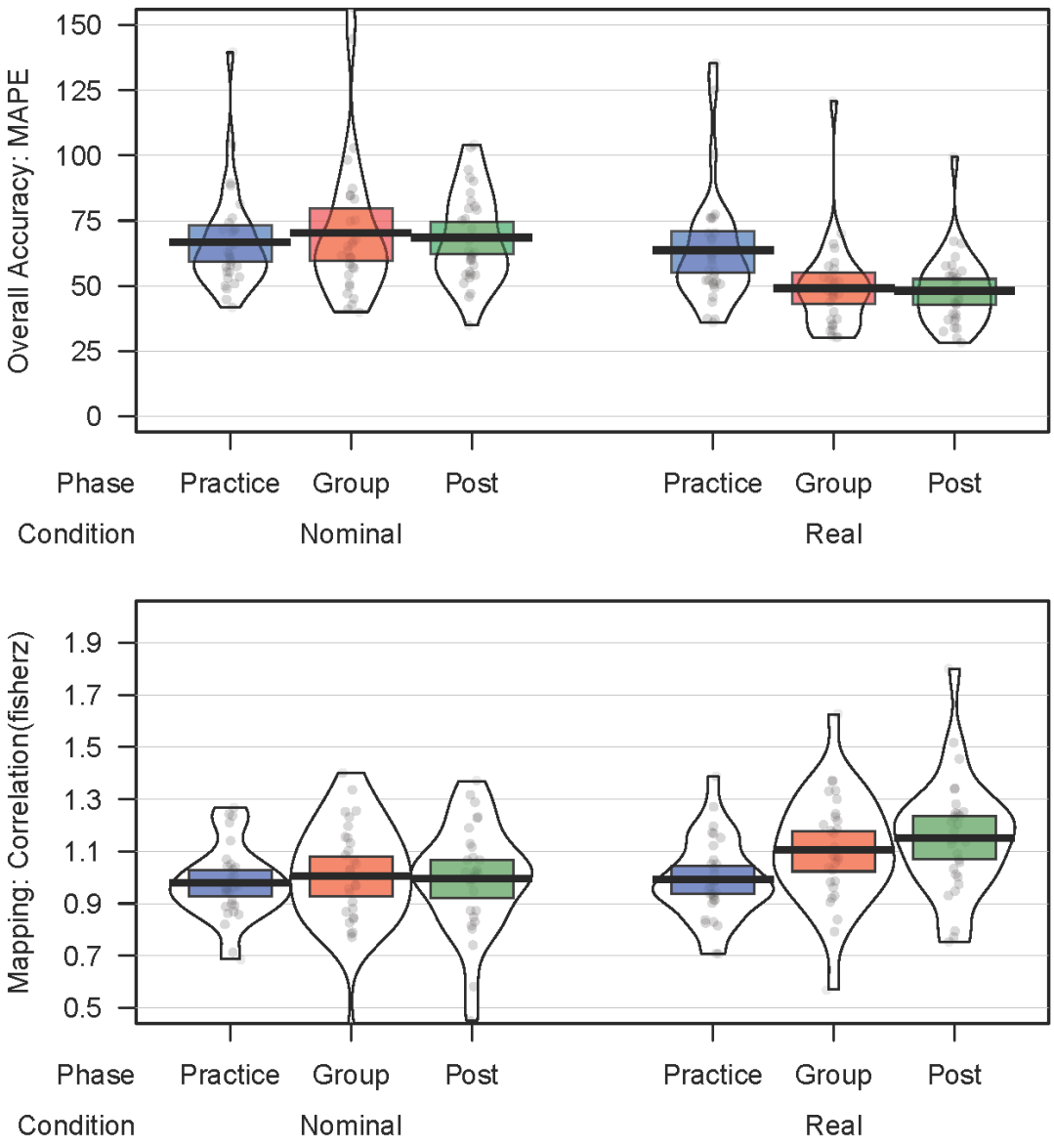

Figure 3: Individual Accuracy (upper panel) and individual mapping knowledge (lower panel) by phase and condition. In the upper panel, the overall accuracy in the form of the mean absolute percentage error is displayed. In the lower graph, mapping knowledge measured via fisher z transformed Spearman`s rank correlation is shown. Note, that the points represent the data points, the bold horizontal black lines represent the means, and the colored rectangles represent the 95% highest density intervals. The beans indicate the distribution of the data, with their width corresponding to the estimated density at the respective level of the y-axis.

Learning gains by group member ability. If the improvements in individual accuracy, mapping, and metric knowledge we observed are, in fact, the result of G-I-transfer, we would expect the initially weaker group members to benefit more than the initially stronger members. We tested this possibility in a series of exploratory analyses. First, we determined the best, medium, and worst member per group separately for overall accuracy, mapping knowledge, and metric knowledge in the practice phase. Then, we tested for differences in the magnitude of the accuracy or knowledge gains using Bayesian ANOVAs with the default settings of the Bayes factor package (Morey & Rouder, 2015), treating group member as a within-subjects factor (best vs. medium vs. worst).

The analysis of accuracy gains revealed a significant effect of group member, F (2, 58) = 16.07, $p < .001$, $\eta_g^2 = 0.27$, $BF_{10} > 100$. Post hoc comparisons showed that the worst members (M = 40.16, SD = 48.86) improved more than the best members (M = -1.14, SD = 12.41), $t(29) = 4.44$, $p < .001$, $BF_{10} > 100$, $d = 0.81$, and the medium members (M = 7.38, SD = 13.51), $t(29) = 3.67$, $p < .001$, $BF_{10} = 33.57$, $d = 0.67$. Descriptively, the medium members (M = 7.38, SD = 13.51) also improved somewhat more than the best members (M = -1.14, SD = 12.41), $t(29) = 2.43$, $p = .022$, $BF_{10} = 2.36$, $d = 0.44$, but the statistical tests yielded an inconclusive result. Consistent with this pattern, separate one-sample Bayesian t-tests against zero revealed an improvement in individual accuracy among the worst members, $t(29) = 4.50$, $p < .001$, $BF_{10} > 100$, $d = 0.82$, and the medium members, $t(29) = 2.99$, $p = .006$, $BF_{10} = 7.40$, $d = 0.55$, whereas the best members' accuracy remained largely unchanged, $t(29) = 0.50$, $p = .819$, $BF_{10} = 0.21$, $d = 0.09$. This pattern of results is displayed in the left panel of Figure 4.

Next, we analyzed the improvement in mapping knowledge. The analysis revealed a significant effect of group member, $F(2, 58) = 14.50$, $p < .001$, $\eta_g^2 = 0.18$, $BF_{10} > 100$. Post hoc comparisons revealed that the worst members' (M = 0.33, SD = 0.28) mapping knowledge improved more than that of the best members (M = 0.01, SD = 0.30), $t(29) = 5.59$,

$p < .001$, $BF_{10} > 100$, $d = 1.02$, and the medium members ($M = 0.14$, $SD = 0.27$), $t(29) = 3.00$, $p = .005$, $BF_{10} = 7.50$, $d = 0.55$. The comparison of the medium members ($M = 0.14$, $SD = 0.27$) and the best members ($M = 0.01$, $SD = 0.30$) yielded an inconclusive result, $t(29) = 1.95$, $p = .061$, $BF_{10} = 1.02$, $d = 0.36$. One-sample Bayesian *t*-tests against zero showed improvements in mapping knowledge for the worst and the medium members, $t(29) = 6.36$, $p < .001$, $BF_{10} > 100$, $d = 1.16$, and $t(29) = 2.78$, $p = .010$, $BF_{10} = 4.70$, $d = 0.51$, respectively, whereas the mapping knowledge of the best members remained, more or less, stable, $t(29) = 0.19$, $p = .854$, $BF_{10} = 0.20$, $d = 0.03$. This pattern of results is displayed in the middle panel of Figure 4.

Finally, we analyzed the improvements in metric knowledge in relation to the group members. Again, there was an effect of group member, $F(2, 58) = 22.78$, $p < .001$, $\eta_g^2 = 0.32$, $BF_{10} > 100$. Post hoc comparisons revealed that the worst members ($M = 46.99$, $SD = 56.81$) improved their individual metric knowledge more than the best members ($M = -9.97$, $SD = 17.17$), $t(29) = 5.36$, $p < .001$, $BF_{10} > 100$, $d = 0.98$, and the medium members ($M = 6.44$, $SD = 15.66$), $t(29) = 4.17$, $p < .001$, $BF_{10} > 100$, $d = 0.76$. The medium members ($M = 6.44$, $SD = 15.66$) did improve more than the best members ($M = -9.97$, $SD = 17.17$), $t(29) = 4.81$, $p < .001$, $BF_{10} > 100$, $d = 0.88$. Separate one-sample Bayesian *t*-tests against zero revealed an improvement in individual metric knowledge for the worst members, $t(29) = 4.53$, $p < .001$, $BF_{10} > 100$, $d = 0.83$. Medium members' metric knowledge improved descriptively, but the test against zero yielded an inconclusive result, $t(29) = 2.25$, $p = .032$, $BF_{10} = 1.71$, $d = 0.41$. Finally, the members' metric knowledge declined slightly, $t(29) = -3.18$, $p = .003$, $BF_{10} = 11.10$, $d = -0.58$. The results are displayed in the right panel of Figure 4.

We conclude this exploratory analysis with an interesting observation. As outlined in the paragraphs above, we did find differential performance gains separately for overall accuracy, mapping knowledge, and metric knowledge. Not surprisingly, group members' relative mapping knowledge and their relative overall accuracy prior to the group discussions

(each expressed as group members' rank within the group) were correlated, $x^2(4) = 19.80$, $p < .001$, $BF_{10} > 100$. The same was true for group members' initial relative metric knowledge and relative accuracy, $x^2(4) = 45.60$, $p < .001$, $BF_{10} > 100$. However, group members' relative mapping and metric knowledge were largely uncorrelated a priori, $x^2(4) = 4.60$, $p = .331$, $BF_{10} = 0.16$. R These results indicate that group members who had superior knowledge in mapping were often not the same participants who had superior metric knowledge. As we found for both types of knowledge that the weaker group members improved, we consider it likely that different group members can learn different things from the group discussions, which constitutes a strong potential for synergy. Since we could not provide a more direct test of this interesting possibility, we return to it in the general discussion.

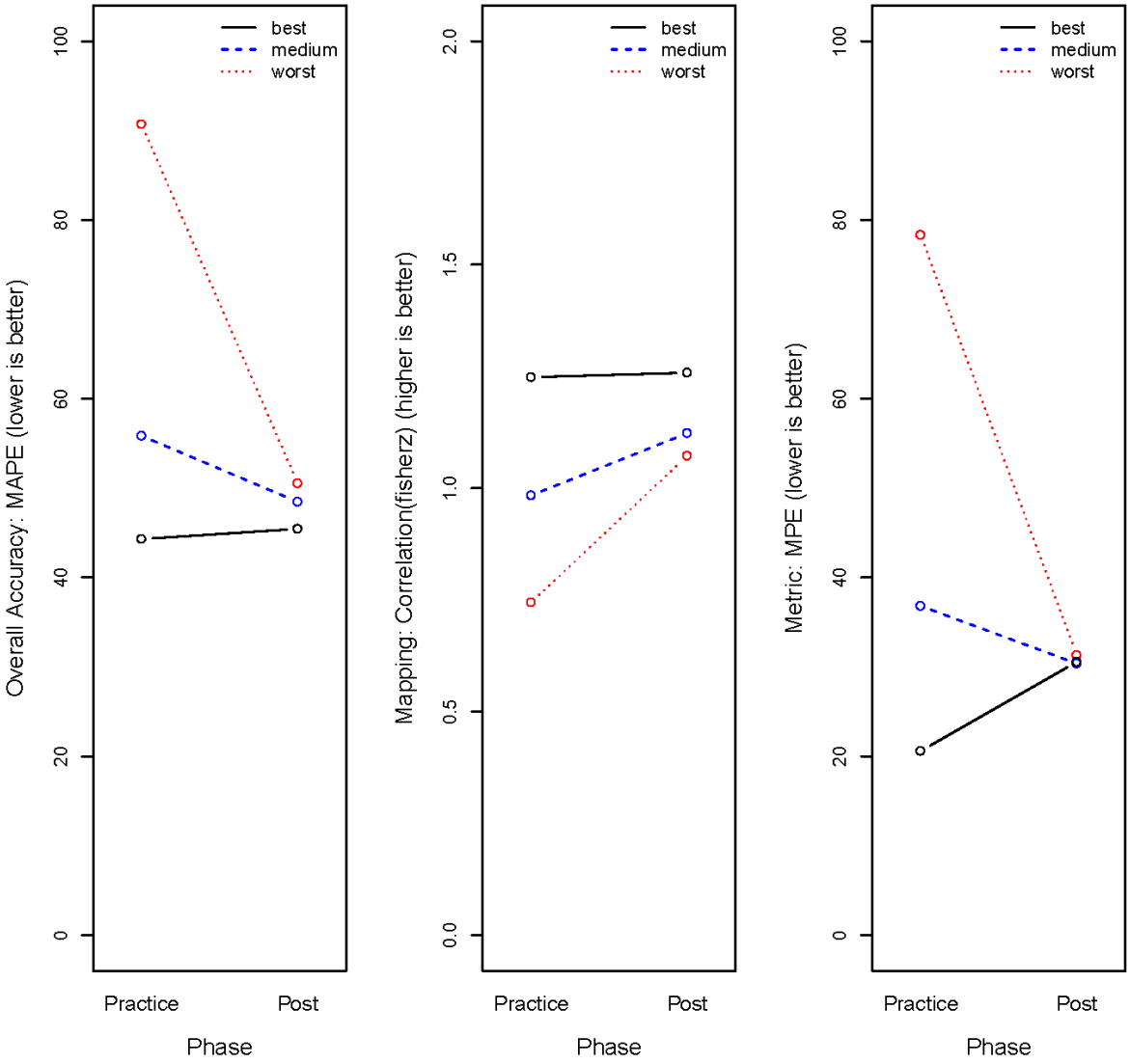

Figure 4: Individual accuracy (left panel), mapping knowledge (middle panel) and metric knowledge (right panel) by phase and group member. The points represent the average performance of the specific type of group members.

Group performance. So far, our analyses show that real group members benefitted from substantial G-I transfer in terms of increased overall individual accuracy, and that this transfer was due to both increased mapping and increased metric knowledge. Given that G-I transfer is one of the sources of synergy in quantitative group judgment, we would expect real groups to outperform nominal groups in Phase 2. To test whether this was the case, we compared the accuracy of the real group judgments with the average of the nominal group

members' individual judgments during this second phase. Remember that the nominal group performance benefits from statistical aggregation but lacks potential benefits of social interaction and interdependence. A two-tailed independent samples Bayesian *t*-test with a point null hypothesis revealed that the group judgments (M = 43.92, SD = 21.92) were much more accurate than the averaged judgments in the nominal groups (M = 70.81, SD = 38.76), $t(45.83) = 3.31$, $p = .002$, $BF_{10} = 19.82$, $d = 0.85$. This result indicates substantial synergy due to interaction in real groups.

Furthermore, we compared the group judgments to the judgments of the nominal groups' best members. To determine nominal groups' best members, we first computed the MAPE scores of nominal group members' second estimates in the group phase and then chose the person with the lowest MAPE. A two-tailed independent samples Bayesian *t*-test with a point null hypothesis revealed that the group judgments (M = 43.92, SD = 21.92) were as accurate as the judgments of the nominal groups' best members (M = 45.65, SD = 10.43), $t(41.49) = 0.39$, $p = .700$, $BF_{10} = 0.21$, $d = 0.10$.

Group mapping knowledge. Similar to the main analysis of the G-I transfer, we compared the mapping knowledge of real groups to that of the nominal groups. To this end, we computed nominal group's mapping knowledge based on the average of the three nominal group members' judgments. A two-tailed independent samples Bayesian *t*-test with a point null hypothesis revealed that real group judgments (M = 1.44, SD = 0.29, $\rho = 0.89$) reflected higher mapping knowledge than nominal group judgments (M = 1.04, SD = 0.45, $\rho = 0.78$), $t(49.43) = 4.05$, $p < .001$, $BF_{10} > 100$, $d = 1.05$.

We also compared real group's mapping knowledge to that of the nominal groups' most knowledgeable members in terms of mapping knowledge. To determine the most knowledgeable member, we computed nominal group members' mapping knowledge based on their second estimates in the group phase and chose the person with the highest rank correlation per group. A two-tailed independent samples *t*-test with a point null hypothesis

revealed that the real groups' mapping knowledge (M = 1.44, SD = 0.29, ρ = 0.89) surpassed that of the nominal groups members with the highest mapping knowledge (M = 1.24, SD = 0.26, ρ = 0.84), $t(58) = 2.79$, $p = .007$, $BF_{10} = 5.63$, $d = 0.72$.

Group metric knowledge. For the sake of completeness, we also compared the metric knowledge of the real groups to that of the nominal groups in Phase 2. Again, we computed the latter based on the average of nominal group members' individual judgments. A two-tailed independent samples Bayesian *t*-test with a point null hypothesis revealed that the metric errors of group judgments (M = 31.04, SD = 24.09) were somewhat lower than the metric errors of nominal group judgments (M = 44.65. SD = 41.39), but this difference was rather small, and the respective *t*-test test yielded an inconclusive result, $t(46.62) = 1.56$, $p = .126$, $BF_{10} = 0.58$, $d = 0.40$.

We also compared real groups' metric knowledge to that of the nominal groups' most knowledgeable members in terms of metric knowledge. The nominal groups' most knowledgeable member in terms of metric was the member with the lowest metric errors across the second individual judgments of Phase 2. Descriptively, the metric errors of real groups (M = 31.04, SD = 24.09) were higher than the metric errors of the nominal groups' most knowledgeable members (M = 20.75, SD = 14.96), but a two-tailed independent samples Bayesian *t*-test with a point null hypothesis yielded an inconclusive result, $t(48.47) = 1.99$, $p = .052$, $BF_{10} = 1.12$, $d = 0.51$. .

Differential weighting. Our final analysis concerns the possibility that real groups achieved synergy (also) by weighting their more expert members or more accurate individual judgments more strongly when forming the consensus estimates (differential weighting). To test for this possibility, we compared the accuracy of real group judgments to the average of the real group members' corresponding individual pre-discussion judgments (i.e., the average model). We excluded the first trial of the group phase from this comparison because, whereas the first group judgment may already benefit from G-I transfer, the individual pre-discussion

judgments for the first trial of the group phase cannot (they were provided prior to any group interaction). Thus, retaining the first trial of the group phase in the analysis might artificially create the impression that groups engaged in differential weighting (see Schultze et al., 2012). A two-tailed paired sample *t*-test with a point null hypothesis revealed that the overall accuracy of the group judgments (M = 44.23, SD = 22.67) was almost identical to that of the average model (M = 44.15, SD = 19.10), $t(29) = 0.07$, $p = .947$, $BF_{10} = 0.14$, $d = 0.00$. Hence, there was no evidence of differential weighting.

Discussion

The results of Experiment 2 confirmed all of our hypotheses. First, we again replicated the generalizability of G-I transfer to multi-cue judgment tasks. Second, and most importantly, we found very strong evidence for G-I transfer of mapping knowledge: Whereas nominal group members' mapping knowledge remained largely unchanged, real group members benefited from group interaction and increased their individual mapping knowledge substantially. Consequentially, the individual mapping knowledge of former real group members exceeded that of former nominal group members in Phase 3. Since we made sure – via parallelization – that there were no initial differences in mapping knowledge or general accuracy between conditions, the only viable explanation for the observed results is G-I transfer. Furthermore, group members were able to improve their metric knowledge via group interaction, thereby replicating previous findings (Stern et al., 2017). In an exploratory analysis, we further replicated synergy through real group interaction, that is, real groups outperformed nominal groups. Since there was no indication at all in our exploratory analyses that group judgments benefitted from differential weighting by expertise or accuracy, we attribute this synergy fully to G-I transfer. Additional exploratory analyses showed that G-I transfer differed in relation to the member's initial performance. In particular, initially weaker group members realized the strongest G-I transfers.

General Discussion

The main goal of the present research was to test whether working on quantitative judgment tasks in interacting groups leads to increases in group members' individual mapping knowledge. We hypothesized that real group members' individual mapping knowledge would increase due to group discussion, whereas the mapping knowledge of nominal group members should remain constant. As a prerequisite for the transfer of mapping knowledge in real groups, we also investigated whether the individual judgment accuracy of real group members benefited from G-I transfer.

In line with our assumptions, the results of our two experiments showed that participants who worked in real (i.e., interacting) groups improved their individual judgment accuracy due to group discussions, while participants who worked alone (i.e., as members of nominal groups) did not. With regard to the transfer of mapping knowledge, we observed a similar pattern. In the two experiments, group members improved their individual mapping knowledge as a consequence of group discussion, while the individual mapping knowledge of participants working alone did not improve over time. Because of baseline differences between the real and nominal groups in terms of mapping knowledge, the results of Experiment 1 were not entirely conclusive regarding the transfer of mapping knowledge. However, the methodologically improved Experiment 2 showed strong evidence of the predicted transfer of mapping knowledge. Furthermore, in Experiment 2, group judgments were more accurate than the average of a comparable number of individual judgments in the nominal condition. This synergy effect could be fully attributed to G-I transfer, as we did not detect any evidence of differential weighting.

To the best of our knowledge, our experiments provide the first conclusive evidence for G-I transfer of mapping knowledge: group members can increase their individual mapping knowledge due to group interaction. In other words, group discussion enables group members to gain an improved understanding of the relationship between cue characteristics on the one hand, and target values on the other hand. In line with previous findings (Stern et al., 2017),

group members were also able to reduce their metric errors via group interaction, and weaker group members benefited the most from the group discussions.

Our results further indicate that previous findings of G-I transfer also generalize to multi-cue judgments tasks. Previous research detected G-I transfer in collective induction (Brodbeck & Greitemeyer, 2000), reasoning tasks (Laughlin, Carey, & Kerr, 2008), mathematical problems (Laughlin & Ellis, 1986), and simple judgment tasks pertaining to general knowledge (Schultze et al. 2012, Stern et al. 2017). Thus, by showing that G-I transfer also occurs in multi-cue judgment, our findings highlight the relevance of this learning process for group performance. Nevertheless, we would like to point out an important difference between multi-cue judgments and simple general world knowledge judgments that highlight the relevance of our current findings: Multi-cue judgments possess a great deal of practical relevance. In most real-world judgment tasks, such as climate change predictions or financial prognoses, the forecasters have knowledge of all relevant cues, but must (learn to) use these cues correctly to make the most accurate prediction. Experiments 1 and 2 showed that group members improved their individual accuracy and mapping knowledge under these exact conditions. In other words, our findings suggest that G-I transfer in quantitative judgment is not restricted to simple general knowledge questions (where, in practice, one would hardly need groups to solve them), but can also occur under more realistic task settings.

The results of Experiment 2 also showed synergy through real group interaction, that is, real groups outperformed nominal groups in terms of overall accuracy. This adds to previous research showing synergy in group judgment (e.g., Bonner & Baumann, 2012; Minson et al., 2017; Schultze et al., 2012; Sniezek & Henry, 1989). In the present study, this synergy can be fully explained by G-I transfer, as we did not find any evidence of differential weighting. We can even go one step further and attribute synergy to G-I transfer of both metric and mapping knowledge. Finally, the results of Experiment 2 indicate that groups

outperform even the most knowledgeable members of the nominal groups with regard to mapping.

Practical implications

Practitioners often need to decide whether they should rely on groups to perform a particular task. In many cases, group performance is considered superior to individual performance. However, research on group performance shows that groups often fall short on these expectations and do not exceed reasonable baselines based on the combination of individual contributions (e.g., Steiner ,1972; Kerr & Tindale, 2004). Even under conditions that might seem optimal for developing synergy, groups often fail to realize this potential (Schulz-Hardt & Mojzisch, 2012). One of the few exceptions where the implementation of groups seems to be highly beneficial are quantitative judgments (e.g., Schultze et al., 2012). Based on our finding that interacting groups substantially outperformed the average of a comparable number of individuals, we can now make an even stronger case for having groups work on quantitative judgments.

Discussing and deciding on quantitative judgments in a group also increases group members' individual ability to perform the task. On the one hand, this implies that organizations may use group work as a means to train forecasters effectively. On the other hand, combining group and individual work in an appropriate manner could help to exploit the benefits of group interactions without having to bear its full costs. One could first have forecasters work in interacting groups to reap the benefits of G-I transfer. Once G-I transfer is complete, one could then disband the groups, have former group members work on the task individually, and then combine their estimates by averaging them (similar to a nominal group). Because we did not find evidence of differential weighting, we can assume that, over time, the average of individual judgments of former group member might become as accurate as the group judgments themselves, and that additional group work is not necessary from this point on.

So far, our focus was on judgmental accuracy. However, accuracy is not always the sole determinant of judgment quality. Specifically, there may be situations in which the mapping of judgments is of even greater importance than the overall accuracy. An example for such a situation is the selection of employees: Usually, the suitability of candidates for a certain position is rated based on various cues such as test scores, previous education, past performance, and letters of recommendations. As organizations want to hire the person who is best suited for the job, the goal of personnel selection is to sort all candidates by suitability (i.e., the best candidate receives the highest score, the second-best candidate receives the second highest score, etc.). In this situation, even large metric errors might be inconsequential, that is, a hiring committee might underestimate all candidates or judge all of them too harshly, but – despite lacking in overall accuracy – still end up with the correct order of applicants. As an added benefit, the mapping of judgments (but not necessarily their calibration in terms of metric) ensures the fairness of the procedure (e.g., Conway, Jako, & Goodman, 1995). Our data suggest that groups may be particularly useful in such contexts. Finding that groups' mapping knowledge exceeded even that of the best members of nominal groups implies that their judgments are not only accurate but also highly consistent**.**

Limitations and directions for future research

There are some limitations worth addressing. The first set of limitations concerns the task and the experimental design we used. Since we applied the same multi-cue judgment task in both experiments, we cannot rule out that our results are somewhat task-specific and do not necessarily generalize to other multi-cue judgment tasks. Additionally, our experimental procedure did not include any form of feedback. However, in many real-world tasks such feedback is available, either immediately or with delay. For example, in forecasting tasks the true value will be known eventually. It is conceivable that feedback in multi-cue judgement tasks might already be sufficient to elevate individual decision-makers' mapping knowledge to a level where there is little room for additional increases due to G-I transfer. If so, it might

still be possible that groups are better at interpreting feedback, thus enabling to learn target-cue relations faster. Accordingly, G-I transfer might not manifest in greater individual mapping knowledge but, rather, in reduced time to obtain it. Therefore, future research is needed to test the generalizability of our findings, especially in multi-cue judgement tasks with (time-delayed) feedback.

The second set of limitations is concerned with the specific conditions relevant to G-I transfer of mapping knowledge. Hence, it remains to be clarified under what conditions mapping knowledge can be transferred between group members. In our experiments, we created ideal conditions for this type of transfer. In particular, we choose a task where mapping can be learned, and we implemented heterogeneous groups to facilitate learning. Future research should investigate the influence of specific task features, as well as the effect of group composition on the transfer of mapping knowledge, because they might yield important insights into the psychological mechanisms involved in G-I transfer. For example, the general individual mapping knowledge prior to group discussion was already relatively high in our task. This may have created a situation with high task demonstrability and sophisticated discussion of the cue-target relations, which might be a prerequisite for G-I transfer of mapping knowledge. Therefore, it is important to investigate the extent of general individual mapping knowledge necessary for having an improvement in mapping knowledge. Likewise, we assigned participants to the groups so that groups were heterogeneous with regard to their mapping knowledge. Future research could investigate whether there is an optimal ability difference between the best and weakest member of a group for the improvement of the individual mapping knowledge.

A final point we consider worth noting here is that group members with the best mapping knowledge were not necessarily also the ones with the best metric knowledge. In fact, one of our exploratory analyses found evidence for the null hypothesis that group members' relative expertise in metric and mapping knowledge was unrelated. The fact that we

still found substantial G-I transfer in both knowledge domains suggests that groups were rather good at leveraging their members' specific strengths by integrating their unique knowledge. Unfortunately, our data did not allow for more detailed tests of this exciting possibility. Therefore, future research should investigate how well, or under which circumstances, groups can realize their potential for synergy in when the relevant metric and mapping knowledge is distributed among group members.

Conclusion

Groups working on multi-cue judgment tasks benefit from G-I transfer, that is, they can outperform a comparable number of individuals because their members' individual ability to perform the task increases due to group discussion. The current study shows that this increase in group members' accuracy has two components. On the one hand, group members can correct their erroneous metric knowledge, that is, they develop a better understanding of what constitutes a possible range of target values. On the other hand, they also gain a deeper understanding of how different target values relate to each other. Our study highlights the importance of studying group performance as a dynamic construct that is subject to group learning.

References

Baumann, M. R., & Bonner, B. L. (2004). The effects of variability and expectations on utilization of member expertise and group performance. *Organizational Behavior and Human Decision Processes, 93*, 89-101.

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making*, *10*(3), 265-276.

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology*, *102*(2), 337.

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.

Brodbeck, F. C., & Greitemeyer, T. (2000). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes and Intergroup Relations, 3,* 159-182.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511.

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In *Psychology of learning and motivation* (Vol. 41, pp. 321-359). Academic Press.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*(1), 149.

Greiner, B. (2015). Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE, *Journal of the Economic Science Association 1*(1), 114-125.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol., 55*, 623-655.

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11*, 319-330.

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

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.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. *R package version 0.9, 9*, 2014.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an

alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, *118*, 24-36

Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology*, *23*, 305-343.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*, 1–28.

Steiner, I. D. (1972). Group Process and Productivity Academic Press. *New York*.

Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). How Much Group is Necessary? Group-To-Individual Transfer in Estimation Tasks. *Collabra: Psychology*, *3*(1).

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, *296*.

Treffenstaedt, C., & Wiemann, P. (2018). Alfred - A library for rapid experiment development (Version 0.2b5). https://doi.org/10.5281/zenodo.1437220

Appendix B.

Manuscript 2

Lippold, M., Schultze, T., & Schulz-Hardt, S. (2021). *The benefit of collaboration: Disentangling the sources of synergy in group judgments.*

Unpublished Manuscript.

The benefit of collaboration: Disentangling the sources of synergy in group judgment

Matthias Lippold[1,2], Thomas Schultze[1,2], & Stefan Schulz-Hardt[1,2]

[1] University of Goettingen

[2] Leibniz ScienceCampus Primate Cognition, Göttingen

Author Note

This research was supported by grants of the Deutsche Forschungsgemeinschaft (DFG) to the second and last authors (SCHU 1279/11-2; SCHU 2813/2-2).

Correspondence concerning this article should be addressed to Matthias Lippold, Gosslarstr. 14, 37073 Goettingen. E-mail: matthias.lippold@uni-goettingen.de

Abstract

Whereas recent research on the *wisdom of the crowd effect* questioned the importance of group interaction by showing that, in many cases, simple aggregations of independent judgments are already highly accurate, we present extensive meta-analytic evidence that interacting groups continuously outperform such aggregations with regard to accuracy in multiple consecutive judgments and, thereby, achieve synergy. In particular, we show that group judgments are more accurate than the average of a comparable number of individual judgments, and that group judgments can be as accurate as the judgments of the best individual member of the nominal groups (i.e., the person with the lowest judgment error). Additionally, our meta-analytic results substantiate two different mechanisms that lead to synergy in quantitative group judgment. On the one hand, group members can learn from each other and, thereby, increase their individual accuracy (G-I transfer). On the other hand, when controlling for G-I transfer, group members are also able to identify the more accurate judgments and assign more weight to them (differential weighting).

Keywords: group judgment, group processes, wisdom of the crowd

The benefit of collaboration: Disentangling the sources of synergy in group judgment

Quantitative judgments such as forecasts and prognoses form the basis for many important political, entrepreneurial, and medical decisions. The quality of those decisions crucially rests on the accuracy of the underlying judgments. For example, medical teams will usually only decide in favor of a particular medical intervention if the forecasted chances of success are high. Accordingly, the accuracy of quantitative judgments has been of broad scientific interest to diverse disciplines, including social psychology, cognitive science, economics, and management science. Often, decision-makers appoint groups rather than individuals to make important quantitative judgments (Gigone & Hastie, 1997), presumably because groups have greater cognitive capacities, possess more information, and can integrate their members' unique knowledge in order to achieve higher accuracy. Hence, a fundamental research question in group judgment is whether groups are, indeed, more accurate judges than individuals (e.g., Gigone & Hastie, 1997, p. 149). Even in the absence of beneficial group processes, group judgments will (at least on average) be more accurate than individual judgments, because averaging a number of independent judgments usually leads to increases in accuracy. This happens due to statistical error cancellation (e.g., Herzog & Hertwig, 2009; Soll & Larrick, 2009) and is (among others) labeled as the "wisdom of the crowds" effect (Surowiecki, 2005). It has been known for a long time that these kinds of accuracy gains do not represent a real group process (Stroop, 1932). Therefore, we can only speak of a benefit of collaboration or *synergy* (Larson, 2010) if group judgments are not just more accurate than individual judgments, but if groups are rather able to make more accurate judgments than the *aggregation* of individual judgments. Earlier reviews on group judgments were somewhat skeptical about the existence of such synergy in group judgments. For example, Hastie (1986) stated that the advantage of the group over the average of individual judgments is only small. Gigone and Hastie (1997) summarized that, according to previous studies, group judgments were just as accurate as the average of individual judgments. However, they also criticized that previous studies did rely on different baselines for analyzing group judgment accuracy and, therefore, a meaningful meta-analytic test was not conducted in their review. In the meantime, a number of studies has provided solid evidence that group judgments can, indeed, be more

accurate than the average of a comparable number of individual judgments (e.g., Keck & Tang, 2018; Lippold, Schulz-Hardt, & Schultze, 2020; Minson, Mueller, & Larrick, 2017; Sniezek & Henry, 1989, Schultze, Mojzisch, & Schulz-Hardt, 2012; Stern, Schultze, & Schulz-Hardt, 2017). In the present research, we will conduct several meta-analyses to determine the extent of the benefit of group judgments over a statistical aggregation of individual judgments. Furthermore, a central aim of our paper is to disentangle different components of these synergetic gains, in order to learn more about their nature, and to identify the mechanisms that lead to these gains. We will address this in two ways: First, we will investigate what type of judgment errors become smaller in groups compared to statistical individual aggregates, in order to learn what groups are particularly good at. Second, we will investigate the mechanisms that help groups achieve these gains: Are these gains due to an improvement of group members' individual capabilities, or are they the result of beneficial coordination in the group, or both? Overall, we aim at presenting a comprehensive analysis of synergy in group judgment.

Synergy in group judgment

As already outlined, even in the absence of any group processes, statistical aggregation will benefit judgment accuracy. As a consequence, comparing the accuracy of a group with the accuracy of a single individual working alone is only of limited value when it comes to evaluating the performance of groups. Specifically, if we want to know whether group processes benefit judgment accuracy - that is, whether groups achieve synergy - we have to compare group judgment accuracy with the accuracy that is achieved without group processes, but in the presence of statistical aggregation. In other words, we have to compare the judgments of N-person-groups with the average judgments of N independent individuals or, as this comparison is often framed, we have to compare groups with the average model. Calculating the average constitutes a baseline for group judgment accuracy that is both theoretically and practically meaningful. Theoretically, as we have outlined, forming the average of independent judgments leads to statistical error cancellation (e.g., Soll & Larrick, 2009). Hence, whenever there is unsystematic error in members' individual judgments, the average model is more accurate than the average individual judgment, and this advantage increases with increasing

group size. Such statistical error cancellation can be powerful, as the literature on the wisdom of the crowds indicates (Surowiecki, 2005). Therefore, beating the average model (i.e., achieving synergy) is quite an achievement. In line with these theoretical considerations, the average model has often been used as the baseline for the detection of synergy in judgment tasks (e.g., Minson et al., 2017; Schultze et al., 2012; Stern et al., 2017). Apart from this, the average model can also be practically meaningful. For example, in cases where groups are unable to use the individual knowledge of the members, or are unable to reach a consensus through communication, the average of all members' judgments constitutes a reasonable alternative for achieving a group result. On top of that, decision makers might decide to save the effort of bringing people together in a group and have them discuss, but to exploit the statistical advantage of error cancellation by eliciting several individual judgments and averaging them afterwards. Therefore, the average model can be a practical alternative to the group work itself. Whereas we will consider the average model as the baseline for the detection of synergy in group judgment, we will also take an additional benchmark into account which can help to determine and better understand the accuracy of group judgments. This additional benchmark is obtained by identifying the best member of the group, or the best individual member of a group of individuals working alone, often termed the best member model. If the group can outperform the best member model, we can assume that no member can be better than the group as a whole. Therefore, the best member model constitutes an important benchmark for the group performance. In research on quantitative judgment tasks, groups have often been compared with the best member models, but the corresponding analyses yielded mixed results. While Gigone and Hastie (1997) stated in their review that group judgments were, on average, less accurate than the best members, more recent studies indicate that groups might be able to perform on a similar level as the best members in quantitative judgments (e.g. Bonner & Baumann, 2008; Laughlin, Gonzalez & Sommer, 2003). Performance wise, the best member model does not have an advantage over the average model, or vice versa. Rather, depending on the environment, the best member's estimate might be more or less accurate than the average model (e.g., Einhorn, Hogarth, & Klempner, 1977). There are, however, important theoretical and practical differences between these models: In theoretical terms, the average model can be seen as the

group judgment that would be obtained in the absence of group processes (Steiner, 1966). The same cannot be said for the best member model, because identifying the best member does, in itself, require a group process. Furthermore, the best member model is hardly a model that can be practically applied (in terms of having the best member rather than the group perform the task), because identifying the best member is only possible after all judgments have been made (and feedback has been obtained). There is, however, a variant of the best member model that could, under certain conditions, also be practically applied: Whereas the best member model is usually calculated on the basis of the performance of the group members during the group phase (which, of course, requires the measurement of individual performance during the group phase), having an individual phase before the group gets together allows for a measurement of initial group member performance. Based on these measurements, an initial best member can be determined, and the performance of this initial best member during the (later) group phase can also constitute a baseline against which the actual group accuracy can be compared. We will term this the initial best member model. Similar to the average model, the initial best member model would constitute a practically meaningful baseline. For example, a company either can delegate a task to the person considered to be best for the job or rely on a group including this member. However, the best member prior to the group interaction does not necessarily have to be the best member during the group interaction. In particular, the performance of the presumable best member will most likely decline within further trials, due to regression to the mean (e.g., Hertel, Kerr, & Messe, 2000). In addition, the individual group members might benefit to different degrees from the group interaction, which might change the performance-related rank order of the individual group members. In spite of these limitations of the best member model and the initial best member model, we consider both to be important additional benchmarks on top of the average model, helping to better evaluate the actual accuracy of groups in judgment tasks. Therefore, we will include both of these models into our meta-analyses, that is, we will compare group accuracy not only with the average model, but also with both best member models. We have now introduced three possible ways to create benchmarks for real group performance in quantitative judgment tasks, namely one primary benchmark for synergy (the average model) and two additional benchmarks that compare

groups with their best members. By regarding all of them, we hope to obtain a rather complete picture of how groups perform and, hence, to successfully address our first research question:

Research Question 1: Do groups systematically achieve synergy in quantitative judgments tasks, that is, do they systematically beat the average model? If so, how large is this synergetic advantage of group work? And how accurate are groups compared to the two best member models outlined above?

By answering this (threefold) question, we will know whether and, if so, to what extent groups manage to outperform different aggregates of individual judgments. However, at this point we will not yet know why they do so. To gain insights with regard to this latter question, we will, on the one hand, disentangle two fundamental types of errors that groups and individuals commit. On the other hand, we will differentiate between two different basic mechanisms that can lead to increased accuracy in groups.

Disentangling Error Components

It is a common approach to disentangle judgment error to further understand group judgments (e.g., Gigone & Hastie, 1997). In the current research, we will rely on the taxonomy by Brown and Siegler (1993). By differentiating between the two dimensions of metric errors, on the one hand, and mapping errors, on the other hand, the Brown and Siegler taxonomy can help us to better understand the errors that groups and individuals make and, hence, the types of improvements that groups can achieve over individuals. Metric knowledge is concerned with the distributional attributes of the criterion, such as the mean, median, and plausible range. Hence, metric errors are made when a judge is incorrectly calibrated on the criterion, leading to a systematic over- or underestimation of entities from the same category. In other words, metric errors relate to the bias of the judge (Einhorn, Hogarth, & Klempner, 1977). For example, one judge might overestimate distances between European cities as she relies on a false reference frame to make her judgments. Her frame of reference might be the distance between the southern and the northern border of Germany. She might erroneously consider this distance to be 5000 kilometers. When she then has to estimate the distance between Copenhagen and

Vienna, she might know that Copenhagen is located north of Germany, and that Vienna is located south of Germany. Therefore, she considers the distance between from Copenhagen to Vienna to be more than 5000 kilometers. Accordingly, she would overestimate many distances between major European cities. The metric error can be measured by calculating the mean error based on all judgments, thereby allowing over- and underestimations to cancel out each other (Stern et al., 2017). Mapping knowledge is concerned with the relation between targets in a domain. In other words, mapping relates to the order of targets. Mapping errors are made if the relative magnitude of the target is determined incorrectly. For example, a judge might erroneously consider the distance between Hamburg and Berlin to be longer than the distance between Copenhagen and Vienna. Mapping can be measured using Spearman's rank correlation between the estimated and the true values. This correlation index indicates to which extent the order of the judgments corresponds with the order of the underlying true values. Differentiating between mapping and metric errors gives us insights into the group judgment processes, particularly with regard to what exactly groups are better at when being compared with the different individual baselines. Previous research on group judgment in relation to metric and mapping is sparse. To the best of our knowledge, only one study did compare group judgments and individuals' baselines with regard to mapping (Lippold et al., 2020). This study showed that groups made smaller mapping errors than the average and best member models. With regard to the metric error component, Lippold and colleagues could not find evidence for an advantage of the group. Gigone and Hastie (1997) did also report no difference between the average model and actual groups with regard to metric errors . However, some research indicates that groups discuss a smaller range of possible values and commit fewer extreme errors as individuals do (Minson et al. 2017), which indicates that the groups might have better metric than individuals. As both judgment error dimensions are important to understand the accuracy of group judgments, we will compare group judgments and the individual benchmarks with regard to both of these error dimensions.

Research Question 2: Do groups systematically achieve synergy with regard to metric errors? And how do their metric errors compare with their best members' metric errors?

Research Question 3: Do groups systematically achieve synergy with regard to the

mapping? And how do their mapping errors compare with their best members' mapping errors?

Disentangling group mechanisms

Two mechanisms have been postulated to explain synergy in group judgments: G-I transfer (e.g., Schultze et al., 2012; Stern et al., 2017) and differential weighting (e.g., Bonner et al., 2007; Sniezek & Henry, 1989). G-I transfer refers to group members' individually improving their task performance as a consequence of prior group interaction (Brodbeck & Greitemeyer, 2000). In group judgments, individual group members can increase their individual accuracy due to previous participation in a group, where task-relevant information has been exchanged between the group members (Schultze et al., 2012; Stern et al., 2017). For instance, the most knowledgeable group member could demonstrate her thought processes to the weaker group members and, thereby, highlight relevant underlying principles. G-I transfer has been shown to occur quite frequently in cognitive tasks, such as collective induction (Brodbeck & Greitemeyer, 2000), or mathematical problems (Laughlin & Ellis, 1986). In multiple judgment tasks, the members' individual performance has a direct influence on group performance (Sniezek & Henry, 1990). Therefore, G-I transfer plays a crucial role in quantitative judgment tasks. In fact, several studies already found evidence for G-I transfer in quantitative judgment tasks (Lippold et al., 2020; Schultze et al., 2012; Stern et al., 2017). Note that group members can, on the one hand, increase their judgment accuracy by adjusting their individual metric and, thereby, reducing their individual metric errors (Stern et al., 2017). One the other hand, group members can also learn specific cue-target relations during group discussion and, thereby, become able to reduce their individual mapping errors (Lippold et al., 2020). Differential weighting leaves group members' individual task-related capabilities unaffected but, instead, takes place at the level of how the group coordinates its members' inputs. Synergy due to differential weighting occurs if group members give more weight to more accurate judgments when forming their consensus group judgment. For example, groups might be able to identify more capable members and weight their input stronger. On the other hand, groups might be able to identify the most accurate individual judgment for a specific judgment task through group discussion and weight these judgments stronger. Note that both

phenomena, G-I transfer and differential weighting, are not mutually exclusive. Rather, both processes play a role in group judgment accuracy. Despite that G-I transfer and differential weighting can be easily distinguished on the conceptual level, disentangling them at the empirical level is more challenging. The majority of (earlier) group judgment studies did use the so-called I-G design, which does not allow for the differentiation between the two mechanisms. In the I-G design, all group members provide all of their individual judgments blockwise before the first group discussion takes place. In a subsequent second block, the group members meet and work on the same tasks as they have done before individually. If, using this design, one finds that the accuracy of the group judgments is higher than the members' averaged individual judgments from the first phase, the superiority of the group cannot not be fully attributed to either differential weighting or G-I transfer. The reason for this is that no individual judgments are elicited during the group phase, meaning that the researcher does not know whether group members' had already improved their capabilities before the second, third, fourth etc. group judgment are made. Because differentiating between the mechanisms that (can) lead to synergy in group judgment lies at the heart of our paper, we will restrict our analysis to studies using the so-called alternating-individual-group design (aI-G design; Schultze et al., 2012). This design, illustrated in Figure 1, consists of, at least, two phases: an individual practice phase and a subsequent group phase. In the practice phase, all participants work alone. In the group phase, individual and group judgments alternate. This means that for each task, the group judgments directly follow the group members' individual judgments. Hence, the individual contribution of each member can be identified during this process, allowing for an assessment of the individual judgment accuracy within each trial and, hence, for changes in individual accuracy as a consequence of previous group interaction. Therefore, this alternating design allows us to differentiate between the G-I transfer and differential weighting.

Using the al-G design, previous studies found evidence that the group member benefit from G-I transfer in judgment task, but at the same time groups often did not benefit from differential weighing (Lippold et al., 2020; Schultze et al., 2012). In particular, when controlling for G-I transfer, most of the time group judgments were not more accurate than the average of the initial group members' judgments right before the corresponding group

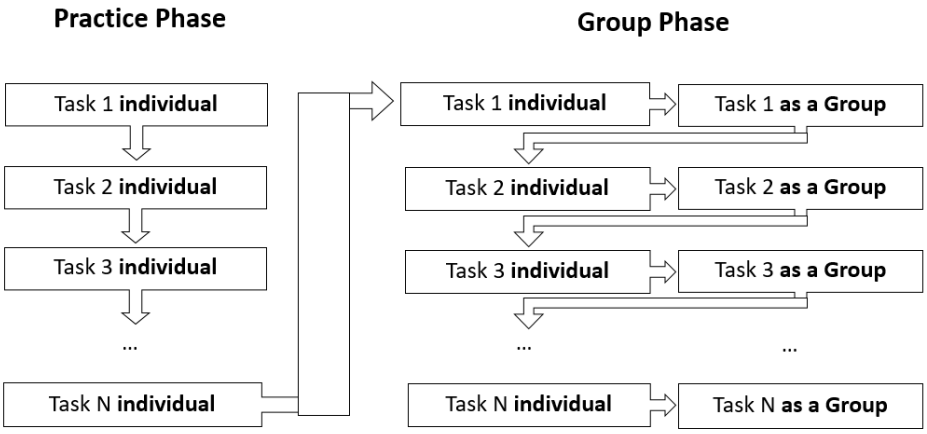

Figure 1. Display of the aI-G design.

discussions. Only under favoring circumstances did some differential weighting occur (Stern et al., 2017). Particularly, differential weighting was detected in tasks where the participants systematically overestimated the weight of objects. In other words, in these tasks the participants had high metric errors, resulting in an overall population bias. In contrast, if the participants had no systematic errors (but rather only random errors), no evidence for differential weighting was found. Hence, in comparison with differential weighting, G-I transfer seems to be the stronger process so far, but this evidence is based on a few studies and, hence, far from being conclusive. Against this background, we will investigate the extent of G-I transfer and differential weighting in judgment tasks:

Research Question 4: To what extent do group members benefit from G-I transfer in quantitative judgment tasks?

Research Question 5: To what extent do groups benefit from differential weighting in quantitative judgment tasks?

Methods

To answer our research questions, we reanalyzed data of altogehter 11 experiments with 855 groups and 2565 participants in total. All of these experiments have been conducted in our own lab, because we know of no other studies on group judgment accuracy that have used the aI-G design (or an alternative design allowing for the identification of G-I transfer

vs. differential weighting, so far.) The experiments are highly similar in nature. The group sizes of the included experiments are identical for all real interacting and nominal groups, containing three members each. Note that three experiments did not include a nominal condition and that we dropped extra conditions which were also implemented in some of the experiments. Further details about each study are provided in Table 1. Based on these experiments, we will conduct a comprehensive analysis of synergy in group judgment. First, we will introduce the calculations of our accuracy measures and the relevant judgment errors. Second, we will present meta-analyses on the comparison between groups and nominal groups with regard to the previously discussed baselines and specific error terms. Then, we will provide meta-analyses on G-I transfer. Finally, we will present meta-analytic comparisons between the group performance and the performance of their members right before the group discussions, which will allow us to investigate differential weighting.

Judgment Error

We based our meta-analyses on three different accuracy measures. Apart from the overall accuracy, we also calculated mapping and metric errors. The criteria were calculated as follows:

Overall accuracy. The overall accuracy was measured depending on the specific experiment. The overall accuracy was measured using the mean absolute percentage error (hereafter, *MAPE*). Note that the MAPE does control for the fact that larger errors are committed at higher true values.

Mapping errors. As our indicator of mapping error, we calculated Fisher-z-transformed Spearman's rank correlations between true values and the respective estimates. Note that this correlation represents the mapping knowledge rather than the mapping error. High values indicate low errors and low values indicate high values.

Metric errors. As the indicator of metric error, we calculated participants' mean percentage error (*MPE*).

Table 1

Details of included experiments

Experiment	Items	N	Estimation Task	Nominal Group Condition	Group Composition Measure	Additional Condition
Lippold et al. in prep	20	100	Distances	yes	randomisation	reference group
Lippold et al. 2020c Exp1	30	103	Used Car Prices	yes	based on heterogenity	reference group
Lippold et al. 2020a Exp1	30	60	Used Car Prices	yes	randomisation	-
Lippold et al. 2020a Exp2	30	60	Used Car Prices	yes	based on heterogenity	-
Lippold et al. 2020b Exp1	30	38	Stock Market	yes	based on heterogenity	-
Lippold et al. 2020d Exp1	20	108	Distances	no	heterogenity and to enhance bias	-
Schultze et al. 2012 Exp1	20	43	Distances	yes	based on heterogenity	-
Schultze et al. 2012 Exp2	20	35	Distances	no	based on heterogenity	feedback
Schultze et al. 2012 Exp3	20	30	Distances	no	based on heterogenity	-
Stern et al. 2017 Exp1	10	42	Distances	yes	based on heterogenity	single interaction
Stern et al. 2017 Exp2	10	56	Weight Estimates	yes	based on heterogenity	single interaction

Note:

Items represent the number of items presented in the group phase. N refers to the number of nominal groups and interacting groups per experiment which were included in the analysis.

Synergy (group vs. nominal group)

Our meta-analyses comparing group judgments with the baselines from non-interacting nominal groups were conducted on the basis of the raw data from each study. For each experiment, we first computed the judgment errors of interest for the groups and the corresponding baseline models of the nominal groups. Then we computed the differences between the groups and the baseline models of the nominal groups using d, a version of Cohen‘s standard d statistic. We then conducted standard random effects meta-analyses using REML with the metafor package (Viechtbauer,2010).

In this section, we compare the accuracy of real interacting groups with non-interacting nominal groups. Figure 2 displays the results of the corresponding meta-analyses. First, we compared the accuracy of the real group judgments with our criterion for synergy -the average of the nominal group members’ individual judgments- during the group phase for each experiment. The corresponding meta-analysis revealed that the groups were more accurate than the average of the individual judgments (d = 0.62, CI = [0.44 - 0.80], z = 6.74, $p < .001$). The effect size is moderate and substantially larger as previously assumed (Hastie, 1986). The results show that groups systematically outperform the average model and, thereby, achieve synergy.

In addition, we also compared the group judgments with the judgments of the best members of the nominal groups. Note that we determined the best members based on the performance during the group phase. The results revealed that the group judgments were as accurate as the judgments of the nominal groups’ best members (d = -0.09, CI = [-0.36 - 0.19], z = -0.62, $p = .536$). In addition, we compared the group performance with the performance of the initial best member. For this purpose, we selected the member with the highest overall accuracy in the practice phase for each nominal group. We then determined the overall accuracy in the group phase for this person. The groups performed slightly better in comparison to the nominal groups’ initial best members (d = 0.23, CI = [0 - 0.45], z = 1.99, $p = .047$). Note that the resulting effect size is about one third of the comparison between group and average model.

To gain a better understanding about what kind of judgment errors groups commit in

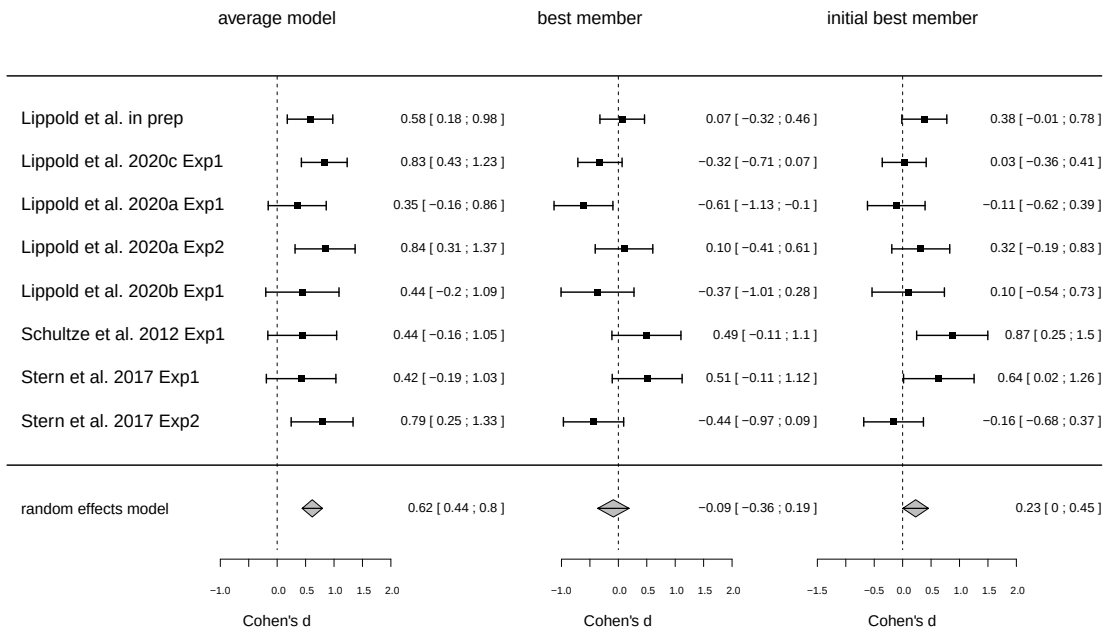

Figure 2. Results of the random-effects meta-analyses of the comparisons of overall accuracy between the groups and the different types of nominal baseline models.

comparision to individuals, we conducted meta-analyses over the mapping and metric error components. First, we compared the mapping errors of real groups to nominal groups (see Figure 3). To this end, we computed nominal groups' mapping errors based on the average of the three nominal group members' judgments(average model), based on the estimates of the nominal groups' members with the lowest mapping errors in the group phase(best member) and based on the estimates of the nominal groups' members who had the lowest mapping errors in the practice phase(initial best member). The three meta-analyses revealed roughly similar results as those for the overall accuracy. The groups outperformed the average of the individuals' judgments (d = 0.57, CI = [0.32 - 0.82], z = 4.42, $p < .001$). The groups were as accurate as the best members of the nominal groups (d = -0.01, CI = [-0.29 - 0.27], z = -0.08, $p = .936$), and they outperformed the nominal groups' initial best members (d = 0.54, CI = [0.36 - 0.72], z = 5.94, $p < .001$).

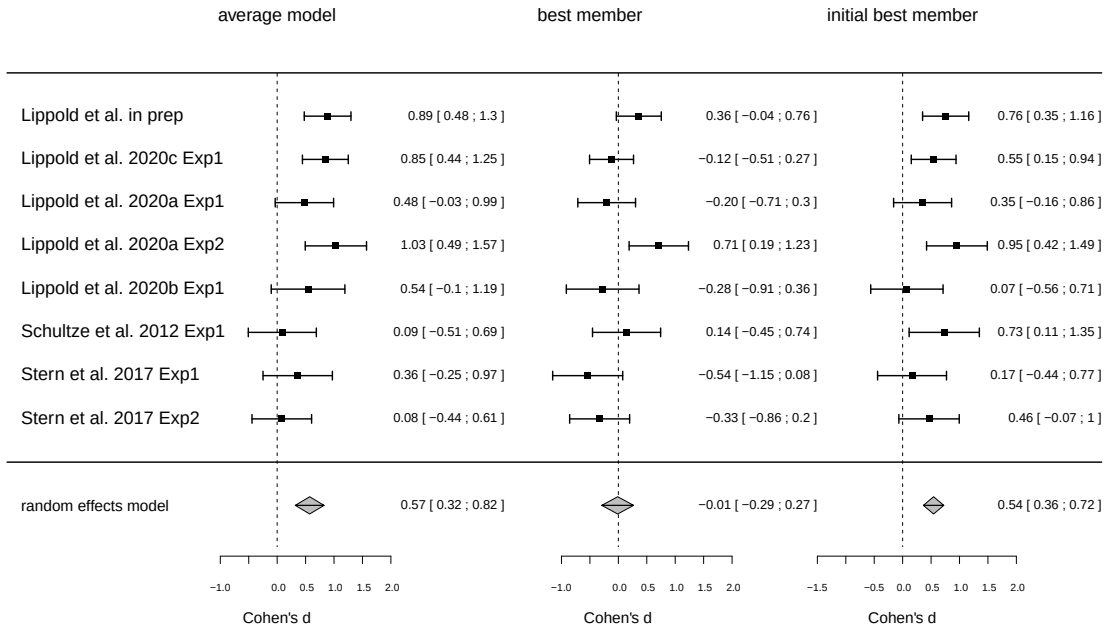

Figure 3. Results of the random-effects meta-analyses of the comparisons of mapping between the groups and the different types of nominal baseline models.

Furthermore, we compared the metric errors of real groups to the metric errors of the nominal groups (see Figure 4). To this end, we computed the three baseline models based on the nominal groups' metric errors. For the best member model, we determined the scores based on the members with lowest metric error in the group phase. For the initial best member model, we selected the members with the lowest metric error in the practice phase. The groups outperformed the average of the individuals' judgments (d = 0.32, CI = [0.11 - 0.53], z = 2.95, $p = .003$). Note that the effect size is about one half of the comparison between group and average model on the overall accuracy. Additonally, the group judgments were less accurate as the judgments of the nominal groups' best members (d = -0.40, CI = [-0.68 - -0.12], z = -2.76, $p = .006$), and as accurate as the nominal groups' initial best members (d = -0.04, CI = [-0.28 - 0.21], z = -0.29, $p = .769$).

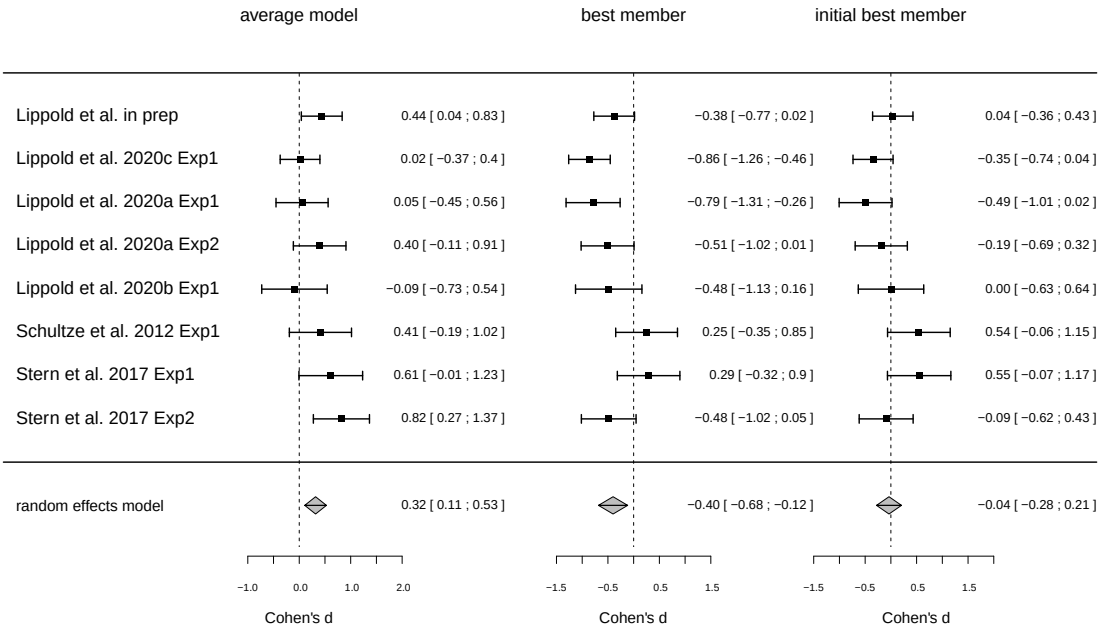

Figure 4. Results of the random-effects meta-analyses of the comparisons of metric between the groups and the different types of nominal baseline models.

Individual Learning (G-I transfer)

To investigate the extent to which group members benefit from G-I transfer, we conducted three meta-analyses. We calculated G-I transfer as the differences between participants' individual scores in the practice phase and their individual scores in the group phase. Positive values indicate an increase in individual capability. For each experiment, we compared the average individual performance increase in the real groups with the average increase in the nominal groups. In order to learn from the other group members, group members must have previously engaged in group work. Hence, we excluded the first trial of the group phase for each experiment, because the individual measures in this trial were taken before any group interaction had taken place.

First, we investigated G-I transfer with respect to the overall accuracy (see Figure 5). The meta-analysis showed that the individual overall accuracy increases more in the real groups than

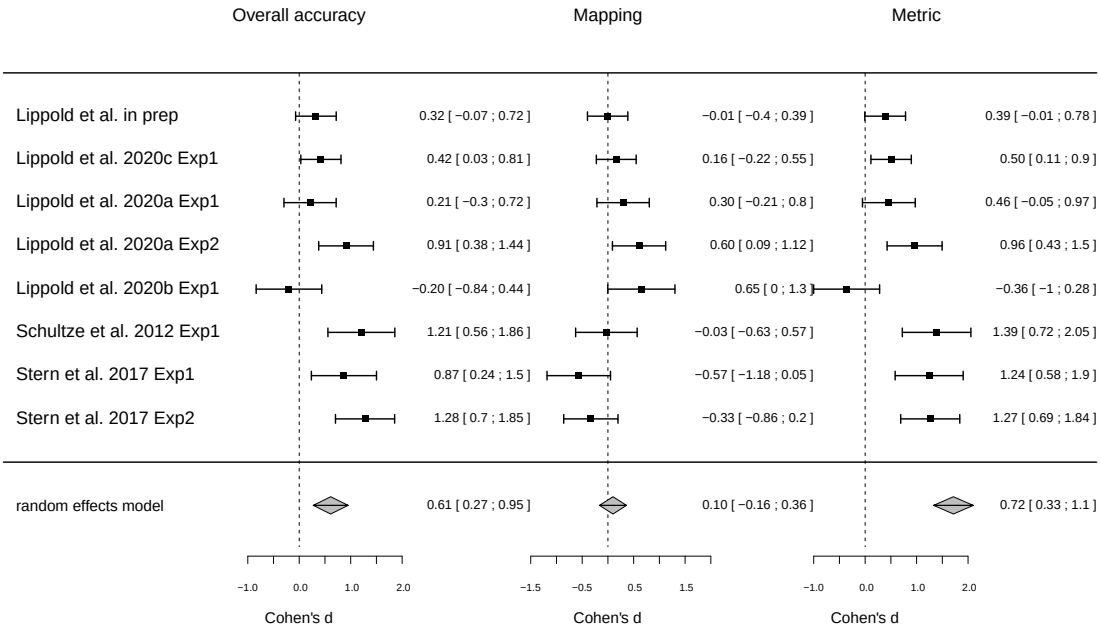

Figure 5. Results of the random-effects meta-analyses of G-I transfer based on the different error components.

in the nominal groups (d = 0.61, CI = [0.27 - 0.95], z = 3.53, $p < .001$). In another meta-analysis, we also found an increase in the individual metric in the real groups in comparision to the nominal groups (d = 0.72, CI = [0.33 - 1.10], z = 3.64, $p < .001$). Note that the resulting effect size was slighly larger than the effect size of the G-I transfer for the overall accuracy. In contrast, the meta-analysis with regard to the transfer of mapping knowledge did not show support for G-I transfer with regard to this error component (d = 0.10, CI = [-0.16 - 0.36], z = 0.74, $p = .462$). However, as Lippold and colleagues (2020) have pointed out, group members' individual improvements in mapping knowledge might be difficult to detect in the group phase of the al-G design. Mapping knowledge is measured via Spearman's rank correlation, which assesses the right order of all included stimuli. Therefore, it is important that the learning process is finished before the mapping knowledge is measured. In contrast to the transfer of metric knowledge, the transfer of mapping knowledge might take longer. Hence, it is unclear which individual trials in the group phase should be used to determine the mapping

knowledge. Therefore, Lippold et al. (2020) assessed the increase in individual mapping with the help of an additional individual phase after the group phase. They calculated G-I transfer with regard to mapping knowledge on the basis of an increase between the practice phase and this additional individual judgment phase. Therefore, we conducted another meta-analysis and included only those four experiments which had such an additional individual phase. We then compared the individual increase of mapping knowledge from the practice to the additional individual phase between the groups and nominal groups. This analysis showed support for the transfer of mapping knowledge (d = 0.45, CI = [0.06 - 0.84], z = 2.28, $p = .023$). Note the transfer of mapping knowledge was accompanied by an higher increase in overall accuracy (d = 0.59, CI = [0.26 - 0.92], z = 3.46, $p = .001$) and a transfer of metric knowledge (d = 0.53, CI = [0.27 - 0.78], z = 4.05, $p =< .001$). Note that the corresponding effect sizes seem all to be on a similar level.

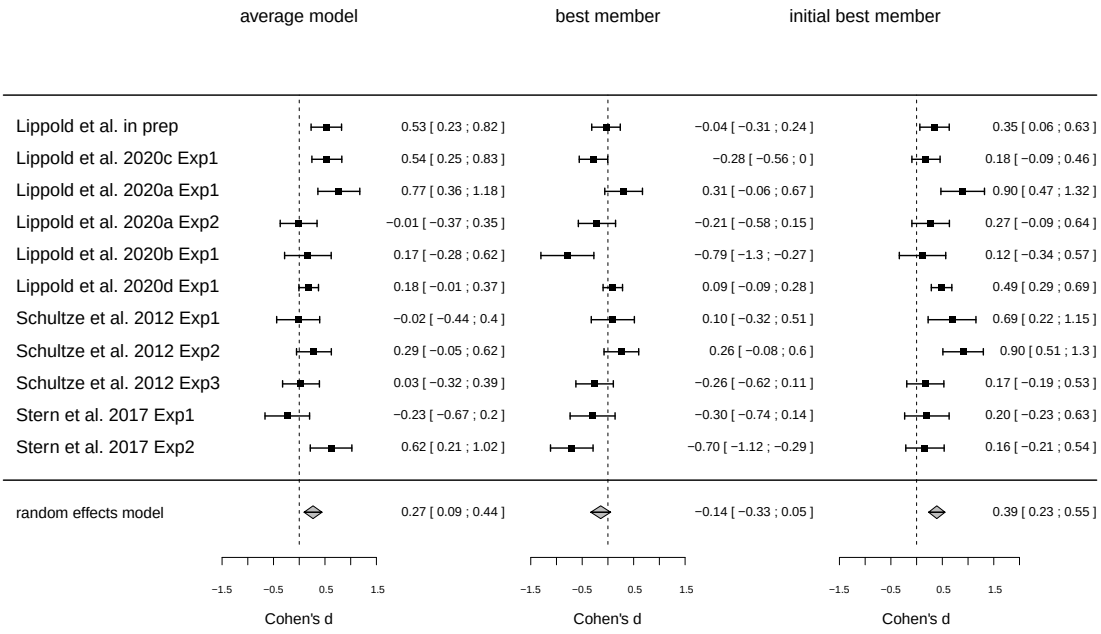

Figure 6. Results of the random-effects meta-analyses of the pairwise comparisons of overall accuracy between the groups and the different types of ingroup baseline models.

Differential weighting (group vs. group members)

The following analyses are concerned with the comparison between the accuracy of the group judgments and the accuracy of the aggregates of the group members' initial judgments. These initial individual judgments were made immediately prior to the group judgments on each trial in the aI-G design. By comparing the group judgments with the aggregates of these pre-discussion judgments of the group members, we can detect synergy as a consequence of effective differential weighting. That is, real groups would achieve synergy by weighting their more expert members or more accurate individual judgments more strongly when forming the consensus estimates. In contrast to the comparison with the nominal groups, we are now able to include all studies conducted, as all of them included a group condition in the al-G design.

To disentangle G-I transfer and differential weighting, we excluded the first trial of the group phase for each experiment from the following comparisons because the first group judgment may already have benefitted from G-I transfer, whereas the individual pre-discussion judgments for the first trial of the group phase cannot benefit from G-I transfer (they were provided prior to any group interaction; see Schultze et al., 2012). We compared the accuracy of real group judgments to the average of the real group members' corresponding individual pre-discussion judgments. The corresponding meta-analysis revealed that the overall accuracy of the group judgments was higher than the overall accuracy of the average of the group members's pre-discussion judgments (d = 0.27, CI = [0.09 - 0.44], z = 3, $p = .003$). In other words, synergy due to differential weighting took place. As with the previous analyses, we also compared group judgment accuracy with the best performing group members. In particular, we determined the best members and the initial best members based on the individual performance of the group members per group. While the groups performed on the level of their most accurate members based on the group phase (d = -0.14, CI = [-0.33 - 0.05], z = -1.41, $p = .158$), the groups exceeded their most accurate initial members (d = 0.39, CI = [0.23 - 0.55], z = 4.82, $p < .001$). The results of these three meta-analyses can be found in Figure 7.

To understand why and how the groups were able to weight their members' judgments effectively, we disentangled the underlying judgment errors further. In particular, we conducted

additional meta-analyses with the mapping and metric error terms as dependent variables. We compared the mapping errors of group estimates to that of the baseline models based on the individual contributions of the group members (see Figure 8). To this end, we computed the individual mapping scores based on the average of the three group members' individual judgments, based on the groups' most knowledgeable members in terms of mapping knowledge in the group phase, and based on the groups' most knowledgeable members in terms of mapping knowledge in the practice phase. The groups outperformed the average of the individual judgments (d = 0.34, CI = [0.23 - 0.44], z = 6.44, $p < .001$). The group judgments were on a similar level as the judgments of their best members (d = 0.19, CI = [-0.04 - 0.42], z = 1.59, $p = .111$), but outperformed the their initial best members (d = 0.88, CI = [0.63 - 1.13], z = 6.84, $p < .001$). Note that the effect sizes for these mapping effects are considerable larger than the effect sizes for the similar comparisons based on the overall accuracy (see previous paragraph).

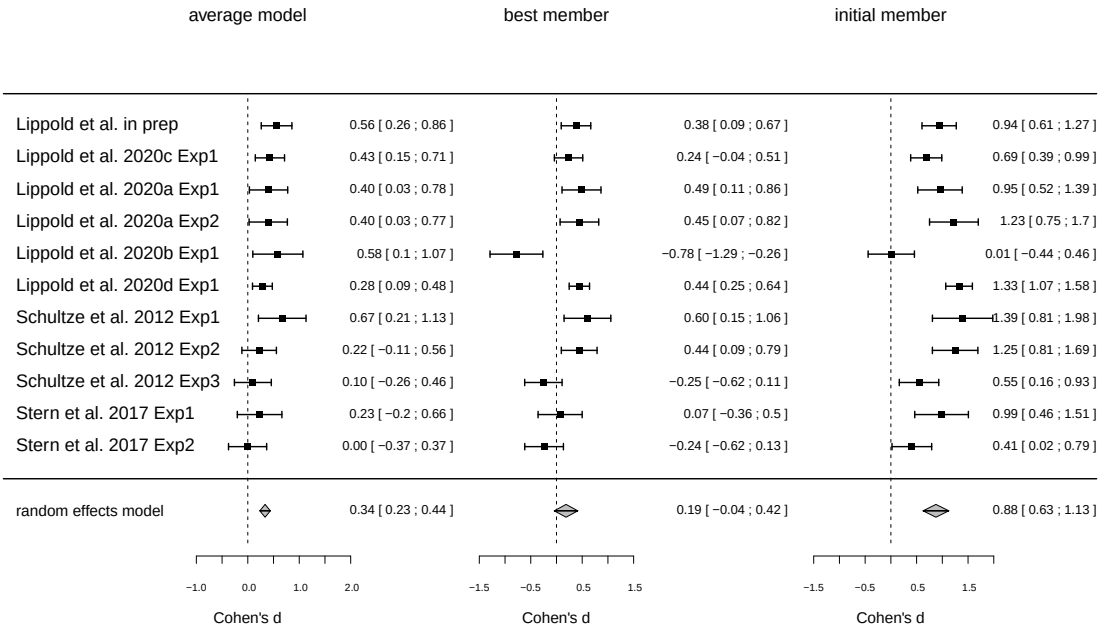

Figure 7. Results of the random-effects meta-analyses of the pairwise comparisons of mapping between the groups and the different types of ingroup baseline models.

Furthermore, we also compared the metric errors of real groups to the averaged metric

errors of their members prior to discussion (see Figure 9). To this end, we computed group members' metric errors based on the average of the three individual members' judgments, based on the groups' most knowledgeable members in terms of metric knowledge in the group phase, and based on the groups' most knowledgeable members in terms of metric knowledge in the practice phase. While the groups outperformed the average of the individual judgments (d = -0.01, CI = [-0.20 - 0.17], z = -0.15, $p = .882$), the group judgments had larger metric errors as the judgments of their groups' best members (d = -0.88, CI = [-1.02 - -0.74], z = -11.92, $p < .001$), and also larger metric errors as their groups' initial best members (d = -0.30, CI = [-0.43 - -0.17], z = -4.52, $p < .001$).

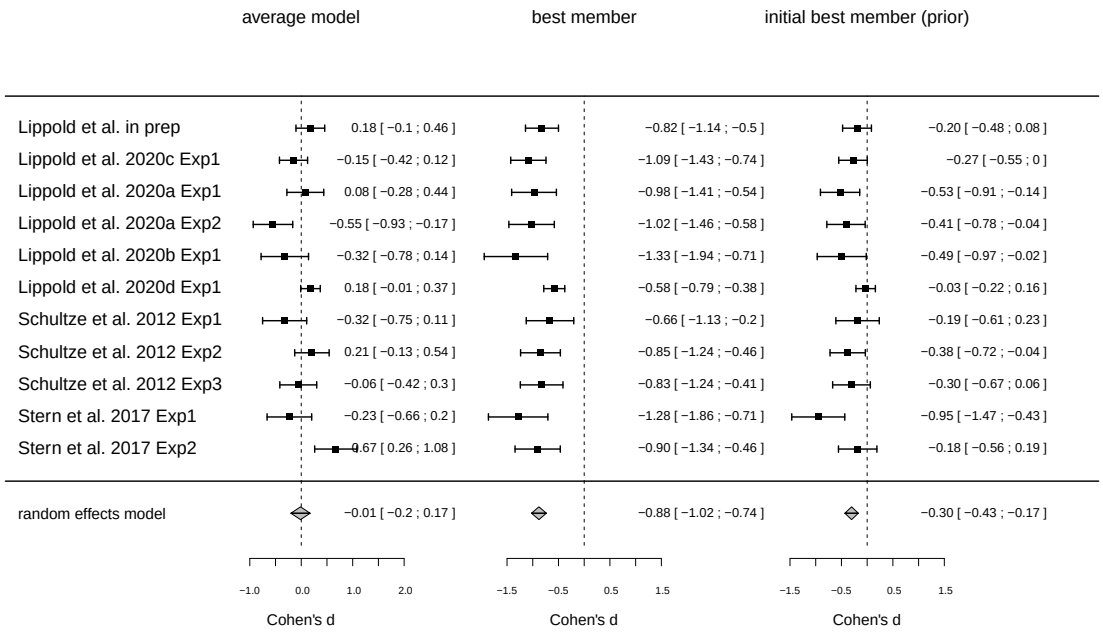

Figure 8. Results of the random-effects meta-analyses of the pairwise comparisons of metric between the groups and the different types of ingroup baseline models.

Discussion

In the present research, we investigated the judgment accuracy of interacting groups in quantitative judgments, with a particular focus on synergy in group judgment, and identified different mechanisms affecting group judgment accuracy. For this purpose, we meta-analyzed

group performance in quantitative judgment tasks from 11 different experiments conducted in our lab. We would have been happy to also include studies conducted in other labs but, as outlined in the beginning, for the purposes of our research question it was essential that the studies used the aI-G design (or any other design that allows for a measurement of individual performance gains in the group phase), and we did not find studies from other labs that fulfilled this precondition. The experiments that went into our meta-analyses included a wide range of quantitative judgments tasks, such as distance and weight estimates, as well as multi-cue judgment tasks.

First, we tested whether and, if so, to what extent groups achieve synergy. To this end, we compared group judgments to the average of a comparable number of individual judgments in the nominal conditions of the corresponding experiments. We chose this baseline, because forming the average of independent judgments made by individuals in the nominal groups does already enhance the judgment accuracy over the accuracy of individual judgments, due to statistical error cancelation. However, this statistical benefit of averaging does not reflect a real group process. Hence, to determine genuine group processes, group judgments need to be more accurate than the average of individuals working alone. The result of the corresponding meta-analysis, which was based on altogether eight experiments, confirmed that groups systematically outperformed the average of individuals working alone. This indicates a considerable advantage of (interacting) groups over nominal groups in quantitative judgment tasks. Against the backdrop of studies from very different fields of group research, this finding shows quantitative judgment tasks to be one of the few cases where the benefits of group work have been successfully demonstrated. In particular, in many other areas of group research, groups usually do not outperform nominal groups; they even often underperform relative to their group's potential (e.g., Kerr & Tindale, 2004; Steiner, 1972). For example, in the seminal experiments by Ringelmann (1913), where groups and individuals conducted physical tasks, or in brainstorming experiments (e.g., Mullen, Johnson, & Salas, 1991), groups predominantly exhibit processes losses (i.e., performance below the group potential). And even under very favorable conditions, when the individual group members possess unique knowledge, which would theoretically allow them to make a superior decision compared to what the members

would decide individually, groups often do not make use of their inherent advantages (Schulz-Hardt & Mojzisch, 2012). In contrast to these other fields of group research, we have now systematic evidence that groups are able to achieve synergy in quantitative judgment tasks. We will next discuss our results with regard to the specific mechanisms underlying this synergy in group judgments, namely G-I transfer and differential weighting.

Group mechanisms affecting accuracy

The first proposed mechanisms leading to synergy in group judgments is G-I transfer (Schultze et al., 2012), that is, an increase in individual task-related capabilities due to collaboratively working on a task in groups. Our results indicate that previous findings of G-I transfer in quantitative group judgment tasks are robust (Lippold et al., 2020; Schultze et al., 2012; Stern et al., 2017). Indeed, our analyses showed that group members were able to improve their individual accuracy due to their participation in a group. To further understand the particular content of this individual learning, we relied on the taxonomy by Brown and Siegler (1993) and decomposed the judgment errors into their metric and mapping components. We were able to confirm previous findings that group members primarily improve their metric during the group discussion (Lippold et al., 2020; Stern et al., 2017), as the meta-analytic effect size of the metric improvement was even slightly larger than the corresponding effect size of the overall G-I transfer (i.e., the improvement of the overall accuracy).

With regard to mapping, the main meta-analyses revealed no evidence that group participation led to a systematic improvement from the practice to the group phase. However, those four experiments that were specifically designed to test for a possible transfer of mapping knowledge by means of an additional individual post phase did, indeed, show an improvement of overall accuracy, metric knowledge, and - most importantly - also mapping knowledge from the practice to the additional post phase. The additional individual post phase was added in those experiments to ensure an unbiased measurement of the individual mapping knowledge. Mapping knowledge is measured across multiple trials. It is, therefore, necessary that the learning process has ended before the individual mapping knowledge is assessed. While it is unknown after how many group trials the transfer of mapping knowledge is completed in the

group phase, no further learning can occur in the additional individual post phase. In the meta-analysis of these four experiments, the corresponding effect size of the improvement of mapping knowledge proved to be slightly smaller than the effect sizes for the overall accuracy improvement and the improvement of metric knowledge. It seems that, via group discussions, members' metric-related judgment errors are corrected somewhat more than mapping-related judgment errors. Presumably, only few pieces of information about the metric need to be exchanged in order to instigate such improvements. For example, the exchange of reference values during the group discussion might help participants to adjust their metric (see Stern et al., 2017, for a similar argument). In sum, group members can improve their individual metric and mapping knowledge during group discussion, which results in a higher overall group accuracy.

While G-I transfer is a mechanism that affects the capabilities of the individual group members, we will now focus on the group's ability to coordinate within the group by combining their individual members input, as differentially weighting their members' inputs in accordance with their accuracy or the members' competence is the second proposed mechanism leading to synergy. When controlling for G-I transfer, previous research was mostly unable to detect an improvement of group judgment accuracy due to differential weighting (Lippold et al., 2020; Schultze et al., 2012). Only under specific circumstances, namely when participants systematically overestimated the target values, it could be demonstrated that groups weighted their individual pre-discussion judgments more effectively than by just averaging them (Stern et al., 2017, Exp. 2). In contrast to most single studies, we found a small, but nevertheless significant overall effect for differential weighting in our meta-analysis. In particular, group members seem to decide on a more accurate consensus group judgment than their average model would, indicating that they were able to give more weight to more accurate individual judgments (or more accurate members) in the group judgment process. As the estimated average effect size was $d = 0.27$, the magnitude of this effect may simply have been too small to be detected in the previous single studies[1].At the same time, when evaluating the size of the differential weighting effect, we have to take into consideration that its magnitude might have

[1] Additionally, the differential weighting effect was stronger in the more recent studies. These studies had more trials in the group phase, which might have led to stronger effect sizes due to higher reliability of measurement.

been limited by the strong G-I transfer that we observed. Because of this transfer, the variability of the members' initial judgments was reduced, as initially weaker members benefit more from G-I transfer than initially stronger members. This should make it harder for the group members to differentiate between their initial individual judgments and weight them based on their members' competence. Hence, contrary to the majority of previous findings, we can assume that the accuracy of group judgments is not only the result of an improvement of individual task-related capabilities, but also the result of coordination within the group, in the form of effective weighting. This leads to the question of how a group is able to weight their members' input effectively.

Functional weighting in groups

Based on our fine-grained analyses, we have identified two weighting strategies which could potentially account for the effective weighting that we observed in the groups. The first strategy involves the identification of the group's most accurate member and relying solely on her input. While it is difficult for a group to be more accurate than the average of their initial judgments, relying solely on the input of the best group member (i.e., the group member who's judgments are most accurate in the group phase) has the potential to be a successful weighting strategy. Averaging is most effective if the included judgments are independent or even negatively correlated (Davis-Stober et al., 2014), given that under these conditions averaging leads to a high degree of statistical error cancelation. Dependencies in the form of information exchange between group members reduce this known benefit of averaging (Lorenz et al., 2011). Hence, while the accuracy of the group members' individual pre-discussion judgments benefited from G-I transfer, the benefit of statistical error reduction should have been reduced due to group discussion. Supporting this idea, in our studies, the best group members were more accurate than the average of the group members' initial individual judgments. Importantly, groups weighted their members' initial judgments as effectively as if they were able to identify their best member during the group phase, and then give full weight to this member's judgments. Therefore, a possible group strategy of fully relying on the best members' input is in line with our meta-analytic results.

However, for two reasons we are not convinced that the groups did actually follow this best-member strategy. On the one hand, it is difficult for a group to identify their best member without corresponding feedback (Kurverts et al., 2020). Even an omniscient authority would only be able determine the best group member in hindsight, when the tasks are completed, and the individual capabilities of all members are known. On the other hand, other motives might prevent groups from solely relying on the performance of one particular group member. For example, there is some evidence that groups might weight their members' judgments in accordance with fairness concerns (Mahmoodi et al., 2015). Similarly, Bonner and Baumann (2012) found that groups weight their members input primarily based on centrality by giving more weight to judgments that are closer to the average of their members' initial judgment. Therefore, we find it more plausible that the groups followed a different weighting scheme that we will describe in the following.

This second weighting strategy relates to our empirical finding that the effective differential weighting effect could be attributed more or less entirely to mapping. Groups weighted their individual judgments in a way that fewer mapping errors occurred compared to what would have happened if they had averaged their initial judgments. Notably, this effect was even larger than the overall differential weighting effect. In other words, groups weighted their initial members' judgments in a way that substantially reduces inconsistencies over trials, leading to an enhanced mapping of the resulting consensus group judgments. Therefore, we propose that the groups follow what we call a *mapping error monitoring strategy*. We assume that the group members primarily weight their members' input based on centrality (meaning that outliers are discounted) but, at the same time, specifically check whether a judgment would be inconsistent with the rank-order of the objects that have already been judged. Consider, for example, a group estimating distances between German cities. Let us assume that this group already made two group judgments. The group estimated the distance between Hamburg and Berlin to be 300 kilometers and the distance between Stuttgart and Hannover to be 600 kilometers. In the third trial, the group now has to judge the distance between Leipzig and Berlin. The three group members make the following initial judgments: 250, 360, and 500 kilometers (the true distance is about 200 kilometers). The average of their judgments would

indicate an estimate of 370 kilometers. However, one group member mentions that both Leipzig and Berlin are located in the eastern part of Germany and that, in comparison to the distance between Hamburg and Berlin, the distance between Leipzig and Berlin therefore should be smaller. As a consequence, the group might agree on a consensus group judgment of 280 kilometers, which is more accurate and more in line with the true rank-order than the average in this case.

This proposed strategy is consistent with previous research emphasizing the role of demonstrability for in-group weighting (Bonner & Baumann, 2012). In particular, Bonner and Baumann showed that only under conditions of high demonstrability, whenever the right solution can be clearly demonstrated, do groups weight their members' input based on accuracy. As it is easier to judge and explain the relation between two objects than to judge and explain a specific value for an object (e.g., Hsee, 1996; Hsee & Zang, 2010), mapping errors are relatively well demonstrable in consecutive judgment tasks. Therefore, by bringing up previous judgments and by appealing on the mapping knowledge of the other group members, a group member might be able to demonstrate a superior solution to the other group members, as illustrated in our example above. Note that groups can bring together the unique knowledge and skills of its members within this weighting scheme, which is a prime example for synergy. For instance, one group member might have superior mapping knowledge and might be able to identify the correct rank-order of the objects in question. However, in order to do that, she needs to be aware of the previously made group judgments. While she might not have a sufficient recollection of all earlier made group judgments, another group member with better memory capabilities might be able to provide these in the group discussion. This contribution by the other group member would enable the group member with the best mapping knowledge to demonstrate possible mapping errors, which would then enhance the accuracy of the group judgments and, as a result, lead to synergy. Our mapping error monitoring strategy seems to be a plausible and effective weighting strategy that groups might apply in consecutive judgment tasks. However, in comparison to the best member model, this strategy in its current form does not allow for a direct test with our current data, as we cannot infer specific point predictions. Therefore, we have to postpone firm conclusions about weighting strategies in group judgments until specified

versions of this general strategy have been formulated and empirically tested.

A plea for groups in quantitative judgments

Practitioners often have to decide whether to assign judgment tasks to a group, to an individual, or to several independent individuals. Based on the results of our meta-analyses, we can recommend the choice of groups for quantitative judgments. Our research shows that groups outperform all baselines that could realistically be applied in practice. The first practical baseline would be to rely on the particular person who is expected to make the most accurate judgments. Note that in companies or organizations, tasks are often allocated based on the assumed competences of the individual members, which are usually inferred from previous task performances. Accordingly, when calculating this baseline, we determined the initial best member based on the members' individual performance before the formation of the groups, and we found that the group judgments were slightly more accurate than the initial best members of the nominal groups were. This suggests that, in group judgment tasks, even the most skilled individuals, identified prior to the task, do not fully reach the performance that interacting groups may achieve. In other words, the individual best member may not substitute a group. Similarly, the group judgments were more accurate than the average of a comparable number of individuals, indicating that the group cannot be substituted (without losses) by simply averaging independent individual judgments - a substitution which the literature on wisdom of the crowds might have suggested (e.g., Armstrong, 2006). In addition to the higher group judgment accuracy, as we have consistently shown, group discussions increase the group members' capabilities to perform the judgment tasks individually. This training function of groups is meaningful as organizations can, for example, use groups as a tool to increase the accuracy of forecasters. On these grounds, we infer that organizations may use group work to generate accurate judgments, and to train individual members of the organization for quantitative judgment tasks.

While the above-mentioned arguments already make a strong case for the implementation of groups for quantitative judgment tasks, our findings indicate that groups have another ace up their sleeves: Groups perform exceptional well with regard to the mapping error dimension.

According to our meta-analyses, groups commit fewer mapping errors than both the average model of the nominal groups and the initial best member model do. Note that mapping itself is highly relevant for judgment quality, and we can think of various circumstances under which the correct mapping of judgments might be even more important than the overall accuracy of these judgments (cf., Lippold et al., 2020). Consider, for example, the case of organ donation. For each patient, doctors have to judge the severity of the patient's condition and the likelihood of a successful transplantation. Based on these ratings, patients are placed on an overall list. To ensure that the right person will receive the next available organ, the order of the patients on this list is crucial. Metric errors and the overall accuracy do only play a subordinate role in this situation. The doctors might overestimate or underestimate the severity of each patient's illnesses and by that commit extremely large overall judgment errors; however, they might still come up with the right order of patients on the list, and this is what will (and has to) determine subsequent action. And this is just one of many examples where the priorities lie on the correct rank-ordering of the objects in question, like personnel selection decisions, picking players in the NFL draft, or selecting financial funds for investing corporate or private money. As outlined above, our meta-analytic results suggest that groups should particularly be used in these situations.

Limitations and future research

Some limitations of the current research need to be addressed here. The first caveat concerns the al-G design with which all our experiments included in the meta-analyses were conducted. By using this design, we were able to disentangle G-I transfer and differential weighting. However, with this design, all group members necessarily form and report independent individual estimates prior to the beginning of each trial of the group discussions. Therefore, this measurement might already constitute a (mild) form of intervention, as all group members must, by means of providing these estimates to the other group members, contribute to the group outcome in this design. In practice, not all group members might contribute to the group in this way. In particular, it might be even the case that some group members do not form an individual estimate at all, and that they might simply "free-ride" during discussion. By

artificially *forcing* all individuals to contribute, the group discussions might be more thorough and elaborate than discussions without such a prior individual estimate. In fact, one recent study suggests that group judgments might be more accurate when the individual members form independent judgments prior to discussion as compared to group judgments made without this kind of intervention (e.g., Minson et al., 2017). Therefore, the accuracy of the group judgments in the meta-analyses might be somewhat exaggerated, and the resulting effect sizes might have a small bias.

Another potential limitation arises from the fact that all experiments included in the meta-analyses solely relied on ad-hoc groups from student samples, working on quantitative judgment tasks that were particularly suitable for the use in the laboratory. This might restrict the external validity of our findings. The groups in the included experiments were formed ad-hoc, meaning the group members did not know each other and only met once in the laboratory for a short time period. In practice, however, working groups might meet on a regular basis and their group members often have a personal history with each other. Therefore, the group interactions in our ad-hoc groups might be different from the group interactions occurring in real working groups. For example, real working groups might be aware of their group members' specific strength and weaknesses. This might allow real working groups to weight their members' inputs more effectively than the ad-hoc groups in our experiments. Additionally, the used tasks in the experiments included in our meta-analyses might not be the most typical quantitative judgment tasks that groups are assigned to outside of the laboratory. In practice, groups are primarily confronted with forecasting tasks, where the events to be judged have not happened yet and the true values cannot be known beforehand, such as predicting the next quarterly figures for a company. The tasks in our experiments, however, mainly consisted of simple general world knowledge judgments and multi cue judgments, and we do not yet know whether our findings hold in the case of, for example, forecasting tasks under uncertainty. Therefore, future research should address the effectiveness of small groups by tasking groups with predictions of real future events. For example, groups and nominal groups can be tasked with the estimation of the value of specific stocks or with the estimation of the final scores of upcoming football games.

Finally, we have to address another limitation. In our analyses, we compared group performance with three baseline models, namely the average model and two different best member models. There are, however, other potential baselines based on the aggregate of independent individual performances that could be taken into account, and that we have not analyzed here. For example, according to the so-called "confidence heuristic" (Thomas & McFadyen, 1995), groups weight their members' input dependent on their individual confidence for each trial. Another example is based on a more sophisticated algorithm like *pivoting* (Palley & Sol, 2019). Pivoting is an aggregation technique that can potentially capitalize on the differentiation of shared and unshared knowledge of individual judges. For this technique, all judges do not only provide their best judgments, but also have to make a guess about the average individual judgments made by other judges. By analyzing the deviations between these two types of responses, it is possible to create aggregates that can be highly accurate. While the confidence heuristic and pivoting have both the potential to be more accurate than our used baselines and, therefore, might represent higher hurdles for the groups, their formation relies on supplementary information. The acquisition of extra information and the forming of these aggregates (for example, by a third person) would lead to extra costs in practice. Therefore, these baselines might not be as practically applicable as relying on the known best individual or as forming the average is. Furthermore, it might be argued that gaining the additional information needed for these more sophisticated baselines is, in itself, a social interactive process - meaning that such baselines would not work as theoretical indicators of what could be achieved in the absence of group processes. Still, future research might want to compare the accuracy of interacting groups with additional baselines to learn more about the effectiveness of groups.

Conclusion

Groups working on judgment tasks are highly accurate. The results of the meta-analyses show that groups outperform a comparable number of individuals, thereby, achieving synergy. This synergetic effect can be attributed to G-I transfer and differential weighting. Our findings highlight the benefits of relying on groups to perform judgment tasks.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting*, (5), 3-15.

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology*, *102*(2), 337.

Bonner, B. L., & Baumann, M. R. (2008). Informational intra-group influence: the effects of time pressure and group size. *European Journal of Social Psychology*, *38*(1), 46-66.

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.

Brodbeck, F. C., & Greitemeyer, T. (2000). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes and Intergroup Relations, 3,* 159-182.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, *121*(1), 149.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.). *Decision research* (Vol. 2). Greenwich, CT: JAI Press.

Hertel, G., Kerr, N. L., & Messé, L. A. (2000). Motivation gains in performance groups: Paradigmatic and theoretical developments on the Köhler effect. *Journal of personality and social psychology*, *79*(4), 580.

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

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, *67*(3).

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343-355.

Keck, S., & Tang, W. (2018). Elaborating or aggregating? The joint effects of group decision-making structure and systematic errors on the value of group interactions. *Unpublished Manuscript*

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol.*, *55*, 623-655.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., ... & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, 5(11), eaaw9011.

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

Laughlin, P. R., Gonzalez, C. M., & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice*, *7*(1), 55.

Larson, J. R. (2010). *In search of synergy in small group performance*. Psychology Press.

*Lippold, M., Schulz-Hardt, S., & Schultze, T. (2020a). *Exploring the minimal conditions for G-I transfer in quantitative group judgments.* Unpublished manuscript.

*Lippold, M., Schulz-Hardt, S., & Schultze, T. (2020b). *Exploring the minimal condition for transfer of mapping knowledge.* Unpublished raw data.

*Lippold, M., Schulz-Hardt, S., & Schultze, T. (2020c). *Exploring the role of bias for group coordination in quantitative judgement task.* Unpublished raw data.

*Lippold, M., Schulz-Hardt, S., & Schultze, T. (2020d). *G-I transfer in forecasting tasks*. Unpublished raw data.

*Lippold, M., Schulz-Hardt, S., & Schultze, T. (2020e). *G-I transfer in multi-cue judgment tasks: Discussion improves group members' knowledge about target relations.* Manuscript submitted for publication.

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*(22), 9020-9025.

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

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.

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and applied social psychology*, *12*(1), 3-23.

Palley, A. B., & Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information is Shared. *Management Science*, *65*(5), 2291-2309.

Ringelmann, M. (1913). Recherches sur les moteurs animés: Travail de l'homme. *Annales de l'Insitut National Agronomique*, *12*, 1-40.

*Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes, 118*, 24-36

Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology, 23,* 305-343.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*, 1–28.

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes, 45*(1), 66-84.

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

Steiner, I. D. (1972). Group Process and Productivity Academic Press. *New York*.

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science, 11*(4), 273-283.

*Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). How Much Group is Necessary? Group-To-Individual Transfer in Estimation Tasks. *Collabra: Psychology, 3*(1).

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group?. *Journal of experimental Psychology, 15*(5), 550.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations, 296*.

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology, 16*(1), 97-113.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48.

Appendix C.

Curriculum vitae

Curriculum vitae

Not available in the online version of this thesis.