ORIGINAL RESEARCH REPORT

How Much Group is Necessary? Group-To-Individual Transfer in Estimation Tasks

Alexander Stern, Thomas Schultze and Stefan Schulz-Hardt

G-I transfer denotes an increase in individual performance due to group interaction, for example, because of acquiring certain skills or knowledge from the other group members. Whereas such G-I transfer has been successfully shown for problem-solving tasks, evidence for G-I transfer on quantitative estimation tasks is scarce. We address this research gap with a focus on how often a group has to interact in order to fully exploit the benefit of this learning effect. Results from two experiments support the idea that a single group interaction is sufficient to induce a stable G-I transfer, which reduces group members' metric error. Smaller metric errors indicate that people improved their representation of the correct upper and lower boundaries, or what range of values is plausible. In contrast to nominal groups, both members of continuously interacting groups and members of groups with only one initial interaction exhibited stable G-I transfer, and the size of this transfer did not significantly differ between the latter two conditions. Furthermore, we found evidence for differential weighting of group members' individual contributions that goes beyond sheer individual capability gains under certain circumstances, namely in tasks with a population bias.

Estimation tasks like forecasting prospective profits for a company, or estimating the expected increase in global temperature, lie at the heart of many far-reaching financial or political decisions. Important quantitative judgments are often made in groups, because groups are considered to be superior to a comparable number of individuals with regard to performance. For example, there is evidence that groups outperform comparison individuals (e.g., Bonner & Baumann, 2012; Bonner, Sillito, & Baumann, 2007; Laughlin, Bonner, Miner, & Carnevale, 1999) or even perform on the level of very challenging baselines like the most accurate group member's estimates (e.g., Einhorn, Hogarth, & Klempner, 1977; Laughlin, Gonzalez, & Sommer, 2003).

Beyond that, and independent of the actual group performance, group interaction might have another beneficial effect: individuals who interact in groups are assumed to exhibit particular social learning effects and thereby solve subsequent similar tasks more accurately than individuals who have no prior group experience. This possible individual capability gain as a consequence of prior collective task performance in a group is called *group-to-individual transfer* (G-I transfer, e.g., Laughlin &

Barth, 1981; Laughlin & Sweeney, 1977). Despite the fact that there is robust evidence for such group learning processes in problem-solving tasks (e.g., Laughlin, Carey, & Kerr, 2008; Laughlin & Ellis, 1986; Stasson, Kameda, Parks, Zimmerman, & Davis, 1991), this phenomenon has been mostly neglected in research on quantitative group estimations. To our knowledge, the only exception is a study by Schultze, Mojzisch, and Schulz-Hardt (2012), which found strong individual performance improvements in quantitative estimations after group interaction. However, it is yet unclear what people learn and whether they need ongoing social interaction to maintain this improved performance. Therefore, in the present research we try to find out which knowledge is transferred when interacting with others. Furthermore, we focus on the repetitions of group interaction in order to accomplish stable individual performance enhancements. In other words, we investigate if one group interaction is sufficient to produce a significant increase in individual accuracy and, most importantly, whether it persists after group members leave the group.

Group-to-individual transfer
Building on the dynamic model of group performance by Brodbeck and Greitemeyer (2000a), we understand group learning as a function of two sources of change in individual resources that can improve groups in a complementary way. On the one hand, group members can improve their capabilities to work efficiently with each

University of Goettingen, Institute of Psychology, Göttingen, DE
Corresponding author: Alexander Stern
(stern@psych.uni-goettingen.de)

other (*learning to collaborate*). For example, group members might develop a shared mental model of the task, or they could acquire knowledge about the expertise of the other group members. On the other hand, and this is what we focus on in this paper, group members can improve their individual task-related skills as a consequence of group interaction, independent of purely individual practice effects (*learning to perform the task*). As already mentioned, this socially induced individual learning is known as *G-I transfer*. Examples of such learning processes are vicarious learning, or exchange of basic principles and strategies for effective task performance (e.g., Laughlin & Jaccard, 1975; Brodbeck & Greitemeyer, 2000a, 2000b).

So far, most research on G-I transfer has been conducted in the domain of problem-solving tasks, with ample empirical evidence for its existence. For example, participants who had worked on mathematical problems in a group later solved the same or other, logically related problems better than individuals working alone (e.g., Laughlin & Ellis, 1986; Stasson et al., 1991). Similarly, participants with prior group interaction exhibited better individual performance in rule induction tasks compared to participants without such group interaction (e.g., Brodbeck and Greitemeyer, 2000b). By using multiple training sessions, Laughlin et al. (2008) addressed the necessary repetitions of group interaction in order to achieve G-I transfer, the major result of which was that one group interaction was sufficient for the occurrence of a stable G-I transfer. In other words, multiple group interactions did not affect the strength of the individual performance enhancements. However, all of these studies worked with tasks that are very likely characterized by high levels of demonstrability, which is considered to be a prerequisite for the occurrence of G-I transfer (Brodbeck & Greitemeyer, 2000a). According to Laughlin and Ellis (1986), one of the core conditions of demonstrability is that the member with the correct answer must have the ability, motivation, and time to demonstrate the correct solution to the other group members. On mathematical problems, this should usually be the case. When the task complexity is moderate, the member with the right solution should be able to explain its correctness. In contrast, on quantitative estimation tasks, it might be difficult for the best group member to demonstrate the high quality of his or her estimate, and for inferior group members to understand its quality. This lack of demonstrability might have some consequences on the occurrence of G-I-transfer. For example, it is more difficult to justify an estimation of New York City having around 8.4 million inhabitants than explaining that 5 times 7 equals 35. Nevertheless, as long as people do not simply guess their estimations on world knowledge tasks, it is generally possible to explain why certain estimations are better than others. This should be especially true when it comes to rather poor estimations. For example, one might explain quite easily why the population of New York City cannot be 200 million when taking into account that the whole United States of America has around 320 million citizens. On the other hand, it should be more difficult to judge whether the city has either 6 or 7 million inhabitants. Therefore, G-I transfer might have somewhat higher hurdles in estimation tasks as compared to problem-solving tasks.

Beyond that, differences in task demonstrability should have another important consequence. When tasks are highly demonstrable, as is often the case with arithmetic problems (e.g. Laughlin & Ellis, 1986; Stasson et al., 1991), the most capable member often determines the group outcome. In other words, one capable group member can be sufficient for solving the task. On this basis, group-level performance often does not benefit from individual capability gains because, even if the less capable group members become more capable over time, it is unlikely that their contributions will add anything beyond that of the most capable member. In addition, it is also unlikely that this most capable member will improve his or her performance in the absence of superior models to learn from. In contrast, on tasks with a somewhat lower demonstrability, like estimation tasks, the accuracy of group estimations usually benefits from taking *all* group members' opinions into account (e.g., in the form of weighted or unweighted averages), because this can help to eliminate or, at least, reduce idiosyncratic errors. In other words, exclusively relying on the most capable member is usually not the best strategy in quantitative estimation tasks (Bednarik & Schultze, 2015; Soll & Larrick, 2009). Consequently, the group as a whole might benefit from improved individual performances of inferior group members. This fact makes the field of quantitative estimation tasks particularly interesting, because here individual capability gains could actually lead to a performance enhancement on the level of the entire group.

To the best of our knowledge, there is only one study combining estimation tasks with a design that allows to detect increases in individual accuracy as a consequence of prior group interaction. Typically, studies in this field use a so-called *I-G design* (individual-group design, e.g., Bonner et al., 2007; Henry, 1993, 1995; Henry, Strickland, Yorges, & Ladd, 1996; Sniezek & Henry, 1989, 1990), meaning that participants first complete a series of quantitative estimation tasks individually (I) and then work on the same series of tasks as groups (G). Unfortunately, this design cannot account for G-I transfer, because individual performance is not measured *after* one or more group interactions have taken place. To address this limitation, Schultze et al. (2012) used an improved *aI-G design* (alternating-individual-group design). Their experiments were separated into two sections: (a) an individual practice phase and (b) a group phase consisting of alternating individual and group estimates of distances between different European capital cities. Consequently, changes in participants' individual accuracy due to the group interaction could be measured on their subsequent individual estimates. With this modified design, the authors found evidence for strong increases in individual accuracy after the first within-group interaction. In other words, group members already improved their individual performance after the first group discussion. In line with the idea of G-I transfer, inferior group members improved in accuracy while the groups' best members' accuracy remained stable. Furthermore, the improved estimation accuracy

was relatively constant after this first major performance enhancement. Hence, it seemed that participants did not substantially benefit from further group interaction. These results are in line with the above-mentioned evidence from group problem-solving research, showing that one group interaction can be sufficient for the occurrence of a stable G-I transfer (Laughlin et al., 2008) in this type of task.

However, the findings of Schultze et al. (2012) leave two important question unanswered. The first one is *what* do people learn when interacting with others? To answer this question, it is useful to differentiate two types of estimation error. As outlined by Brown and Siegler (1993; see also Brown, 2002), one's knowledge in a judgment domain can be decomposed into two components: *metric knowledge* and *mapping knowledge*. Metric knowledge is a general understanding of the appropriate scaling, that is, whether people have an accurate representation of the correct upper and lower boundaries, or what range of values is plausible. For example, knowing that Germany has a length of approx. 900 kilometers, and that the equator has a length of roughly 40,000 kilometers, helps when estimating distances, and prevents us from making judgments that are completely implausible. In contrast to that, mapping knowledge is an accurate representation of the relative magnitude of possible target values. In other words, mapping knowledge allows us to put different target values of the same kind in the correct order. Most people know that the distance between London and Paris is shorter than the distance between London and New York, without necessarily having a good guess about the actual distances.

Interacting with others when working on estimation tasks should affect these two sources of estimation error differently, and to a different extent. Previous studies imply that providing people with frames of reference can strongly increase their estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Collaborating with others during quantitative estimations could have exactly this effect: During their task-related communication, group members provide the reasoning for their individual estimates and illustrate the validity of certain benchmarks (Schultze et al., 2012). With regard to the two above-mentioned sources of error, such reference values should mainly improve metric knowledge quite rapidly and thereby diminish particularly implausible estimates. In contrast, reducing one's mapping error during social interaction should be more difficult than understanding differences in group members' metric error. When estimating distances between European cities one can only recognize that another group member has a different mapping error, for example, when realizing that he or she always overestimates distances between southern European cities and always underestimates distances between northern European cities. In other words, one need to precisely remember multiple estimates of other group members in order to recognize such differences. Consequently, the process of reducing one's mapping error should be very slow and only possible after a long period of cooperation.

The second open question has to do with the stability of the learning process. Specifically, we do not yet know whether the G-I transfer is stable even if the group is completely disbanded, or whether continuous social interaction is needed for its maintenance. In other words, it is crucial to find out how the individual performance develops after the last group interaction. In the experiments of Schultze et al. (2012), participants continuously alternated between working on the estimation tasks individually and in groups, that is, they remained in a group context until the end of the experiment. Hence, so far there is no research about the temporal stability of G-I transfer in quantitative estimation tasks after members have left the group, and whether, under these conditions, one single group interaction leads to an equally strong individual performance enhancement compared to continuous group interaction.

Answering this question is not only relevant to gain a more conclusive theoretical understanding of the mechanisms that underlie G-I transfer. Rather, it is also a crucial question for practical purposes: The results of the Schultze et al. (2012) study suggest that it might be sufficient to have just one group interaction to fully exploit the benefits of having groups work on quantitative estimation tasks. As bringing group members together and having them discuss and decide on an issue costs more effort than just collecting and averaging individual judgments, truncating group interaction right after the first group judgment would, obviously, save a lot of resources. However, this would only pay off if the benefits of this interaction (i.e., the G-I transfer) do not fade away relatively soon after this first interaction. As Schultze et al. (2012) do not provide an empirical test for a sustainable beneficial G-I transfer after one single group interaction, we want to address this research gap in the current study.

Therefore, it is crucial to (a) replicate the finding of a strong increase in individual accuracy after just one group interaction, (b) analyze if group members increase their metric knowledge after interacting with others and (c) check whether their individual performance enhancement remains stable even if the first group interaction is also the last, that is, if all subsequent individual trials take place without any further group interactions in between. In other words, our study investigates what people actually learn during a group interaction and whether a single group interaction is sufficient to achieve a stable improvement in individual performance, or whether a robust transfer requires ongoing group interaction. We present two experiments exploring these issues. In each, we report all measures and manipulations. Furthermore, no participants were excluded from analyses.

Hypotheses
Schultze et al. (2012) found evidence that one group interaction might be sufficient for a strong increase in individual accuracy in quantitative estimation tasks. Our first aim is to test whether this effect is replicable by comparing two experimental group conditions (differing in the number of group interactions) to a control condition with nominal groups, that is, an equivalent number

of non-interacting individuals. We predict that one group interaction is sufficient to achieve a significant increase in individual accuracy (G-I transfer):

> **Hypothesis 1:** *Group members' individual accuracy will increase after the first group interaction. Both members of continuously interacting groups as well as members of groups with only a single group interaction will manifest these performance enhancements, whereas individual accuracy in the nominal groups will not improve at all.*

More importantly, we aim to answer the question of whether the stability of the G-I transfer requires ongoing group interaction. One single group interaction might be sufficient to achieve a stable performance enhancement. However, it is still possible that permanent group interaction is crucial for the stability of the increased individual estimation accuracy. In other words, the individual performance could deteriorate when the group is disbanded. As a consequence, we formulate two competing hypotheses regarding this research question:

> **Hypothesis 2a:** *The increase in individual accuracy after the first group interaction is stable even when the group is disbanded.*

> **Hypothesis 2b:** *The increase in individual accuracy after the first group interaction deteriorates after the group is disbanded.*

Furthermore, we are interested in whether there are differences between the two experimental conditions. Once again, there are two different possibilities, both of which we consider to be plausible. On the one hand, if the individual estimation accuracy is similarly stable after one as compared to many group interactions, the individual estimation accuracy should be (more or less) equally strong in both conditions. On the other hand, the increase in individual accuracy might deteriorate after the group is disbanded, or continuous group interaction could additionally foster the G-I transfer. This, in turn, should lead to stronger performance enhancements for members of continuously interacting groups. Accordingly, we also formulate two competing hypotheses for this issue:

> **Hypothesis 3a:** *The increase in individual accuracy over the course of the individual trials is equally strong for members of single-interaction groups and continuous-interaction groups.*

> **Hypothesis 3b:** *Members of continuous-interaction groups will manifest a stronger increase in individual accuracy than members of single-interaction groups.*

Finally, we aim to test *what* group members learn when interacting with others. We do not expect transfer of mapping knowledge since the process of reducing one's mapping error should be very slow and only possible after a long period of cooperation. In contrast, if the exchange of reference values underlies individual performance

enhancements, group members should mainly reduce their metric error. Hence, we hypothesize:

> **Hypothesis 4:** *Interacting with others will reduce group members' metric error. Both members of continuously interacting groups as well as members of groups with only a single group interaction will manifest this transfer of metric knowledge, whereas non-interacting individuals will not improve their metric error.*

Although the focus of our study is on individual performance after group interaction, for exploratory purposes we will also investigate group performance in comparison to individual performance. As hypothesized, group members might benefit individually from group interaction, which, in turn, could make groups better than an equivalent number of individuals. Hence, we will conduct an exploratory test of whether such surplus at the group level occurs in our study. Furthermore, we will also look at the possible occurrence of differential weighting strategies, that is, groups weighting more competent members more strongly. In addition to G-I transfer, such weighting strategies might also contribute to the quality of group judgments complementarily. So far, the only study that controlled for individual performance enhancements did not find any evidence for differential weighting (Schultze et al., 2012). Nevertheless, this latter finding need not necessarily be generalizable, because Schultze et al. only used one specific type of task. Consequently, we want to analyze if groups engage in differential weighting strategies on different quantitative estimation tasks. However, since individual capability gains are the focus of our study, we refrain from formulating hypotheses and, instead, analyze these two questions in an exploratory manner.

Experiment 1
In Experiment 1, we aimed to investigate whether the previously found individual performance enhancements due to group interaction (Schultze et al., 2012) are replicable, and whether or not this increase in accuracy requires ongoing group interaction. In other words, we wanted to find out whether a single group interaction has the same beneficial effect as multiple interactions. For this purpose, we compared continuously interacting groups with groups that were disbanded after their first within-group interaction, and with nominal groups. Members of both continuous-interaction and single-interaction groups should provide more accurate judgments individually due to the G-I-transfer. Furthermore, the experimental design allows us to examine whether G-I-transfer is equally strong after multiple group interactions in comparison to just one, which could then be interpreted as evidence for its stability beyond the group context.

Method
Participants, design and task
One hundred eighty-three German or German-speaking students (112 women, 70 men, one participant did not report his or her gender) with an average age of 21.43 (*SD* = 3.28) years participated in the experiment, with three persons each forming a real or nominal group.

The sample size was based on a previous relevant study (Schultze et al., 2012). Experiment 1 used a mixed design with group type (continuous-interaction, single-interaction, no interaction) as a between subjects variable and task trial (or, for some analyses, trial block) as a within subjects variable.

The participants worked on a set of distance estimations between different European capital cities. This is the same task that has been used by Schultze et al. (2012). It was chosen based on two pretests ($N = 40$ and $N = 38$) revealing that there were stable differences in participants' individual performance (mean Spearman's Rho = .28, $p < .001$ and mean Spearman's Rho = .35, $p < .001$, respectively), which is a prerequisite for learning processes (Schultze et al., 2012). We measured accuracy with the mean absolute percent error (MAPE). In group judgment research, the MAPE is a common measure of accuracy (e.g., Sniezek & Henry, 1989, 1990) and indicates the average deviation of the estimates from the true values. The average MAPE scores in the two pretests of Experiment 1 indicated that pretest participants' deviated, on average 60.22 percent ($SD = 63.55$) and 48.31 percent ($SD = 21.14$) from the true values. Furthermore, we checked whether participants' estimates were evenly distributed around the true values or whether the task contains a systematic *population bias*, that is, whether they tended to over or underestimate the true value. For this purpose, we calculated participants' mean percent deviation from the true values (thereby allowing overestimations and underestimations to cancel out each other). Corresponding t-tests against zero revealed no significant differences ($M = 16.31$, $SD = 79.72$), $t(39) = 1.29$, $p = .203$, $d = 0.29$, and ($M = -2.59$, $SD = 38.94$), $t(37) = -.41$, $p = .685$, $d = 0.09$, respectively, indicating that this task contained no substantial population bias.

Procedure

In each experimental session, six to nine participants were invited and randomly guided to one of three lab rooms, where they were placed at separate tables. Participants were informed about the task and the procedure of the experiment. They were instructed that the experiment consisted of two phases with ten distance estimates each: an individual practice phase and a group phase. Hence, participants knew from the beginning that they were going to interact unless the number of participants showing up was not divisible by three. In this case, excess participants were assigned to the individual control condition. The specific distances that the participants should estimate were not identical between the two phases but were, as the pretests indicated, on average, of similar difficulty.[1] In the practice phase, participants were asked to work on ten trials individually, and they were told that the goal was to estimate the airline distances between cities in kilo meters as accurately as possible. Furthermore, the experimenter asked them not to communicate or to exchange notes. There was no time limit, but participants usually took between ten and fifteen minutes to finish this phase. Once they were done, the experimenter collected the data and computed the MAPE for each participant. Afterwards, the participants were assigned to three-person-groups.

Whenever possible, we aimed for some heterogeneity in group members' skill level, as a certain amount of heterogeneity is necessary for individual capability gains and differential weighting strategies. To this end, groups were composed so that there was a minimum difference of 10 percentage points between the MAPE scores of the most capable and medium group member as well as between the medium and least capable group member. Hence, participants' practice phase MAPE influenced the assignment to the three conditions.[2]

The second phase of the experiment differed depending on the experimental condition. In the continuous-interaction and the single-interaction condition, three participants each formed a group and were asked to take a seat at a shared table. The groups received four questionnaires containing the estimation tasks, one for each group member to write down his or her individual estimates, and one for the group estimates. The difference between the two group conditions was the number of group interactions. In continuous-interaction groups, the group members first worked on a specific distance estimate individually and then discussed their individual estimates in order to come up with a consensus estimate. Afterwards, they proceeded with the next trial of the group phase in the same fashion. Participants were told that they were neither allowed to inspect their estimates of trials they had already worked on, nor to revise these previous estimates. Furthermore, they were reminded not to communicate or exchange notes when working on their individual estimates. The single-interaction groups only interacted on the first trial of the second phase. Again, before interacting as a group, each group member had to come to an individual estimation. After discussion of the first task and making a group judgment, the single-interaction groups were disbanded, and their members were placed at separate tables where they worked on the remaining nine trials independently and without further discussion. The individual judgments of the nine trials on which participants in the single-interaction condition worked on their own were later averaged to form hypothetical group judgments. In the nominal group condition, participants worked on all ten estimates of the second phase individually, that is, they worked at separate tables and were not allowed to communicate or exchange information. Subsequently, the judgments of the three nominal group members were averaged to create the nominal group judgments. Participants had no guidelines regarding how to work on a particular task and, again, there was no time limit.

In each condition, the experimenter explained that the accuracy of the estimates during the second phase would determine the amount of money the participants would receive for participating in the experiment. In addition to a show-up fee of 5 Euro, there was an accuracy-based bonus payment ranging from 0 to 5 Euros.[3] After completing phase 2, participants were asked to fill in a final questionnaire containing a suspicion check. In the meantime, the experimenter calculated the MAPE score of the second phase to determine the bonus payment. Before the participants were dismissed, they were thanked for their participation and debriefed.

Results and discussion

Group-to-individual transfer

In order to test for individual performance enhancement in terms of G-I transfer, we analyzed whether group interaction led to improved subsequent individual estimations. For reasons of simplification, we compared the differences in individual MAPE scores between the individual practice phase and the group phase in the three experimental conditions and did not add the two phases as a within-subjects factor. The trial right before the first group interaction was treated as the last trial of the individual practice phase, since this trial still took place before any effects of group interaction could have occurred.[4] Positive values of the accuracy difference measure represent an increase in accuracy from phase one to phase two. We conducted a 3 (*group type*: continuous-interaction vs. single-interaction vs. no interaction) × 3 (*group member*: most capable vs. medium vs. least capable) ANOVA with experimental condition as between-subjects factor and group member as within-subjects factor. The analysis revealed a significant main effect of group type, $F(2, 58) = 4.84$, $p = .011$, $\eta_p^2 = .14$. LSD post hoc comparisons showed that the performance enhancements where roughly similar in the continuous-interaction and the single-interaction condition ($M = 19.28$, $SD = 20.68$ vs. $M = 16.13$, $SD = 17.09$), $p = .611$.[5] However, participants increased their performance significantly more in both the continuous-interaction groups and the single-interaction groups than in the non-interacting nominal groups ($M = 19.28$, $SD = 20.68$ vs. $M = 1.77$, $SD = 20.12$), $p = .005$, and ($M = 16.13$, $SD = 17.09$ vs. $M = 1.77$, $SD = 20.12$), $p = .023$, respectively. Furthermore, separate post hoc *t*-tests against zero revealed a significant performance enhancement in the continuous-interaction condition, $t(20) = 4.27$, $p < .001$, $d = 0.94$, as well as in the single-interaction condition, $t(18) = 4.11$, $p = .001$, $d = 0.95$. The nominal group condition, in contrast, showed no significant changes in the individual MAPE scores from the first to the second phase, $t(20) = 0.40$, $p = .692$, $d = 0.09$. Because participants in the nominal group condition did not improve their performance between the two phases, we can assume that there are no substantial individual practice effects in the task we used, which mirrors the findings of Schultze et al. (2012). Accordingly, in line with Hypothesis 1, the increase in judgment accuracy in the two group conditions should be the result of G-I transfer, supporting the idea that one group

interaction is sufficient to increase individual estimation accuracy. The results further indicate that, in general, the performance enhancements were equally strong after one as after multiple interactions. In other words, groups working on the distance estimates were able to exchange all information necessary to induce the full amount of increase in individual accuracy already during their first group discussion, supporting Hypothesis 3a.[6]

The ANOVA further revealed a main effect of group member, $F(2, 57) = 26.14$, $p < .001$, $\eta_p^2 = .31$, which was qualified by an interaction of group member and group type, $F(4, 116) = 4.51$, $p = .002$, $\eta_p^2 = .13$, in line with the idea that differences in G-I transfer can only be observed in the two group conditions and not in the non-interacting control condition. Separate post hoc paired-samples *t*-tests for the continuous-interaction and the single-interaction condition indicated that the accuracy improvements differed between all levels of group members' judgment accuracy, all $ts(20) > 3.58$, all $ps < .002$, all $ds > 0.78$, for the continuous-interaction condition and, all $ts(18) > 3.12$, all $ps < .006$, all $ds > 0.71$, for the single-interaction condition (for an overview of all individual performances changes, see **Table 1**). In contrast, and as expected, there were no significant differences in performance changes as a function of the particular group member's capability in the nominal-group condition, all $ts(20) < 1.27$, all $ps > .222$, all $ds < 0.28$. Additionally, post hoc *t*-tests against zero revealed that in both continuous-interaction and single-interaction groups only the medium and the least capable members improved their estimation accuracy from phase one to phase two, all $ts(20) > 3.44$, all $ps < .003$, all $ds > 0.75$, and all $ts(18) > 2.95$, all $ps < .009$, all $ds > 0.67$, respectively. In contrast, the most capable group members' performance remained, more or less, stable in both conditions, $t(20) = 0.18$, $p = .863$, $d = 0.04$, and $t(18) = 0.89$, $p = .384$, $d = 0.20$, respectively. These differences in increased accuracy depending on group members' judgment accuracy indicate that group members understand from whom to learn or at least whom to ignore. Apparently, this understanding already occurs during the very first within-group interaction. Superior group members seems to share task relevant knowledge that can and should be learned. Consequently, only inferior group members can benefit from G-I transfer. In contrast to this improved estimation accuracy, in the nominal-group condition, none of the group members significantly changed their performance

Table 1: Group members' individual performance changes by group type in Experiment 1.

Group type	most capable		medium		least capable	
	M	*SD*	*M*	*SD*	*M*	*SD*
continuous-interaction	−0.43	11.18	11.40	15.16	46.87	48.56
single-interaction	2.56	12.52	12.42	18.36	33.36	30.01
no interaction	−4.03	11.00	3.89	27.76	5.29	37.27

Performance change was measured as the difference between (nominal) group members' MAPE scores during trials 1 to 11 and the corresponding MAPE scores during trials 12 to 20. Positive indicate a reduction in MAPE scores and, thus, performance increases.

between the two phases, all $ts(20) < 1.68$, all $ps > .108$, all $ds < 0.37$. Hence, the interaction effect of group member and condition is the result of stronger performance enhancements of inferior group members after interacting with superior group members. This finding further supports the idea of G-I transfer and indicates that the most capable group members were the source of the learning process. In contrast, the medium and least capable group members seemed to benefit from the most capable members, leading them to approximate their levels of individual accuracy.

In addition, we conducted a more detailed temporal analysis of the observed individual performance enhancements in the two interacting group conditions to determine when the major increase in individual accuracy occurs. For this purpose, we compared group members' averaged individual accuracy of the trials before the first group interaction (trial 1–11) with the trial immediately after the first group interaction (trial 12), and then the averaged individual accuracy of the remaining 8 trials (trial 13–20). Doing so allowed us to analyze whether the performance enhancement after the first group interaction is relatively stable over time. Accordingly, we conducted a 2 (*group type*: continuous-interaction vs. single-interaction) × 3 (*trial block*: practice phase vs. trial after first group interaction vs. remaining 8 trials) repeated measures ANOVA. This analysis revealed a main effect of trial block, $F(2, 37) = 35.57$, $p < .001$, $\eta_p^2 = .48$, and no main effect of group type, $F(1, 38) = 0.32$, $p = .574$, $\eta_p^2 < .01$, or interaction of group type and trial block, $F(2, 37) = 0.33$, $p = .724$, $\eta_p^2 < .01$. Post hoc paired-samples t-tests showed that in the continuous-interaction condition the average individual accuracy discontinuously increased after the first group interaction ($M = 48.63$, $SD = 19.97$ vs. $M = 26.30$, $SD = 14.17$), $t(20) = 6.58$, $p < .001$, $d = 1.44$,

and remained (more or less) stable afterwards ($M = 26.30$, $SD = 14.17$ vs. $M = 29.73$, $SD = 7.01$), $t(20) = -1.01$, $p = .324$, $d = 0.22$. Moreover, the same was true for the single-interaction condition ($M = 47.99$, $SD = 15.01$ vs. $M = 29.69$, $SD = 11.66$), $t(18) = 5.05$, $p < .001$, $d = 1.16$, and ($M = 29.69$, $SD = 11.66$ vs. $M = 32.13$, $SD = 8.29$), $t(18) = -1.13$, $p = .274$, $d = 0.26$, respectively (see also **Figure 1**). This result suggests that a single group interaction is sufficient to ensure the stability of the observed G-I transfer, in line with Hypothesis 2a.[7]

Changes in metric and mapping error

Beyond that, we were interested in what members of interacting groups actually learn. To this end, we calculated the mean overall deviation (MOD) (Brown & Siegler, 1993), which is a measure of metric property defined as the absolute difference between the median estimate across all items and the true overall median and is therefore less susceptible to outliers than the arithmetic mean. Accordingly, lower values indicate a lower metric error.

However, the magnitude of participants' judgment errors varied strongly with the respective true values. Hence, we worked with the percentage error instead of the absolute deviation from the true values. Nevertheless, the pattern of results remains unchanged when working with the median absolute error instead of the median absolute percentage error in both experiments. Similar to the analyses of group members' MAPE scores, we compared the differences of individual MOD scores between the individual practice phase and the group phase in the three experimental conditions. Thus, positive values indicate decreasing metric errors. Again, the trial right before the first group interaction was treated as the last trial
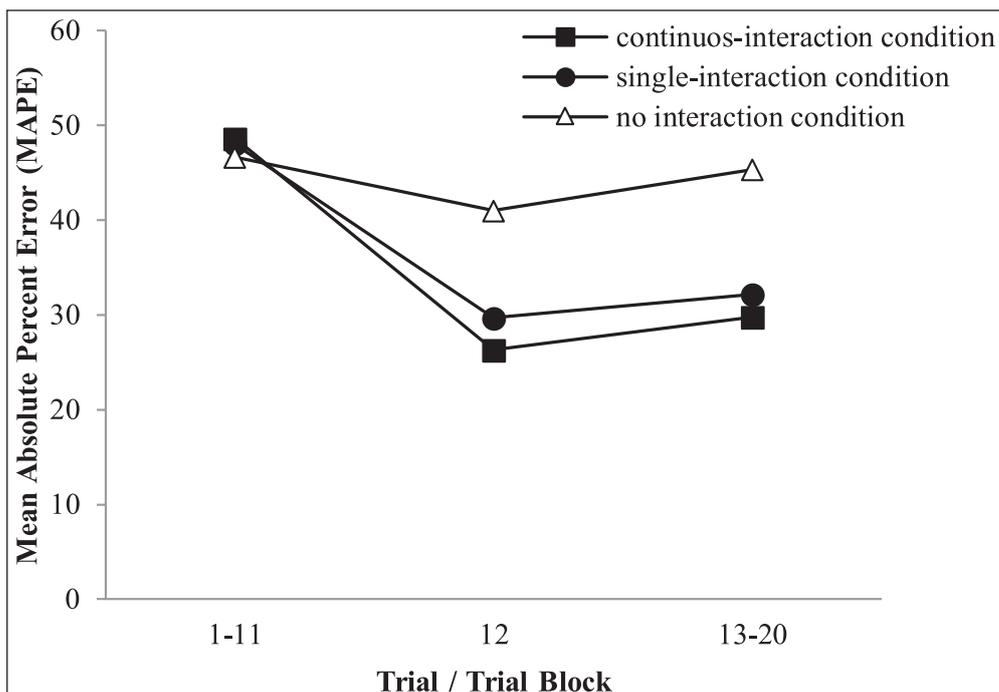

Figure 1: Mean absolute percent error (MAPE) of individual estimates by group type during Experiment 1. Lower scores indicate greater accuracy.

of the individual practice phase and we averaged across group members for reasons of simplicity.[8] We calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor, which showed significant differences, $F(2, 58) = 9.50$, $p < .001$, $\eta_p^2 = .25$. Whereas participants in the continuous-interaction condition and in the single-interaction condition decreased their metric errors ($M = 24.38$, $SD = 23.95$, and $M = 15.38$, $SD = 22.83$), this error even increased for participants in the nominal group condition ($M = -3.52$; $SD = 15.91$). The differences between the two interacting groups condition and the nominal group condition were significant, $p < .001$ and $p = .007$. In contrast, the difference between continuously interacting groups and single interaction groups fell short of significance ($p = .184$) although, descriptively, the metric error reduction was more pronounced among the former than among the latter. In sum, we found evidence that interacting with others reduces group members' metric error, which is line with Hypothesis 4.

Besides the metric error, we were also interested in possible changes in mapping errors, i.e. whether participants were able to put different target values in the correct order. Therefore, we computed rank-order correlations, which represent the correlation between the ranks of estimates with the ranks of true values (Brown & Siegler, 1993), and calculated the difference between group members' Fisher z-transformed averaged rank-order correlation coefficients (Spearman's rho) during the group phase and the individual practice phase. Accordingly, positive values indicate decreasing mapping errors. Similar to the metric error analysis, we calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as between-subjects factor, which revealed no significant differences, $F(2, 54) = 1.71$, $p = .191$, $\eta_p^2 = .06$. Hence, whether participants interacted with others or not had no effect on their mapping error.

Exploratory analyses: Group-level data
In an exploratory fashion, we investigated whether interacting groups outperformed nominal groups. For this purpose, we calculated the MAPE score for the 10 trials of the group phase. In the continuous-interaction condition, these MAPE scores were based on the groups' consensus estimates, whereas in the nominal-group condition these scores were calculated as the average of the three nominal-group members' individual estimates. In the single-interaction condition, the groups' average MAPE score was a composite measure of the group estimate in the first trial of the second phase and the averaged individual estimates in the remaining nine trials. Based on these calculations, our first analysis was an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as a between-subjects factor. This analysis showed no significant effect of group type, $F(2, 58) = 1.56$, $p = .219$, $\eta_p^2 = .05$. Descriptively, the results indicate that both the continuous-interaction groups ($M = 25.97$, $SD = 9.80$) and the single-interaction groups ($M = 25.49$, $SD = 11.07$) performed somewhat better than the nominal groups ($M = 32.05$, $SD = 17.46$). However, due to the rela-

tively high variances within the conditions, the superiority of interacting over non-interacting groups fell short of significance.

Furthermore, we tested whether the performance of continuously interacting groups exceeded the average model that was calculated on the basis of their members' individual estimates right before each of the group trials (i.e., the estimates that already benefitted from G-I transfer). If this were the case, it would indicate that groups differentially weight the proposals of superior members more strongly than the proposals of weaker members. We excluded the first trial of the second phase in order not to artificially penalize the average model and averaged across the remaining trials.[9] The corresponding paired samples t-tests showed that the actual group performance was slightly (but not significantly) inferior to the average model ($M = 25.09$, $SD = 9.51$ vs. $M = 23.42$, $SD = 8.25$), $t(20) = 1.11$, $p = .282$, $d = 0.24$. In general, our results indicate that groups in Experiment 1 were not able to outperform the average model.

However, we cannot yet rule out that the results of Experiment 1 might be task specific, because we used the same task as Schultze et al. (2012). For example, our participants might have had a more or less accurate representation of the map of Europe, which, in turn, could have facilitated learning processes when receiving an accurate point of reference. Beyond that, the task was characterized by low population bias, meaning that there was no systematic trend for participant to overestimate or to underestimate the true values. Consequently, group members could have accomplished individual performance enhancements similar to G-I transfer, by simply centering their individual estimates (because in a task with low population bias the central group member is likely to be the most accurate member). Therefore, the question is whether our findings would still hold if the task were more difficult and if participants would tend to over- or underestimate the true value. Furthermore, the sequence of trials in Experiment 1 – and also in both experiments by Schultze et al. (2012) – was in a fixed order. However unlikely, we cannot rule out that differences in the difficulty of the different trials had an influence on the magnitude of the observed G-I transfer or changes in metric and mapping error. Hence, to validate our findings in terms of replicability and generalizability, we conducted a second experiment with a different task type and a randomized trial order.

Experiment 2
To generalize our findings, we conducted a second experiment with the same design but a different task type, namely estimating the weights of different objects. The task was considerably more difficult and characterized by a strong population bias (see section task and procedure). In spite of these differences, we expected to replicate the results of Experiment 1 with respect to individual performance enhancements. Particularly when taking into account that evidence on G-I transfer in quantitative estimation tasks is extremely scarce so far, we consider a replication of our results as being indispensable.

Method

Participants and design

A total of 252 German or German-speaking students (152 women, 100 men) with an average age of 23.25 (SD = 4.53) years participated in the experiment, with three persons each forming a (real or nominal) group. Experiment 2 used the same mixed factorial design as in Experiment 1, with the group type (continuous-interaction, single-interaction, no interaction) as a between subjects variable and task trial (or trial block) as a within subjects variable.

Task and procedure

The procedure of Experiment 2 was identical to Experiment 1, with the following exceptions: First, participants were asked to estimate the weight of different small items (e.g., hammer, dustpan, or umbrella) that were present in the room, without being allowed to touch or lift them. We chose this task based on two pretests (N = 30 and N = 29) revealing that there were stable differences in participants' individual performance (mean Spearman's Rho = .39, p = .031 and mean Spearman's Rho = .49, p = .008, respectively). Furthermore, this task was evidently more difficult than the task of Experiment 1: The average MAPE scores in the two pretests of Experiment 2 were markedly above the corresponding scores in the two pretests of Experiment 1 (M = 293.02, SD = 234.63 and M = 398.71, SD = 261.07 vs. M = 62.22, SD = 63.55 and M = 48.31, SD = 21.14). Beyond that, in contrast to the pretests of Experiment 1, participants had a strong tendency to overestimate the true values. When calculating participants' mean percent deviation from the true values, these average deviations were significantly greater than zero (M = 281.36, SD = 241.84), t(29) = 6.37, p < .001, d = 1.16, and (M = 395.80, SD = 263.79), t(29) = 8.08, p < .001, d = 1.50, respectively, indicating a large population bias. The second change was that we aimed to rule out that the results obtained in Experiment 1 were in any way due to the fixed order of trials. To this end, we randomly created four different trial orders in Experiment 2 by splitting the 20 trials into two task blocks of 10 trials each. Half of the participants worked on the first block in the individual practice phase and on the second block in the group phase, whereas this order was reversed for the other half. In each sequence, we additionally reversed the order of trials within the two blocks for half of the participants.

Results and Discussion

Group-to-individual transfer

Similar to Experiment 1, we started by testing for increased accuracy of group members' individual estimates consistent with G-I transfer. For this purpose, we again calculated individual performance enhancements by subtracting the individual MAPE scores of the group phase from those of the individual practice phase. Again, the first trial of the group phase (i.e., the trial right before the first group interaction) was counted as the last trial of the individual practice phase, since this trial could not, by definition, be affected by any group interaction. With the MAPE scores as dependent variable, we conducted a 3 (*group*

type: continuous-interaction vs. single-interaction vs. no interaction) × 3 (*group member*: most capable vs. medium vs. least capable) ANOVA with group type as between-subjects factor and group member as within subjects factor.[10] This analysis revealed a main effect of group type, F(2, 81) = 8.39, p < .001, η_p^2 = .17. LSD post hoc comparisons showed that the performance enhancements in the continuous-interaction condition were somewhat stronger than those in the single-interaction condition, but the comparison did not reach conventional levels of significance (M = 87.33, SD = 62.90 vs. M = 49.24, SD = 100.60), p = .077. As the more detailed temporal analysis reported below will clarify, this descriptive difference is indeed most likely due to random variation. As in Experiment 1, participants in both the continuous-interaction groups as well as in the single-interaction groups increased their performance significantly more than participants in the nominal groups (M = 87.33, SD = 62.90 vs. M = 0.79, SD = 70.48), p < .001, and (M = 49.24, SD = 100.60 vs. M = 0.79, SD = 70.48), p = .026, respectively. Furthermore, post hoc t-tests against zero showed a significant increase in individual accuracy in the continuous-interaction condition, t(27) = 7.35, p < .001, d = 1.39, as well as in the single-interaction condition, t(27) = 2.59, p = .015, d = 0.49.[11] The nominal-group condition, in contrast, showed virtually no change in MAPE scores from the first to the second phase, t(27) = 0.06, p = .953, d = 0.01, indicating that, similar to Experiment 1, participants in the nominal-group condition did not improve their performance in terms of practice effects. Hence, the increases in individual accuracy in the other two conditions are the result of G-I transfer, in line with Hypothesis 1. Furthermore, participants did not manifest a significantly stronger G-I transfer after multiple group interactions as compared to a single group interaction, which supports Hypothesis 3a over Hypothesis 3b.[12]

Beyond that, the ANOVA revealed a main effect of group member, F(2, 80) = 40.00, p < .001, η_p^2 = .31, and an interaction of group member and group type, F(4, 162) = 4.72, p = .001, η_p^2 = .10. Separate post hoc paired-samples t-tests for the continuous-interaction and the single-interaction conditions showed that the accuracy improvements differed between all levels of group member expertise for the continuous-interaction condition, all ts(27) > 4.87, all ps < .001, all ds > 0.92, and for the single-interaction condition, all ts(27) > 2.44, all ps < .022, all ds > 0.46. Again, post hoc t-tests against zero revealed significant performance increases for the medium and the least capable members in both interacting group conditions, all ts(27) > 2.22, all ps < .035, all ds > 0.41 (for an overview of all individual performance changes, see **Table 2**). In contrast, there was a tendency for the most capable group members' performance to slightly deteriorate in the continuous-interaction condition, t(27) = −1.87, p = .073, d = 0.35, and even more profoundly in the single-interaction condition, t(27) = −3.13, p = .004, d = 0.59. A similar analysis of the non-interacting nominal groups unexpectedly revealed a significant difference in performance changes between the most capable and the least capable group members, t(27) = 2.90, p = .007, d = 0.55, as well as marginal differences between the medium

Table 2: Group members' individual performance changes by group type in Experiment 2.

| | Group member | | | | | |
| | most capable | | medium | | least capable | |
Group type	M	SD	M	SD	M	SD
continuous-interaction	−29.02	82.30	83.80	94.62	207.22	106.77
single-interaction	−58.72	99.41	45.89	109.36	160.34	239.04
no interaction	−11.88	71.25	−33.21	168.91	46.25	106.30

Performance change was measured as the difference between (nominal) group members' MAPE scores during trials 1 to 11 and the corresponding MAPE scores during trials 12 to 20. Positive indicate a reduction in MAPE scores and, thus, performance increases.

and least capable group members, $t(27) = 2.03$, $p = .052$, $d = 0.38$. However, these differences are unlikely to stem from systematic learning effects; instead, they are likely the result of regression to the mean. Specifically, the least capable group members significantly increased their performance between the two phases, $t(27) = 2.30$, $p = .029$, $d = 0.44$, whereas there were no significant performance changes for the most capable and medium group members, all ts$(27) < 1.05$, all ps > .307, all ds < 0.20. This finding also suggests that not all performance changes in the interacting group conditions can be necessarily attributed to social learning processes. At least a small part might also be ascribed to statistical regression. However, post hoc t-tests revealed that the medium and least capable group members in both the continuous-interaction and the single-interaction condition increased their estimation accuracy more strongly than the medium and least capable members of the nominal groups, all ts$(54) > 2.08$, all ps < .043, all ds > 0.55. Hence, the interaction effect of group member and condition mainly derives from stronger performance enhancements of inferior group members after interacting with others.

Similar to Experiment 1, we were interested in a temporal analysis of the individual performance enhancements in the two interacting group conditions. Therefore, we again compared individual accuracy (averaged across group members) before any group interaction had taken place (trial 1–11) with the trial after the first group interaction (trial 12), and with the averaged individual accuracy of the remaining 8 trials (trial 13–20). The 2 (*group type*: continuous-interaction vs. single-interaction) × 3 (*trial block*: practice phase vs. trial after first group interaction vs. remaining 8 trials) repeated measures ANOVA showed a main effect of trial block, $F(2, 53) = 21.28$, $p < .001$, $\eta_p^2 = .28$, and no main effect of group type, $F(1, 54) = 0.63$, $p = .432$, $\eta_p^2 = .01$, or interaction of group type and trial block, $F(2, 53) = 1.69$, $p = .190$, $\eta_p^2 = .03$. Post hoc paired samples t-tests revealed that the average individual accuracy in the continuous-interaction condition discontinuously increased after the first group interaction ($M = 237.04$, $SD = 81.49$ vs. $M = 148.29$, $SD = 110.84$), $t(27) = 5.29$, $p < .001$, $d = 1.00$, and remained (more or less) stable afterwards ($M = 148.29$, $SD = 110.84$ vs. $M = 149.89$, $SD = 85.76$), $t(27) = −0.09$, $p = .932$, $d = 0.02$. The same was true for the single-interaction condition, with a major performance enhancement directly after

the single group interaction ($M = 229.48$, $SD = 98.59$ vs. $M = 180.04$, $SD = 111.37$), $t(27) = 2.33$, $p = .027$, $d = 0.44$, and virtually no further changes in individual estimation accuracy ($M = 180.04$, $SD = 111.37$ vs. $M = 180.26$, $SD = 112.08$), $t(27) = −0.02$, $p = .987$, $d < 0.01$. This result, illustrated in **Figure 2**, once more suggests that a single group interaction is sufficient to induce stable G-I transfer, which supports Hypothesis 2a.[13]

Figure 2 also indicates stronger individual performance enhancements in the continuous-interaction condition as compared to the single-interaction condition. However, as this analysis shows, it is highly unlikely that this difference is due to the sustained group interaction in the former condition, because the full difference is already present after the first group trial – in other words, at a point in the experiment where the procedure has yet been identical for both conditions – and it remains stable afterwards. Hence, by chance, participants in the former condition seem to have reacted somewhat more strongly to the first group interaction than participants in the latter condition.

Changes in metric and mapping error

Similar to Experiment 1, we were interested in what participants learn when interacting with others. To this end, we calculated an ANOVA with group type (continuous-interaction vs. single-interaction vs. no interaction) as a between-subjects factor and differences between the averaged group members' individual median percentage error between the individual practice phase and the group phase as the dependent variable. This analysis revealed a significant main effect of group type, $F(2, 81) = 10.40$, $p < .001$, $\eta_p^2 = .20$. Additional LSD post hoc comparisons showed no significant differences in metric error reduction between the continuous-interaction and the single-interaction condition ($M = 68.47$, $SD = 60.50$ vs. $M = 41.43$, $SD = 82.25$), $p = .176$. In contrast, participants' reductions in metric error were significantly stronger in both the continuous-interaction groups and the single-interaction groups than in the non-interacting nominal groups ($M = 68.47$, $SD = 60.50$ vs. $M = −19.77$, $SD = 74.68$), $p < .001$, and ($M = 41.43$, $SD = 82.25$ vs. $M = −19.77$, $SD = 74.68$), $p = .003$, respectively. Hence, interacting with others reduced group members' metric error, which supports Hypothesis 4.

Furthermore, we analyzed changes in participants' mapping errors averaged across group members. To this end, we calculated an ANOVA with group type
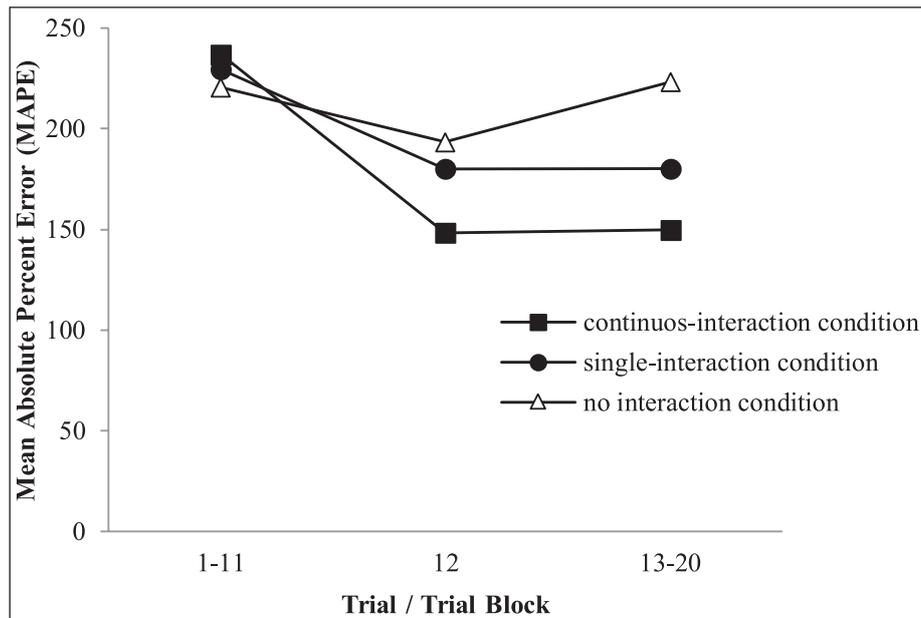
Figure 2: Mean absolute percent error (MAPE) of individual estimates by group type during Experiment 2. Lower scores indicate greater accuracy.

(continuous-interaction vs. single-interaction vs. no inter-action) as a between-subjects factor and the difference between participants' Fisher z-transformed rank-order correlation coefficients (Spearman's rho) during the group phase and the individual practice phase as the dependent variable. This analysis revealed no significant differences between the group types, $F(2, 81) = 1.12$, $p = .307$, $\eta_p^2 = .03$. Hence, there were no systematic difference in partici-pants' mapping error changes.

Exploratory analyses: Group-level data
Similar to Experiment 1, we also analyzed group perfor-mance for exploratory purposes and checked whether interacting groups' judgments were more accurate than those of nominal groups. Accordingly, we calculated the groups' average MAPE score for the 10 trials of the group phase in the same way as in Experiment 1. Hence, in the single-interaction condition, the groups' average MAPE score was a composite measure of the group estimate of the first trial of the second phase and the averaged indi-vidual estimates of the remaining nine trials. Afterwards, we ran an ANOVA with the group type (continuous-interaction vs. single-interaction vs. no interaction) as a between-subjects factor and (nominal or real) group per-formance as the dependent variable. This analysis revealed a significant effect of group type, $F(2, 81) = 4.02$, $p = .022$, $\eta_p^2 = .09$. LSD post hoc comparisons showed that the accuracy of both the continuous-interaction groups and the single-interaction groups were superior to the nomi-nal groups ($M = 142.30$, $SD = 82.06$ vs. $M = 220.21$, $SD = 110.29$), $p = .006$, and ($M = 175.08$, $SD = 114.42$ vs. $M = 220.21$, $SD = 110.29$), $p = .106$, respectively, even though the latter comparison did not reach conventional levels of significance. Beyond that, continuous-interaction and single-interaction groups did not differ significantly with regard to accuracy ($M = 142.30$, $SD = 82.06$ vs. $M = 175.08$, $SD = 114.42$), $p = .238$.

Finally, we were interested in the possible occurrence of functional differential weighting strategies. To this end, and similar to Experiment 1, we tested whether interacting groups outperformed the average of their members' indi-vidual estimates after controlling for G-I transfer. A paired samples t-tests showed that, on average, group estimates were significantly more accurate than the average model; the difference in accuracy was about 11 percentage points ($M = 133.99$, $SD = 80.86$ vs. $M = 144.07$, $SD = 87.09$) $t(27) = -3.37$, $p = .002$, $d = 0.64$. Hence, continuously interacting groups were apparently willing and able to assign differ-ent weights to their members' individual estimates, and they did so in an effective manner, allowing them to out-perform the average of their members' estimates. Hence, Experiment 2 constitutes, to our knowledge, the first evi-dence for functional differential weighting after control-ling for G-I transfer.

These findings raise an interesting question: Why did we find evidence for group members weighting their individual contributions differentially in Experiment 2, but not in Experiment 1, and also not in the study by Schultze et al. (2012)? Of course, at this point we can only speculate about this, but we find it, at least, plau-sible that parts or all of this divergence could be due to differences between the tasks that were used in these experiments: Whereas in Experiment 1 we used the same distance estimation task that had been used by Schultze et al. (2012), and with similar results (no dif-ferential weighting), in Experiment 2 we introduced a new weight estimation task. As already stated, this task was characterized by a large population bias, whereas the distance estimation task of Experiment 1 contained no such bias. Now, the presence vs. absence of a popula-tion bias should have consequences for whether or not the group members' individual values can be expected to bracket the true value: If there is no population bias, that is, if over- and underestimations cancel out

on average, the group members' individual estimates should often bracket the true value. In contrast, if all group members systematically overestimate (or underestimate) the true value, bracketing is less likely to occur. In line with this, we found that the percentage of overall cases where the individual estimates bracketed the true value in Experiment 1 was far above the bracketing rate in Experiment 2 ($M = 62.62$, $SD = 19.85$ vs. $M = 33.10$, $SD = 19.51$). Because more bracketing means that averaging more often leads to accurate results, differential weighting had better chances to pay off in Experiment 2 as compared to Experiment 1.

General discussion
In the present study, we tested whether group members cooperatively working on estimation tasks benefit from G-I transfer. Based on previous findings (Laughlin et al., 2008; Schultze et al., 2012), we expected individual performance enhancements due to group discussion. More specifically, we postulated a stable increase in group members' individual accuracy that persists even when the group is disbanded. Furthermore, we aimed to clarify whether a single group interaction is sufficient to produce G-I transfer, and whether further within-group interaction induces additional performance enhancements. Beyond that, we also wanted to shed some light on what exactly group members learn when interacting with others. We expected a transfer of metric knowledge that should lead to a better calibration of group members' estimates, but also tested the possibility that interaction improves group members' mapping knowledge. In an exploratory manner, we checked for the superiority of groups over an equivalent number of individuals, and for the occurrence of possible differential weighting strategies that improve the group judgments beyond the level of individual capability gains.

In line with our assumptions, the results of our experiments provide evidence for socially induced learning processes as a consequence of group interaction on quantitative estimation tasks. The estimates of superior group members served as a benchmark towards which the inferior group members adjusted their subsequent individual estimates. More precisely, group members reduced their metric but not their mapping error. Importantly, the individual increases in accuracy remained stable even if we disbanded groups after their first group discussion. Since participants in the nominal group condition were not able to enhance their performance over time, the aforementioned result can be interpreted as unequivocal evidence for G-I transfer. Furthermore, additional group interactions did not lead to further increases in individual accuracy, suggesting that groups were able to exchange all information necessary to induce this G-I transfer during their first group discussion. Beyond that, we found first evidence that after group members benefitted from G-I transfer groups are indeed able to assign more weight to more accurate judgments under certain circumstances. We will get back to this finding after having discussed our central results regarding G-I transfer.

Group-to-individual transfer
Our main aim was to test for the relevance of group interaction for the subsequent estimation accuracy of the individual group members. Our results allow us to draw several conclusions: First, one group interaction is sufficient to induce a stable G-I transfer, since in both experiments the performance remained on the improved level even after the group was disbanded. These results are in line with the assumptions of Schultze et al. (2012) and, to the best of our knowledge, constitute the first unambiguous evidence for the stability of this performance enhancement on estimation tasks, thereby mirroring similar findings from the field of problem-solving tasks (Laughlin et al., 2008). This finding also rules out an alternative explanation for the performance changes. Specifically, the anticipation of group discussion could have led to increased feelings of accountability because group members knew that they would have to justify their individual estimates during the discussion. Accordingly, they might have put more effort into their individual estimates, which might have resulted in greater individual accuracy. If this were the case, group members' effort and, thus, their individual accuracy should have reverted to the level we observed prior to the group phase rather than remaining constant. Second, the replicability of the G-I-transfer with a different estimation task provides first evidence for the generalizability of this phenomenon – at least when tasks share some characteristics such as stable differences in expertise and at least some degree of demonstrability.

Third, the fact that we found strong metric error reductions, and that the performance enhancement already occurred after one group interaction on two different task types allows some speculations regarding what people learn when interacting with others on estimation tasks, and what information has to be exchanged in order to produce the observed G-I-transfer. In our opinion, the most likely explanation for the strong reduction of group members' metric error and the rapid increase in estimation accuracy is the exchange of reference values during the first group interaction. As we know from previous research, frames of reference play an important role when it comes to increases in estimation accuracy (e.g., Bonner & Baumann, 2008; Bonner et al., 2007; Laughlin et al., 1999; Laughlin et al., 2003). Such reference values are relatively easy to communicate and to retain, and they should have a beneficial effect on all subsequent judgments in the same domain. As Schultze et al. (2012) discuss, points of reference might provide a basis for better calibration and could enable group members to reduce their individual estimation bias. With this additional information, individuals might be able to improve their individual accuracy, even without any further benchmarks, on subsequent trials. For example, group members might communicate the length of Germany from north to south (approx. 900 km) as a reference value, which might help them when estimating the distance between London and Rome and will prevent very inaccurate estimates. In other words, accurate benchmarks could also serve as a source of error checking. This assumption might also explain the lack of G-I-transfer in an earlier study that used an experimental

design somewhat similar to ours. In one condition of this study, Sniezek and Henry (1990) asked participants to estimate the prices of different automobile models individually before and after interacting with others. During this interaction, group members were allowed to exchange all information relevant to the task with exception of numeric estimates. In other words, they could not provide reference values and, therefore, not reduce their metric error. Nevertheless, future research should systematically investigate the exact nature of the information required to induce G-I-transfer and test whether group interaction provides accuracy gains beyond the exchange of accurate reference values.

Superiority of group judgments and differential weighting

Not only the individual group members but also the groups as a whole seem to benefit from G-I transfer. The results of our experiments generally support the idea that interacting groups outperform nominal groups in quantitative estimation tasks. Although the respective comparisons only reach conventional levels of significance for continuously interacting groups in Experiment 2, where groups were able to benefit from G-I transfer and differential weighting, descriptively groups were more accurate than nominal groups in both experiments. One reason why this comparison was not significant in Experiment 1 is the remarkably good estimation accuracy of the nominal groups in this experiment ($M = 32.05$, $SD = 17.46$). In contrast, the performance of nominal groups in the (in large parts similar) first experiment of Schultze et al. (2012), who found a significant superiority of interacting groups over nominal groups with the same task type, was notably lower ($M = 39.54$, $SD = 38.08$). Presumably, by chance, nominal groups in our first experiment might have consisted of individuals whose idiosyncratic biases cancelled each other out more frequently than in the experiment of Schultze et al. (2012).

The results regarding the occurrence of differential weighting differ between our two experiments. In Experiment 1, where simple averaging was a rather effective strategy due to the lack of a population bias and small remaining differences in individual accuracy, we replicated the results of Schultze et al. (2012) that groups do not outperform the average of their group members' contributions. However, the fact that we found no evidence for differential weighting in Experiment 1 does not mean that group members necessarily weighted their individual contributions equally. The lack of bias in participants' estimates only implies that differential weighting would not have improved the group judgments substantially and, therefore, would not be detectable with a performance-based assessment of the weighting strategy. Since the actual group performance was inferior to the average model descriptively, we can conclude that to the extent that groups engaged in differential weighting, they did not benefit from it in Experiment 1. In contrast, groups engaged in effective differential weighting in Experiment 2, which employed a task favoring weighting by expertise or accuracy over averaging, due to a strong population

bias (e.g., Davis-Stober, Budescu, Dana, & Broomell, 2014; Einhorn et al., 1977). This allowed interacting groups in Experiment 2 to outperform the simple average of their members' individual estimates. Taken together, our results suggest that groups can – to some degree – engage in rather functional weighting strategies. Our findings, thus, provide an interesting basis for systematic research on the weighting strategies groups employ – for example, by investigating how various task and group characteristics relevant to the effectiveness of differential weighting influence the choice of the weighting strategy and its impact on group performance.

Limitations and directions for future research

Although our experiments provide evidence for stable G-I transfer in quantitative estimation tasks, we should also mention some limitations. Despite the fact that we used two different tasks with quite different characteristics that consistently produced G-I transfer, we still cannot exclude the possibility that other types of estimation tasks might yield different results. In the tasks we used, metric errors were rather common and – at times – extreme, as indicated by systematic idiosyncratic biases of group members in both experiments. This is particularly evident in Experiment 2, though, where participants systematically overestimated the weights of the small objects by an average factor of three. Mapping errors, on the other hand, might have been less pronounced, because most participants presumably had a rough recollection of the geographical location of the EU's member countries (if not necessarily the location of the capital cities within the countries), allowing them to distinguish long distances from short ones. Likewise, they could tell that a small plastic comb weighed less than a small metal hammer. Hence, the tasks we used had a great potential for metric error reductions whereas it impeded the occurrence of mapping error reductions as a consequence of interacting with others. Admittedly, this presumed combination of relatively low mapping and high metric errors might not generalize to all estimation tasks. For example, forecasting tasks, like predicting the return on a capital investment, are mainly characterized by mapping errors. In this case, the previous and current values of the variable that is to be predicted constitute rather reliable reference values that minimize the individual metric error. In contrast, there are several factors that should predominantly affect the mapping knowledge component. For example, when predicting the future market rate of a certain stock, one has to learn general market trends, as well as the previous prosperity and future plans of certain companies to reduce one's mapping error. All of these knowledge components and cues might be transferable through group discussion, quite similar to the exchange of reference values. However, in this case, G-I transfer should take more time to emerge, and also more time to fully develop. Therefore, it is crucial to replicate our findings with different types of quantitative estimation tasks, preferably tasks with a high ecological validity like forecasting tasks, or even in a real world setting. In general, our findings should be replicated with tasks of different complexity and different estimation

biases to form an overall picture regarding the strength of G-I transfer on the one hand, and functional differential weighting strategies on the other hand, under different circumstances.

Furthermore, it remains an open question as to how groups manage to identify their group members' expertise or the accuracy of their judgments in order to know whom to learn from. Previous evidence regarding ad hoc groups' ability to recognize expertise is rather contradictory. On the one hand, some studies indicate that groups are capable of identifying their most capable members (e.g., Baumann & Bonner, 2004; Bonner et al., 2007; Henry et al., 1996; Libby, Trotman, & Zimmer, 1987). On the other hand, there is also evidence of groups failing to recognize the specific expertise of their members (e.g., Littlepage, Robinson, & Reddington, 1997, studies 1 and 2; Littlepage, Schmidt, Whisler, & Frost, 1995; Trotman, Yetton, & Zimmer, 1983). The fact that, in our experiments, the most capable members' performance remained largely stable, whereas the performance of the medium and least capable members considerably increased, speaks to the groups' ability to assess their members' expertise or at least the quality of their judgments. One possible determinant for the ability to recognize expertise might be the plausibility of individual estimations. As Yaniv and Kleinberger (2000) discuss, individuals might identify particularly poor estimates as out of the bounds of plausibility, even if people cannot generate correct estimates themselves. In others words, group members might have been reasonably good in realizing whom to ignore. This could also explain why there was no negative individual learning in our experiments. Nevertheless, further research should address the question of which circumstances facilitate the recognition of expertise or the accuracy of certain estimates, and what cues are relevant for groups to determine the relative expertise of their members.

Finally, we do not yet know whether group interaction is really *indispensable* to induce the phenomenon of G-I transfer. Since our results reveal strong learning effects after just one group interaction, this raises the question of whether similar processes might be possible even without any direct communication. As Farrell (2011) suggests, individual accuracy can be improved by knowing the estimates of other persons, without any form of group interaction (in terms of free information exchange). In other words, it is questionable whether discussing individual estimates with other people is crucial to individual learning, or whether the knowledge about others' judgments might be sufficient to achieve the same or at least a similar beneficial effect, at least in some tasks. Hence, a promising line of future research is to identify which factors are indispensable for individual learning effects and by which means group interaction might additionally strengthen these processes.

Conclusion

In accordance with the idea of G-I transfer, group members can learn relevant knowledge in quantitative estimation tasks by cooperatively working with others. One group interaction seems to be sufficient for an increase in metric knowledge that leads to more accurate individual judgments, whereas further group interaction does not foster additional capability gains. Furthermore, under certain circumstances, effective weighting strategies when combining those individual estimates with a group judgment might occur. Thus, we know now that a single group discussion can robustly improve group members' individual judgment accuracy and can also lead to an improved collaboration, although the specific mechanisms underlying these improvements are still an open topic for future research.

Data accessibility Statement

All the participant data and analysis scripts for experiment 1 and 2 can be found on this paper's project page on the Open Science Framework.

https://osf.io/edfqv/?view_only=6057215b0d2f40c38 3ca47f31e84d3b5.

Competing Interests

The authors have no competing interests to declare.

Authors Contributions

Contributed to conception and design: Alexander Stern, Thomas Schultze, Stefan Schulz-Hardt.

Contributed to acquisition of data: Alexander Stern.

Contributed to analysis and interpretation of data: Alexander Stern, Thomas Schultze, Stefan Schulz-Hardt.

Drafted and/or revised the article: Alexander Stern, Thomas Schultze, Stefan Schulz-Hardt.

Approved the submitted version for publication: Alexander Stern, Thomas Schultze, Stefan Schulz-Hardt.

Notes

[1] Furthermore, participants were asked to rate their estimation confidence on a six-point Likert scale after every judgment. However, because this confidence measure revealed no relevant effects for our research question we refrain from reporting any analyses regarding this measure. Nevertheless, we will be glad to publish our data for further analyses.

[2] The assignment to the three conditions did not produce accuracy differences between the experimental conditions, $F(2, 58) = 0.05$, $p = .953$, $\eta_p^2 < .01$, that is, the average MAPE-scores of the individual training phase (prior to the experimental manipulation) were virtually identical for the three conditions. The same

was true for the variance of individual accuracy within the composed (nominal) groups, $F(2, 58) = 0.73$, $p = .487$, $\eta_p^2 = .02$.

[3] For the continuous-interaction groups, this bonus was based on the accuracy of the group judgments. For single-interaction groups, it was based on the group judgment of the first trial in the second phase and the average of their individual estimates for the remaining nine trials. For nominal group members, the bonus depended solely on their individual accuracy during the second phase. However, participants were only aware of a general performance bonus and were not informed about the details of how this bonus was computed in each of the experimental conditions (to prevent that individual task motivation might differ between the experimental conditions as a consequence of differences in bonus computation).

[4] We calculated a paired-samples t-tests between participants' average individual accuracy on the ten trials of the practice phase and the first trial of the second phase for the two group conditions. This analysis revealed no differences at all ($M = 48.44$, $SD = 17.02$ vs. $M = 47.18$, $SD = 33.32$), $t(39) = 0.30$, $p = .762$, $d = 0.05$. This allows us to rule out an alternative explanation, namely that participants' anticipation of group interaction might have increased the accountability they felt for their estimates and – thereby – the accuracy of these estimates.

[5] We used LSD post hoc comparisons because we aimed for the highest possible statistical power when comparing the two interacting group conditions. More conservative post hoc comparisons might have disguised significant differences. As a consequence of the higher statistical power, the absence of significant differences between the two interacting group conditions is an even more informative finding.

[6] Beyond that, participants' gender did not significantly affect the dependent variable, $t(180) = 1.90$, $p = .059$, $d = 0.29$, although, descriptively, performance changes were somewhat more pronounced for women than for men. However, there was no association between experimental condition and participant gender, $\chi^2(2) = 0.33$, $p = .859$. In other words, female and male participants distributed equally across conditions. Moreover, there was no significant relation between participants' age and the dependent variable, $r < .01$, $p = .990$.

[7] Additionally, we split the 20 trials in four blocks (trial 1–6, 7–11, 12–16, 17–20) to analyze more evenly sized blocks of trials. Corresponding post hoc paired-samples t-tests only revealed significant changes in group members' individual accuracy right after the first group interaction in both the continuous-interaction condition ($M = 49.86$, $SD = 22.39$ vs. $M = 31.13$, $SD = 9.92$), $t(20) = 3.84$, $p = .001$, $d = 0.88$, and the single-interaction condition ($M = 50.67$, $SD = 18.37$ vs. $M = 31.18$, $SD = 8.88$), $t(18) = 4.65$, $p < .001$, $d = 1.09$ (all other ts < 1.58; all other ps > .130). Furthermore, there was no significant linear trend towards an individual accuracy decrease or improvement after the first

group interaction for the continuous-interaction nor the single-interaction condition (all Fs < 1.48, all ps > .240). These analyses further support the assumption of immediate performance enhancements right after the first group interaction with no additional improvements on subsequent trials.

[8] The pattern of results when adding group member as a within-subjects factor mirrors the general finding of stronger performance enhancements of inferior group members, such that metric error reductions were stronger for inferior group members in both experiments.

[9] Individuals cannot yet benefit from G-I transfer on the first trial of the second phase. The group judgments, however, can, lead to an "unfair" advantage over the averaged previous individual judgments, which, in turn, would distort the results regarding whether or not groups engage in differential weighting strategies.

[10] We conducted the same ANOVA with the four randomly created trial orders as an additional between-subjects factor. This analysis revealed no significant interaction of trial order and group type, $F(6, 72) = 0.46$, $p = .838$, $\eta_p^2 = .04$, indicating that differences between the three conditions cannot be attributed to differences in the order of trials. Therefore, we dropped the order of trials as a between-subjects factor in all analyses.

[11] As a closer inspection of the data showed that a substantial part of the descriptive difference between performance enhancements in the continuous-interaction and the single-interaction condition derives from one participant in the latter condition whose performance dramatically decreased by 670 percentage points. When excluding this participant's group, the difference between the two group conditions decreases markedly ($M = 87.33$, $SD = 62.90$ vs. $M = 60.12$, $SD = 84.08$), $p = .170$, although some moderate difference remains. Furthermore, the effect size of the post hoc t-tests against zero in the single-interaction condition increases, $t(27) = 3.72$, $p < .001$, $d = 0.71$.

[12] Again, participants' gender didn't significantly affect the individual performance change from phase one to two, $t(250) = 0.75$, $p = .456$, $d = 0.01$. Furthermore, there was no significant relation between participants' age and the dependent variable, $r = .11$, $p = .073$, although, descriptively, younger participants' performance increased more than that of older participants. Furthermore, the results revealed differences in participants' age between the conditions that fell short of significance, $F(2, 249) = 2.29$, $p = .057$, $\eta_p^2 = .02$. However, LSD post hoc comparisons showed that the participants were significantly older in the single-interaction condition than in the nominal-group condition, $p = .017$, whereas there were no differences between the continuous-group condition and the other two (all ps > .205). Accordingly, participants' age should, if at all, lead to somewhat smaller performance enhancements in the single-interaction condition than in the nominal-group condition and, thereby, work against our hypothesis. Hence, the weak relationship between performance changes and participants' age should not interfere with our results. Finally, there were no

differences between the three conditions concerning (nominal) group members' baseline accuracy, $F(2, 81) = 0.29$, $p = .751$, $\eta_p^2 < .01$, nor regarding the variance of accuracy within the groups, $F(2, 81) = 0.42$, $p = .658$, $\eta_p^2 = .01$.

[13] Again, we split the 20 trials in four blocks (trial 1–6, 7–11, 12–16, 17–20) to analyze more evenly sized blocks of trials. Post hoc paired-samples t-tests revealed significant changes in group members' individual accuracy right after the first group interaction in both the continuous-interaction condition ($M = 249.91$, $SD = 110.13$ vs. $M = 153.24$, $SD = 90.39$), $t(27) = 7.13$, $p < .001$, $d = 1.35$, and the single-interaction condition ($M = 237.55$, $SD = 125.16$ vs. $M = 179.27$, $SD = 103.94$), $t(27) = 3.31$, $p = .003$, $d = 0.63$ (all other ts < 1.32, all other ps $> .198$). Beyond that, neither the continuous-interaction nor the single-interaction condition showed a significant linear trend towards an individual accuracy decrease or improvement after the first group interaction (all Fs < 1.53, all ps $> .227$), which further supports the idea of a stable G-I transfer.

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. DOI: https://doi.org/10.1016/j.obhdp.2003.12.004

Bednarik, P., & **Schultze, T.** (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making, 10*, 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*, 46–66. DOI: https://doi.org/10.1002/ejsp.400

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*, 337–350. DOI: https://doi.org/10.1037/a0025566

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

Brodbeck, F., & **Greitemeyer, T.** (2000a). A dynamic model of group performance: Considering the group members' capacity to learn. *Group Processes & Intergroup Relations, 3*, 159–182. DOI: https://doi.org/10.1177/1368430200003002004

Brodbeck, F. C., & **Greitemeyer, T.** (2000b). Effects of individual versus mixed individual and group experience in rule induction on group member learning and group performance. *Journal of Experimental Social Psychology, 36*, 621–648. DOI: https://doi.org/10.1006/jesp.2000.1423

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In Ross, B. H. (Ed.), *Psychology of learning and motivation* (Vol. 41). New York: Academic Press, pp. 321–360. DOI: https://doi.org/10.1016/S0079-7421(02)80011-1

Brown, N. R., & **Siegler, R. S.** (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*, 511–534. DOI: https://doi.org/10.1037/0033-295X.100.3.511

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

Einhorn, H. J., Hogarth, R. M., & **Klempner, E.** (1977). Quality of group judgment. *Psychological Bulletin, 84*, 158–172. DOI: https://doi.org/10.1037/0033-2909.84.1.158

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences, 108*, E625. DOI: https://doi.org/10.1073/pnas.1109947108

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion confidence judgments. *Organizational Behavior and Human Decision Processes, 56*, 11–27. DOI: https://doi.org/10.1006/obhd.1993.1043

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes, 62*, 190–197. DOI: https://doi.org/10.1006/obhd.1995.1042

Henry, R. A., Strickland, O. J., Yorges, S. L., & **Ladd, D.** (1996). Helping groups determine their most accurate member: The role of outcome feedback. *Journal of Applied Social Psychology, 26*, 1153–1170. DOI: https://doi.org/10.1111/j.1559-1816.1996.tb02290.x

Laughlin, P. R., & **Barth, J. M.** (1981). Group-to-individual and individual-to-group problem-solving transfer. *Journal of Personality and Social Psychology, 41*, 1087–1093. DOI: https://doi.org/10.1037/0022-3514.41.6.1087

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

Laughlin, P. R., Carey, H. R., & **Kerr, N. L.** (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations, 11*, 319–330. DOI: https://doi.org/10.1177/1368430208090645

Laughlin, P. R., & **Ellis, A. L.** (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*, 177–189. DOI: https://doi.org/10.1016/0022-1031(86)90022-3

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. DOI: https://doi.org/10.1037/1089-2699.7.1.55

Laughlin, P. R., & **Jaccard, J. J.** (1975). Social facilitation and observational learning of individuals and cooperative

pairs. *Journal of Personality and Social Psychology, 32*, 873–879. DOI: https://doi.org/10.1037/0022-3514.32.5.873

Laughlin, P. R., & **Sweeney, J. D.** (1977). Individual-to-group and group-to-individual transfer in problem solving. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 246–254. DOI: https://doi.org/10.1037/0278-7393.3.2.246

Libby, R., Trotman, K. T., & **Zimmer, I.** (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology, 72*, 81–87. DOI: https://doi.org/10.1037/0021-9010.72.1.81

Littlepage, G., Robison, W., & **Reddington, K.** (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes, 69*, 133–147. DOI: https://doi.org/10.1006/obhd.1997.2677

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & **Frost, A. G.** (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology, 69*, 877–889. DOI: https://doi.org/10.1037/0022-3514.69.5.877

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. DOI: https://doi.org/10.1016/j.obhdp.2011.12.006

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

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

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*, 780–805. DOI: https://doi.org/10.1037/a0015145

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & **Davis, J. H.** (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*, 25–35. DOI: https://doi.org/10.2307/2786786

Trotman, K. T., Yetton, P. W., & **Zimmer, I. R.** (1983). Individual and group judgments of internal control systems. *Journal of Accounting Research, 21*, 286–292. DOI: https://doi.org/10.2307/2490948

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