ORIGINAL ARTICLE

Group decisions based on confidence weighted majority voting

Sascha Meyen[1]* , Dorothee M. B. Sigg[1], Ulrike von Luxburg[2,3] and Volker H. Franz[1]

Abstract

Background: It has repeatedly been reported that, when making decisions under uncertainty, groups outperform individuals. Real groups are often replaced by simulated groups: Instead of performing an actual group discussion, individual responses are aggregated by a numerical computation. While studies have typically used unweighted majority voting (MV) for this aggregation, the theoretically optimal method is confidence weighted majority voting (CWMV)—if independent and accurate confidence ratings from the individual group members are available. To determine which simulations (MV vs. CWMV) reflect real group processes better, we applied formal cognitive modeling and compared simulated group responses to real group responses.

Results: Simulated group decisions based on CWMV matched the accuracy of real group decisions, while simulated group decisions based on MV showed lower accuracy. CWMV predicted the confidence that groups put into their group decisions well. However, real groups treated individual votes to some extent more equally weighted than suggested by CWMV. Additionally, real groups tend to put lower confidence into their decisions compared to CWMV simulations.

Conclusion: Our results highlight the importance of taking individual confidences into account when simulating group decisions: We found that real groups can aggregate individual confidences in a way that matches statistical aggregations given by CWMV to some extent. This implies that research using simulated group decisions should use CWMV instead of MV as a benchmark to compare real groups to.

Keywords: Group discussion, Group decision, Confidence weighted majority vote, Wisdom of the crowd

Significance statement

The question of how a group determines an overall group decision from the individual votes of its group members is pervasive and likely as old as mankind. It is at the basis of democratic voting rules and is also prevalent with new urgency in the age of the Internet, where often many individual votes, or ratings, are available that one wants to combine to an optimal overall group decision—without there being the possibility of real group discussions. From a theoretical point of view, the situation is clear: Individual confidences should be taken into account and

confidence weighted majority voting (CWMV) is the statistically optimal aggregation procedure (under quite general assumptions). However, in research on group decisions, CWMV is not routinely used for comparison to real group performances, but instead the simpler majority vote (MV) that ignores the individual confidences. Therefore, it is currently not clear whether real groups weigh individual votes in the same way CWMV does. Real groups may be limited in their capacity to take individual confidence ratings into consideration or may rely on different strategies. We compared real group decision to simulated group decisions based on the CWMV and MV procedures. We found that real groups weigh individual confidences in a way that can be well described by CWMV. These results suggest that basic research as

*Correspondence: sascha.meyen@uni-tuebingen.de
[1] Experimental Cognitive Science, Department of Computer Science, University of Tübingen, Tübingen, Germany
Full list of author information is available at the end of the article

well as online-based aggregation of individual votes or ratings could benefit from using CWMV instead of MV.

Background

Under uncertainty, groups make more accurate decisions than individuals (Koriat 2015; Mannes et al. 2014): Medical students achieve more accurate diagnoses in groups than individually (Hautz et al. 2015); medical diagnoses improve when groups of independent doctors are involved (Kurvers et al. 2016; Wolf et al. 2015); groups of students make more accurate judgments about criminal cases than individuals (van Dijk et al. 2014); groups detect lies more accurately than individuals (Klein and Epley 2015); groups achieve higher IQ scores than individuals [referred to as wisdom of the crowd (Bachrach et al. 2012; Vercammen and Burgman 2019; Kosinski et al. 2012)], etc. Exceptions occur when group members have widely different levels of competence (Galesic et al. 2018; Puncochar and Fox 2004; van Dijk et al. 2014). Nevertheless, groups generally outperform individuals.

Although some of the above-mentioned studies also used real groups (Hautz et al. 2015; Klein and Epley 2015; van Dijk et al. 2014), all of these studies simulated group decisions: Individuals gave responses that were then statistically aggregated into one simulated group response without a real group discussion occurring. A crucial aspect is therefore the choice of aggregation method that is used to simulate group decisions. One frequently used method is majority voting (MV; Hastie and Kameda 2005; and see for example Klein and Epley 2015; van Dijk et al. 2014; Kosinski et al. 2012; Kurvers et al. 2016; Sorkin et al. 2001).

In MV, the most frequent individual decision (vote) is taken as the simulated group decision. By design, MV weighs all individual responses equally. Note, however, that real groups typically perform better than simulated groups using MV (Bahrami et al. 2010; Birnbaum and Diecidue 2015; Klein and Epley 2015; Sniezek and Henry 1989). This shows that MV cannot capture all the processes that are at work in real group decisions.

In particular, MV overlooks that individuals can estimate how accurate their own decisions are in many situations (Brenner et al. 1996; Fleming et al. 2012; Griffin and Tversky 1992; Martins 2006; Zehetleitner and Rausch 2013; Regenwetter et al. 2014) even though there are also situations in which they cannot (Klein and Epley 2015; Koriat 2012b, 2017; Litvinova et al. 2019). When reliable confidence estimates are available, they can influence real group discussions: It is plausible that individuals share their sense of confidence during group interactions (Bang et al. 2014) such that votes from confident individuals are weighted more than those of less confident individuals.

There are methods that have taken confidence ratings from individuals into account. One of the most prominent is the maximum confidence slating algorithm by Koriat (2012a, b). In this algorithm, the most confident individual decides the vote. Another approach for dealing with multiple confidence ratings is to not only consider the most confident individual but a small subgroup of the top most confident individuals (Mannes et al. 2014), or to average all confidences (Litvinova et al. 2020). However, these methods to simulate group decisions do not strictly follow the mathematically optimal way to aggregate confidences.

The theoretically optimal method to aggregate individual confidences is confidence weighted majority voting (CWMV; Grofman et al. 1983; Nitzan and Paroush 1982)—assuming that individuals can accurately assess confidences in their independently formed decisions. CWMV aggregates these independent responses (votes and confidences) in the mathematically optimal way by giving more weight to reliable than unreliable votes. Thus, statistically aggregating individual responses into a simulated group decision using CWMV rather than MV may reflect real groups better and provide a more appropriate benchmark.

Do real, interacting groups weigh individual confidences in a way that is reflected by simulating a group discussion using CWMV? It is not clear whether real group decisions are adequately represented by CWMV, since CWMV is only sporadically applied in the current research. Bahrami et al. (2010) found that group performance of dyads is well predicted by CWMV. Hautz et al. (2015) found that real dyads performed better than CWMV, which predicts the group response of a dyad to be that of the most confident member. CWMV is also discussed in animals from an evolutionary perspective (Marshall et al. 2017). However, to our knowledge, no study has yet considered groups with more than two members comparing decisions from real group discussions versus simulated decisions using CWMV on a trial-by-trial basis.

In our experiment, we investigated whether CWMV simulations can predict real group decision of triads (groups of three). We compared simulated group decisions to real group decisions on a trial-by-trial basis. Our groups consisted of three individuals because we wanted to investigate whether real groups weigh confidences in a way that is adequately reflected by CWMV. In contrast, using only dyads, CWMV simulates the group decisions to be the vote of the more confident individual (similar to maximum confidence slating) and CWMV can only contribute by predicting a dyad's combined confidence based on the individual responses. But triads can display qualitatively different behaviors than dyads: While it is

sometimes the case that the most confident individual determines the group decision in triads, triads also allow for the possibility that the most confident individual is overruled by the two other group members when they are sufficiently confident in the alternative choice. Thus, we want to clarify whether real groups of three weigh individual votes in a way that can be characterized by CWMV.

Before describing our experiment, we will give a more formal description of the simulation methods MV and CWMV. We will present a formal cognitive model (e.g., see Forstmann et al. 2011) that allows us to measure in how far real groups deviate from CWMV simulations.

Majority voting (MV) versus confidence weighted majority voting (CWMV)

CWMV assumes that multiple individuals report independent decisions (votes) as well as confidence ratings. These confidence ratings indicate how reliable individual decisions are. CWMV weighs the decisions by the confidence ratings in a theoretically optimal way to simulate a group decision (Grofman et al. 1983; Nitzan and Paroush 1982). This section shortly introduces the basic mathematical notation, first of MV and then of CWMV.

Let a group consist of n individuals. The task is to decide between multiple (usually two) options from which exactly one is correct. For example, consider $n = 3$ students trying to determine whether a suspect of a criminal case is guilty or not (cf. van Dijk et al. 2014). First, each individual forms a decision y_i which is either $+1$ (not guilty) or -1 (guilty). Second, in a real, interactive group discussion, the individual group members reach a common decision y_g.

The real-world group decision y_g can be simulated by statistically aggregating the independently formed individual responses. MV simulates the group decision to be that of the majority of individuals, $y_g^{MV} = \text{sign}(\sum_{i=1}^{n} y_i)$. MV (as well as CWMV) assumes that individual responses are independent from each other given the ground truth, that is, individuals must form their decision only based on material that is not systematically shared between members. To illustrate a violation of this assumption, consider as another example a group of radiologists forming their individual diagnoses based on one and the same X-ray. They will not come to fully independent conclusions about the true state of the patient's condition because their opinions will be commonly influenced by the quality of the X-ray. In the worst case, multiple individual responses are fully dependent offering no more information than one single response. In our experiment, independence will be ensured by design in order to study CWMV—even though many real-world situations will not allow for such a controlled environment.

When individuals report confidence ratings, c_i, MV can be improved upon by using CWMV instead. These confidence ratings are assumed to be in the form of estimates for the probability of their decision being correct, $c_i = P(y_i \text{ is correct})$. In some situations, individuals can make such estimates (Griffin and Tversky 1992; Martins 2006; Regenwetter et al. 2014; Koriat 2012a) and, under specific circumstances, assessing confidences is essentially the same as estimating the relative frequency of being correct (Brenner et al. 1996; Pouget et al. 2016). CWMV transforms these confidences into optimal weights, which are the logarithmic odds (log odds), $w_i = \log(c_i/(1 - c_i))$. See Nitzan and Paroush (1982) as well as Shapley and Grofman (1984) and find an intuitive account for using logarithmic odds as weights at the end of this section. Using these weights, CWMV simulates the group decision by

$$y_g^{\text{CWMV}} = \text{sign}\left(\sum_{i=1}^{n} w_i y_i\right). \qquad (1)$$

Similar to the individual confidence ratings, real groups can also report how confident they are in their group decision c_g. CWMV can simulate these group confidences based on the individual confidences by

$$c_g^{\text{CWMV}} = \frac{1}{1 + \exp(-|\sum_{i=1}^{n} w_i y_i|)}. \qquad (2)$$

To illustrate the computation of CWMV, consider again the three students deciding whether a suspect is guilty. Say, Student 1 votes for the suspect being innocent, $y_1 = +1$, but Students 2 and 3 believe the suspect to be guilty, $y_2 = -1$ and $y_3 = -1$. Aggregating these decisions using MV determines the simulated group decision to be guilty, $y_g^{MV} = \text{sign}((+1) + (-1) + (-1)) = -1$. Additionally, Student 1 reports being quite confident in their vote such that the probability of their judgment being correct is 76%, $c_1 = 0.76$. In contrast, Students 2 and 3 are very unsure with a confidence of only 51%, $c_2 = c_3 = 0.51$. Using CWMV to integrate these individual responses into a simulated group decision, the individual confidences are first transformed into weights with Student 1 having a higher confidence and, thus, a larger weight: $w_1 = \log(0.76/0.24) = 1.15$ versus $w_2 = w_3 = \log(0.51/0.49) = 0.04$. Then, CWMV leads to a different simulated group decision than MV finding the suspect not guilty, $y_g^{\text{CWMV}} = \text{sign}((+1.15) + (-0.04) + (-0.04))$ $= \text{sign}(+1.07) = +1$. Moreover, CWMV simulates the group's confidence in their verdict to be 75%, $c_g^{\text{CWMV}} = 1/[1 + \exp(-|(+1.07)|)] = 0.75$. That is, the confident response from Student 1 is only slightly attenuated by the unconfident, opposing responses from

Students 2 and 3 as might be realistic in a real group discussion. This example corresponds numerically to Scenario II from our experiment, which we use to study in how far real groups are better represented by MV or CWMV simulations, see Table 1.

A technical note: The weights in CWMV are log odds because the logarithm is used as a convenient trick to transform a multiplication into a weighted sum. When computing the probabilities of a suspect being guilty or not, the basic probability theory gives that odds ($o_i = c_i/(1 - c_i)$) can be multiplied. In our example, the odds of the suspect being innocent are $o_1 = 0.76/0.24$ and $o_2 = o_3 = 0.51/0.49$. Multiplying these odds results in group odds, $o_g = o_1 \times o_2^{-1} \times o_3^{-1} = 3$ (o_2 and o_3 are inverted because Student 2 and 3 vote for guilty). Observe that the group odds are indeed equivalent to the 75% group confidence computed by CWMV from above, $c_g/(1 - c_g) = 0.75/0.25 = 3$. By applying the laws of logarithm, multiplication of these odds is transformed into a sum of the log odds: $\log(o_1 \times o_2^{-1} \times o_3^{-1}) = (+\log(o_1)) + (-\log(o_2)) + (-\log(o_3))$, which allows to derive Eqs. 1 and 2. Note further that, when an individual is absolutely certain in their decision ($c_i = 0$ or $c_i = 1$), the odds o_i and weights w_i are undefined. In this case, by convention, the simulated group is set to be absolutely certain as well ($c_g = 0$ or $c_g = 1$). But if two participants came to opposite decisions and were both absolutely certain, by convention, their two responses would be discarded and the third individual's vote would decide (this situation did not occur in our experiment).

Given this formal framework of CWMV, the purpose of this study is to investigate how well individual responses (y_i and c_i) aggregated into simulated group responses (y_g^{CWMV} and c_g^{CWMV}) represent the real group responses from actual group discussions (y_g and c_g) on a trial-by-trial basis. We will modify Eqs. 1 and 2 using formal cognitive modeling in order to characterize how real groups deviate from these CWMV simulations.

Methods

Participants

A total of 21 participants (11 females, mean age = 21.4, range = 19–26) completed the experiment in seven groups of three. All were students who received either course credit for 30 min of participation or payment (4 EUR, equivalent to 4.5 USD). All participants had normal or corrected-to-normal vision and provided written informed consent prior to participation.

Stimuli and procedure

We adopted a procedure that has been established by Griffin and Tversky (1992) and extended it to a group setting. The experiment consisted of three practice trials followed by 12 experimental trials. Each trial consisted of an individual phase and a group phase, see Fig. 1. Participants viewed rapid stimulus sequences consisting of 11 to 13 red and blue disks. Their task was to guess whether the stimulus sequence was generated by a fair coin (producing in expectation 50% red and 50% blue disks) or a biased coin (producing 60% red and 40% blue disks). Participants were instructed that both, the fair and the biased coin, are a priori equally likely. Griffin and Tversky (1992) showed that, in this task, participants' individual confidence ratings are well calibrated.
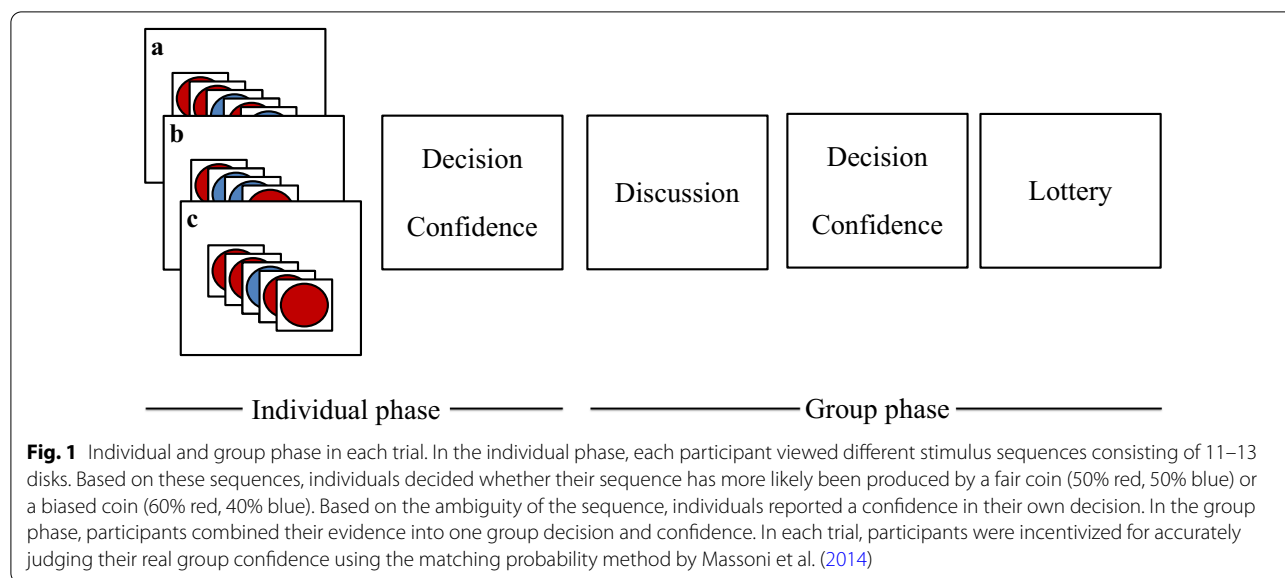
Participants viewed different stimulus sequences simultaneously at individual laptops. Their viewing distance to the screen was approximately 60 cm. Each disk was presented for 100 ms with a diameter of 2.2 cm corresponding to a viewing angle of 2.1°. Disks were intermitted by a 100-ms blank interval creating the impression of a rapid stream. This presentation prevented participants from performing explicit mathematical calculations so that they could only obtain an intuitive sense of confidence.

Depending on which coin better matched the stimulus sequence, participants made a decision for either the fair or the biased coin. Some stimulus sequences were more ambiguous than others providing different

Table 1 Ideal decisions and confidences

Scenario	Individual						Group	
	A		B		C			
	y_1^*	c_1^* (%)	y_2^*	c_2^* (%)	y_3^*	c_3^* (%)	y_g^*	c_g^* (%)
I	Fair coin	87	Fair coin	70	Fair coin	62	Fair coin	96
II	Biased coin	76	Fair coin	51	Fair coin	51	Biased coin	75
III	Biased coin	88	Biased coin	54	Fair coin	81	Biased coin	66
IV	Fair coin	81	Biased coin	58	Biased coin	72	Fair coin	54

In each trial, we applied one out of four scenarios (I–IV), which is defined by three stimulus sequences (A, B and C). Each of the three participants from a group viewed one stimulus sequence. Each individual stimulus sequence entails an ideal decision y_i^* and ideal confidence c_i^* that can be derived from probability computations. The ideal individual responses from each scenario determine the groups' ideal decision y_g^* and confidence c_g^* (see "Methods" section for an example calculation corresponding to Scenario II)

Meyen *et al. Cogn. Research* (2021) 6:18

Page 5 of 13

Fig. 1 Individual and group phase in each trial. In the individual phase, each participant viewed different stimulus sequences consisting of 11–13 disks. Based on these sequences, individuals decided whether their sequence has more likely been produced by a fair coin (50% red, 50% blue) or a biased coin (60% red, 40% blue). Based on the ambiguity of the sequence, individuals reported a confidence in their own decision. In the group phase, participants combined their evidence into one group decision and confidence. In each trial, participants were incentivized for accurately judging their real group confidence using the matching probability method by Massoni et al. (2014)

levels of confidence that participants reported on a visual analog scale from 50% ("I am completely unsure. The other option is equally likely.") to 100% ("I am completely sure. My decision is definitely correct."). Participants were instructed to report their subjective probability with which they believed their decision to be correct. In contrast to verbal scales, where participants give responses such as "somewhat likely" or "almost certain," this numeric scale is necessary because numeric values (here, c_i) are required to simulate group decisions (see Eq. 1). However, it is noteworthy that a numerical scale can prompt participants to engage in formal thinking, which they would otherwise have not (Windschitl and Wells 1996).

The presented stimulus sequences determined ideal individual responses, which reflect posterior probabilities that can be computed using probability theory. Table 1 shows which responses the stimulus sequences would produce if participants were ideal observers. For example, assume that a participant saw the disk sequence red, red, blue, red and red. A fair coin would have produced such a sequence with a likelihood of $p_{\text{fair}} = 0.5^5 = 3\%$ and the biased coin with $p_{\text{biased}} = 0.6^4 \cdot 0.4^1 = 5\%$. Because the biased coin was more likely to produce this stimulus sequence, the ideal decision is for the biased coin denoted by $y_i^* = +1$ (the asterisk denotes ideal values). The ideal confidence was $c_i^* = p_{\text{biased}}/(p_{\text{fair}} + p_{\text{biased}}) = 5\%/(3\% + 5\%) = 62\%$.

Note that scheduling individual reports before a group discussion (as in our experiment) improves group performance and prevents contamination of individual reports by the group decision (Sniezek and Henry 1990). That is, individual reports remain independent because

participants interacted only after they gave their individual responses.

After the individual phase, participants entered the group phase. Since participants had viewed different stimulus sequences that were produced by the same coin, they engaged in a group discussion to aggregate the individually gathered evidence and produce a real group response. Similar to the individual responses, groups reported a decision and rated their confidence in that decision. We label these responses based on real group discussions *reported group decision* and *reported group confidence* and later compare them to the *simulated group decision* and *simulated group confidence*, which we obtain from statistically aggregating individual responses using CWMV. Groups were allowed to give a group response not earlier than 30 s and discussions usually did not last longer than 2 min.

The ideal group responses, y_g^* and c_g^*, can be determined by adding the number of red and blue disks from all three stimulus sequences shown to the participants. Then, the same calculations as for ideal individual responses can be applied to compute the ideal group responses. Alternatively and equivalently, aggregating ideal individual responses using CWMV (Eqs. 1, 2) also produces the ideal group responses because CWMV aggregates confidences in the theoretically correct way.

Across the 12 experimental trials, there were four Scenarios I–IV. Each scenario was defined by three stimulus sequences: A, B and C. Table 1 shows the ideal decision and confidence for each stimulus sequence in each scenario as well as the ideal group responses. Each participant saw one of those sequences from the current scenario. These scenarios were repeated three times in a

Meyen *et al. Cogn. Research* (2021) 6:18

Page 6 of 13

randomized order for a total of 12 trials, and the stimulus sequence that participants saw (A, B and C) was rotated so that participants viewed different stimulus sequences when a scenario was repeated. Importantly, Scenarios II and IV were designed so that MV and CWMV yield different predictions because the most confident individual should—according to CWMV—outweigh the relatively unconfident majority.

At the end of the group phase in each trial, the group was incentivized for giving an accurate group confidence rating. They entered a lottery in which the group could win money depending on how accurate the reported group confidence was. This lottery, the matching probability method, was conceived by Massoni et al. (2014), see also Dienes and Seth (2010). The probability to win in this lottery is maximized if the group confidence is neither under- nor overestimated. Participants were instructed about the rules of this lottery, and it was emphasized that chances to win are best when confidence ratings reflect the probability of the group decision to be correct. In each trial, groups could win 0.60 EUR (approximately 0.66 USD). Across 12 experimental trials, groups could win a total maximum of 7.20 EUR (7.90 USD) in addition to their compensation for participation. The sum was split equally among the three participants of the group. We did not apply this lottery for individual confidence ratings because these have already been shown to be reliable (Griffin and Tversky 1992) so that incentivation was not necessary in the individual phase. In contrast, incentivation was applied in the group phase because we assumed that it is important to additionally motivate participants there and keep them engaged in the group discussions.

Formal cognitive modeling of CWMV

CWMV is the theoretically optimal way of aggregating individual responses. Real groups on the other hand may deviate from CWMV in various ways. To measure these deviations, we introduce four parameters into the CWMV framework in order to capture different aspects in which real groups deviate from CWMV:

- σ_i: *precision of individuals* in recovering the ideal confidence in their reported confidence ratings,
- β: *equality effect*, or, tendency of groups to weigh individual votes more equal or more extreme than CWMV would based on the individual confidences,
- γ: *group confidence effect* determining whether groups tend to over- or underestimate their confidences, and
- σ_g: *precision of groups* in determining the group confidence in accordance with CWMV simulations based on the individual confidence ratings.

We estimate individuals' precision, σ_i, in recovering the true strength of evidence of the displayed stimuli sequences. We assume that individuals are not able to determine the ideal confidence but, instead, their actual responses will scatter around the ideal values. We describe this by an error term ϵ_i:

$$c_i = c_i^* + \epsilon_i.$$

This error term ϵ_i is normally distributed with mean zero (reflecting no absolute bias in individual confidence reports in accordance with Griffin and Tversky 1992) and standard deviation σ_i. This standard deviation characterizes individuals' precision in recovering the true confidence. An ideal observer would be perfectly precise and make no errors, $\sigma_i = 0$, whereas larger values of σ_i indicate less precision.

Individuals make incorrect decisions if the actual confidence deviates below the 50% threshold resulting in the complementary confidence toward the incorrect decision.

$$y_i = \begin{cases} y_i^* \text{ (correct)}, & \text{if } c_i \geq 0.5. \\ -y_i^* \text{ (incorrect)}, & \text{otherwise.} \end{cases}$$

In our experiment, we used a half scale ranging from 50 to 100% toward the decision made by the participant. For correct estimation, we transform the reported confidences into a full scale ranging from 0 to 100% toward the correct decision (see Olsson 2014) by inverting confidences toward the incorrect decision. For example, when an individual responded incorrectly with a confidence of 60%, we transform the confidence to $c_i = 0.4$ (40% toward the correct alternative) in order to estimate ϵ_i in each trial and thereupon σ_i.

Note that confidence ratings cannot be higher than 100%, which potentially causes a ceiling effect (Griffin and Brenner 2004). However, in our experiment, ideal confidences for individual responses only range up to a maximum of 88% (Scenario III, Individual A in Table 1) so that there is enough room for positive deviations, ϵ_i, to avoid a large ceiling effect here.

Furthermore, we introduce the parameter β to estimate the equality effect capturing whether real groups weighted individual responses in a way that deviates from CWMV. This parameter acts upon the weights w_i as an exponent:

$$y_g^{\text{CWMV}}(\beta) = \text{sign}\left(\sum_i w_i^\beta y_i\right). \tag{3}$$

As the name suggests, the equality effect models groups assigning more equalized weights than naive CWMV, which is conceptually similar to the approach by

Mahmoodi et al. (2015), but our model is technically different because we incorporate it into the CWMV framework. Here, the equality effect can vary between zero and infinity, $\beta \in [0; \infty]$. In the edge case of $\beta = 0$, every weight would be transformed equally to $w_i^0 = 1$ producing the special case of (unweighted) MV. On the other hand, $\beta = 1$ would leave weights unchanged, $w_i^1 = w_i$, and would produce undistorted CWMV. Values in between, $0 < \beta < 1$, would represent some compromise in which individual confidences are considered to some extent, but groups tend to equalize those weights. On the other side of the spectrum, larger values of $\beta > 1$ would represent an exaggeration of differences between weights so that the most confident individual's vote has a disproportionately large impact. In such situations, the most confident individual would tend to decide the vote single-handedly, which is equivalent to the predictions from maximum-confidence slating (Koriat 2012b).

We additionally estimate whether groups under- or overestimate their group confidence, which is captured in the group confidence effect γ:

$$c_g^{\mathrm{CWMV}}(\beta, \gamma) = \frac{1}{1 + \exp(-\gamma |\sum_i w_i^\beta y_i|)}. \tag{4}$$

The group confidence effect allows for a nonlinear scaling of the group confidences. This parameter can also vary between zero and infinity, $\gamma \in [0; \infty]$, where $\gamma < 1$ represents groups underestimating their confidence relative to the ideal statistical aggregation of individual responses, whereas $\gamma > 1$ represents an overestimation of group confidences. The special case of $\gamma = 1$ recovers undistorted (naive) CWMV.

Note that the equality effect β modifies individual weights and can potentially change the simulated group decision. In contrast, the group confidence effect γ only modifies a group's final confidence. (Hence, it does not appear in Eq. 3 where the simulated group decision is determined.) These two parameters capture deviations from naive CWMV simulations in a descriptive manner. For cautionary accounts against normative interpretations, see Gigerenzer (2018), Le Mens and Denrell (2011) and Neth et al. (2016).

Finally, we introduce an error term to the group confidence $c_g = c_g^{\mathrm{CWMV}}(\beta, \gamma) + \epsilon_g$. This error term ϵ_g acts similar to the error term of individual confidence ratings. It is normally distributed with mean zero and standard deviation σ_g, where smaller values indicate higher precision of the group discussion process matching the ideal aggregation.

For estimation, the individual precision σ_i was measured by computing the average of sample variances across individuals and taking the square root. For the group

parameters, we performed a grid search in which we varied β and γ in [0, 2] and σ_g in [0, 0.3] (larger values produced worse fits) with step sizes of 0.01. For each group, we chose the parameter combination that produced the maximum likelihood for the observed data using Eqs. 3 and 4 to predict the real group responses.

We validated this approach by conducting multiple parameter recovery simulations as suggested by Wilson and Collins (2019): We simulated data based on our model for fixed values of σ_i, β, γ and σ_g and demonstrated that our estimation approach recovered the ground truth parameters, see open material for details.

Results

We compared the average and median performance of real versus simulated groups, see Fig. 2 and see Additional file 1: Table S1 for estimates of each group. Real groups reported the correct (ideal) decision in 76.2% (SEM = 3.4%) of the trials (Mdn = 75.0%, IQR = 75.0−83.3). CWMV adequately simulated the average performance of real groups with 76.2% (SEM = 2.8%, Mdn = 75.0%, IQR = 70.8−83.3). In contrast, simulating group decisions using unweighted MV produced a lower accuracy of 66.7% (SEM = 3.6%, Mdn = 66.7%, IQR = 62.5−75.0) compared to CWMV with a mean difference of $M = 9.5\%$ (SEM = 3.4%), $t(6) = 2.83$, $p = 0.030$. Comparing MV to real groups yielded a trend toward the same difference, $M = 9.5\%$ (SEM = 4.6%), $t(6) = 2.07$, $p = 0.084$. We conducted two-sided, exact binomial tests to confirm this pattern: MV simulations were less accurate than CWMV simulations ($p = 0.016$) and real group decisions ($p = 0.016$).

Real versus ideal responses

Individual confidence ratings were well aligned with the ideal confidences, see Fig. 3a. The average correlation between reported versus ideal confidences across individuals was $\bar{r} = 0.73$, 95% CI [0.64, 0.80]. (We used Fisher's z-transformation for combining correlations into averages.) This finding replicates Griffin and Tversky (1992), showing that individual participants are able to evaluate the ambiguity in the presented stimulus sequences and report their confidences in form of subjective probabilities. Estimating the precision of individuals, we observed that reported confidences scattered around ideal confidences with a standard deviation of $\sigma_i = 13.3\%$, $SD = 6.6$, 95% CI [9.8, 16].

However, confidence reports showed systematic deviations. In hard (difficult) trials with low ideal confidences, individuals overestimated those confidences. This is reflected in regression lines on average being at $M = 55\%$, 95% CI [50.2, 58.8], where they should be at 50%. Additionally, high confidences were underestimated.
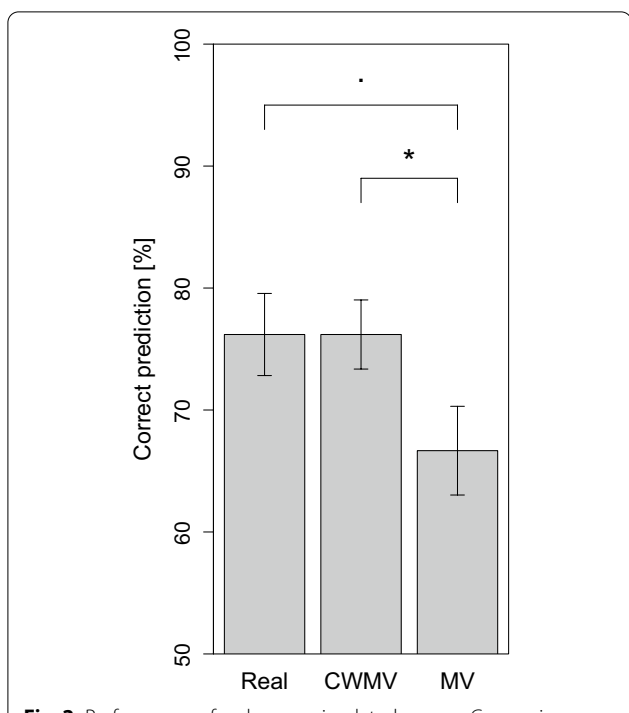
Fig. 2 Performance of real versus simulated groups. Comparing the accuracy of real group decisions to simulated group decisions using either CWMV or MV for aggregation of the individual decisions and confidences. Simulated groups based on CWMV predict the performance of real groups very well, while simulated groups based on MV underestimate the performance of real groups. Error bars indicate standard errors of the mean computed across groups. *$p < 0.05$. ˙$p < 0.1$

The average slope of regression lines was lower than the ideal value 1, $\bar{b} = 0.78$, 95%CI [0.61, 0.95]. A slope of 1 would have indicated that ideal and reported confidences increased equally, whereas, here, the observed slope below 1 indicated that increasing the true evidence strength from the presented stimulus sequences only led to a diminished increase in confidence.

Group confidence ratings showed a somewhat similar pattern, see Fig. 3b. (We again present median values.) The average correlation between reported and ideal group confidences was high, $\bar{r} = 0.71$, 95% CI [0.57, 0.80], Mdn = 0.71, IQR = 0.64−.79, but there was a relatively large root-mean-squared error, RMSE = 0.16. Real groups did not deviate from ideal values at low confidences: The regression lines at the ideal 50% were $M = 47\%$, 95% CI [40.7, 53.7], Mdn = 48.6%, IQR = 42.5−50.6. Nevertheless, groups (similar to individual participants) underestimated high confidences, resulting in an attenuated average slope relative to the ideal value of 1, $\bar{b} = 0.79$, 95% CI [0.58, 0.99], Mdn = 0.77, IQR = 0.73−0.96. The large RMSE reflects this divergence for high confidences. Exact binomial tests confirmed these results: All groups had a correlation above 0 and a slope below 1, both $p = 0.016$, but intercepts scattered around 50%, $p = 0.336$. Note that we avoid common problems of regression in the context of over- versus underconfidence estimation since our regressions use the fixed ideal confidences as independent variables (*x*-axes in Fig. 3), which exhibit no estimation error that would otherwise have lead to a biased analysis (Fiedler and Krueger 2012; Olsson 2014).
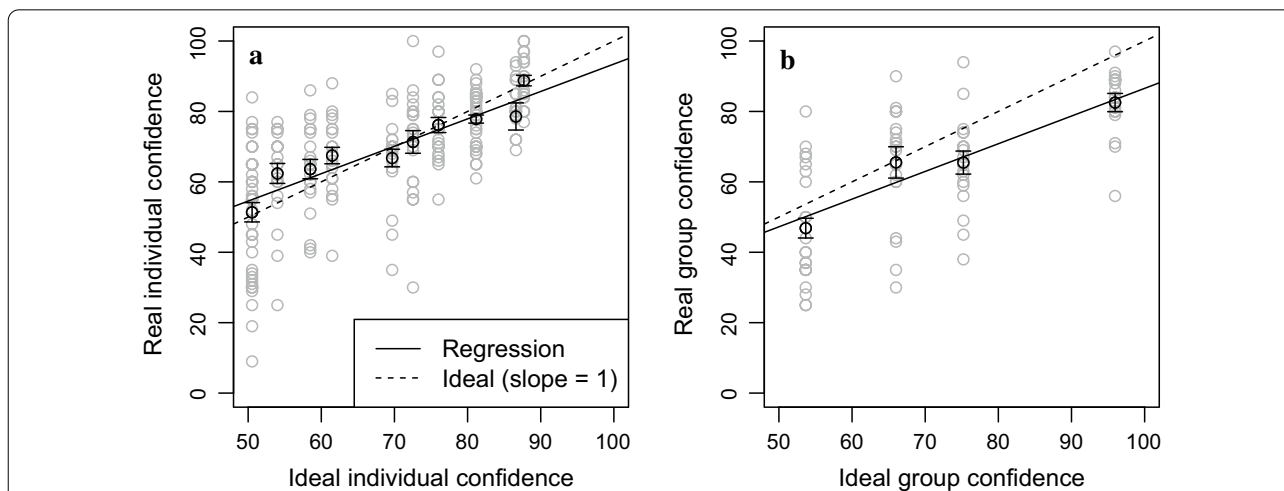

Fig. 3 Comparing ideal versus reported confidences from individuals and groups. Ideal confidence (*x*-axis) ranges from 50 to 100% in accordance with Table 1. In contrast, reported confidences (*y*-axis) range from 0 to 100% because we flipped confidence ratings in cases where an incorrect decision was given (e.g., a reported confidence of 60% toward the incorrect decision is displayed as a confidence of 40% here). In **a**, reported confidences from individuals (*y*-axis) are compared to the ideal values (*x*-axis; cf. c_1^*, c_2^*, and c_3^* from Table 1). Similarly in **b**, reported confidences from groups (*y*-axis) are compared to the ideal values (*x*-axis; cf. c_g^* from Table 1). Black points indicate mean values—averaged across individuals in **a** and across groups in **b**—for each ideal value. Grey points indicate single trial responses. Error bars indicate standard errors of the mean

Meyen *et al. Cogn. Research* (2021) 6:18

Page 9 of 13

Real versus simulated group responses

Responses from real, interacting groups were well predicted by simulated responses using CWMV. Naive CWMV (Eqs. 1, 2) produced an average correlation between reported and simulated confidences of $\bar{r} = 0.83$, 95% CI [0.56, 0.94], Mdn = 0.82, IQR = 0.64−.92. Despite the high correlation, there was still a large discrepancy, RMSE = 0.17, reflecting deviations of real responses from naive CWMV, see Fig. 4a.

We applied our formal cognitive model to estimate in how far real groups deviated from naive CWMV. The equality effect β was on average $M = 0.67$, $SD = 0.30$, 95% CI [0.38, 0.95], Mdn = 0.74, IQR = 0.38−0.95. This indicates that groups used confidences similar to CWMV but tended toward equalizing those weights. Votes from confident individuals were given more impact on the group decision compared to unconfident individuals but not to the extent suggested by CWMV. We observed trials in which the most confident individual is overruled by the majority and the tipping point at which this happened was earlier than what naive CWMV simulations predict. This observation is captured in the equality effect estimate being smaller than 1, $\beta = 0.67 < 1$.

It is noteworthy that, since β estimates are always larger than zero, it is a priori expected to obtain an above zero average simply due to random errors. To account for this, we performed a randomization test where we randomly permutated individual confidences and estimated β from the resulting data set. Since the confidences in these randomized data sets are not indicative of the group's decision, the true equality effect is zero here. From 1000 of such randomizations, we found that 95% of the obtained β estimates were below 0.4. This confirms that groups in our experiment (with $\beta = 0.67$) did take confidences into account ($\beta > 0$) but only to an attenuated extent ($\beta < 1$).

The group confidence effect γ was on average $M = 0.53$, $SD = 0.09$, 95% CI [0.45, 0.61], Mdn = 0.62, IQR = 0.55−0.74, indicating that real groups tend to underestimate ($\gamma < 1$) their confidence compared to CWMV simulations based on the individual responses. In Fig. 4b, this underestimation effect corresponds to a predicted curve (solid line) below the ideal values (dashed line).

The average group precision was $\sigma_g = 11\%$ (root mean square; $SD = 4$) with Mdn = 10%, IQR = 7−12%. This precision of group confidences is comparable to the precision of individual confidences.

The adapted CWMV model using $\beta = 0.67$ and $\gamma = 0.53$ predicted confidences that are correlated with reported confidences to the same degree as naive CWMV, $\bar{r} = 0.84$, 95% CI [0.68, 0.93], Mdn = 0.84, IQR = 0.72−.92. But in absolute terms, this adapted CWMV model matched the reported confidences better (RMSE = 11%) than naive CWMV (RMSE = 17%, mentioned above), $t(6) = 5.24$, $p < 0.002$, because the
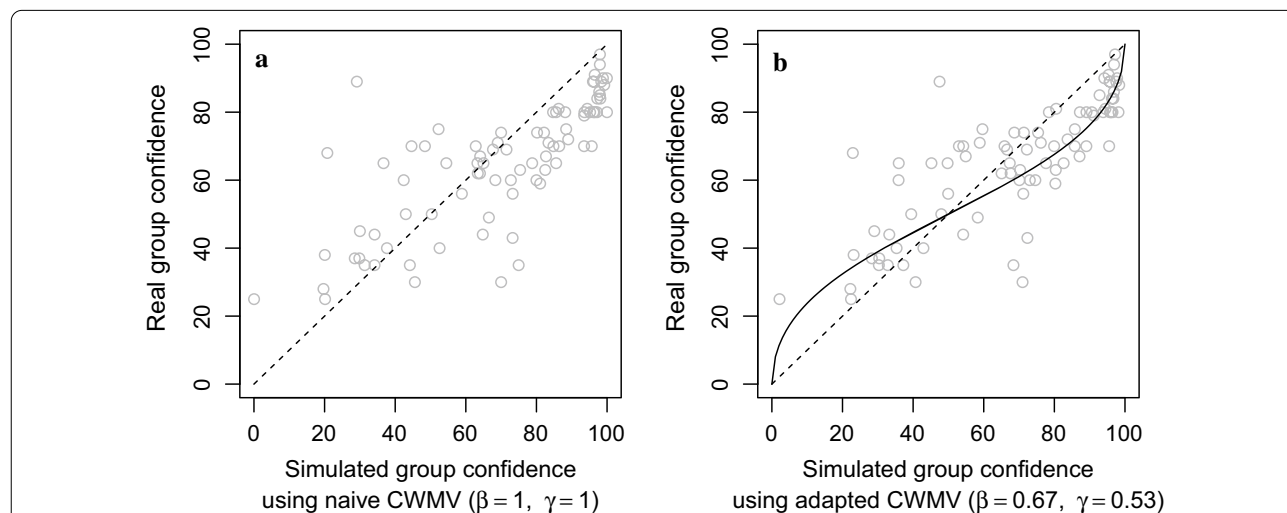

Fig. 4 Comparing real versus simulated group responses from statistically aggregating individual responses. We used individual responses to simulate group confidences via CWMV (*x*-axis). These simulations predict responses from real, interacting groups (*y*-axis). In **a**, we used naive CWMV as in Eqs. 1 and 2. The dashed line represents predictions from naive CWMV. This is equivalent to our formal cognitive modeling with equality effect $\beta = 1$ and group confidence effect $\gamma = 1$. In **b**, we estimated the equality effect, $\beta = 0.67$, and group confidence effect, $\gamma = 0.53$, see Eqs. 3 and 4. This model predicts real group responses (solid line) but incorporates the fact that real groups treated individual votes more equal and displayed an underconfidence effect. In both subfigures, confidence ratings are inverted for incorrect responses. For example, the point (34%, 35%) in **a** corresponds to a trial with a simulated confidence of 66% and a reported confidence of 65% with both decisions being the same but incorrect; hence, both confidences were inverted

adapted model simulates group responses with an equality and underconfidence effect.

Note that going from Fig. 4a to Fig. 4b, points are shifted along the *x*-axis because the equality effect $\beta = 0.67$ changes simulated confidences and can even change the simulated decision (points crossing the 50% border in the *x* direction). The extent of these shifts depends on the exact constellation of individual confidences. On the other hand, the group confidence effect $\gamma = 0.53$ only maps the resulting, simulated confidences in a nonlinear way to the reported confidences (solid curved line). This parameter reflects that groups were less confident in their decisions than what naive CWMV predicted.

Model comparison of group response simulations

To evaluate our adapted CWMV model, we compared the full model to three special case models in which we fixed one parameter at a time (first $\gamma = 1$, second $\beta = 0$ and third $\beta = 1$). For this comparison, we computed the Bayesian information criterion (BIC; Schwarz 1978). Using the Akaike information criterion (Akaike 1973) instead of BIC yielded qualitatively identical results. Smaller BIC values indicate a better fit relative to the number of parameters in the model. For the full model, the total score (sum across groups) was $BIC_{full} = -101$.

As a first comparison, we pitch the full model against a model that fixes the group confidence effect $\gamma = 1$ but keeps the equality effect β free. This model assumes that groups may only deviate from naive CWMV in terms of how they assign weights to the individual votes but exhibit no general over- or underconfidence. Here, the total score was $BIC_{\gamma=1} = -59$ indicating a worse fit as compared to the full model. The Bayes factor resulting from the BIC scores of the two models (e.g., see Farrell and Lewandowsky 2018, Chapter 11) clearly supported the full model, $BF_{full/\gamma=1} > 1000$. This supports the notion that group responses are best characterized by an overall underconfidence effect.

The second comparison fixes $\beta = 0$ but keeps γ free. This model is equivalent to an (unweighted) MV with group confidence effect. Here, the total score was $BIC_{\beta=0} = -63$, again supporting the full model, $BF_{full/\beta=0} > 1000$. This indicates that participants incorporate confidence ratings in the group discussion.

For the third comparison, we fix $\beta = 1$: This model assumes that real groups weigh individual votes exactly according to undistorted CWMV but still allows for an overall confidence effect of the group since γ is free. This model was on par with the full model, $BIC_{\beta=1} = -102$, with an inconclusive Bayes factor, $BF_{full/\beta=1} = 0.71$. This indicates that, according to the BIC criterion, fixing $\beta = 1$ did not perform worse (when accounting for the additional free parameter) than the full model, which keeps β free. On the other hand, when performing a model fit comparison irrespective of the number of parameters (Farrell and Lewandowsky 2018, Chapter 10), the full model performs better than that with fixed $\beta = 1$, $\chi^2(7) = 16.9$, $p = 0.018$. To confirm that incorporating the equality effect β as a free parameter in our model conveys an advantage even when weighing parsimony against model fit, future research with increased sample sizes is necessary.

Discussion

Including confidence ratings in the theoretically optimal way using CWMV increases the simulated group performance over MV. Real groups are more accurately represented by CWMV when individuals provide reliable and independent confidence ratings. Even though real groups consider confidence ratings similar to CWMV, they tend to treat individual responses more equally giving more confident individuals less impact on the group decision than naive CWMV simulations, which is consistent with an equality bias (Bang and Frith 2017; Mahmoodi et al. 2015). Additionally, groups tend to underestimate their confidences.

In our study, individuals were overconfident in hard (difficult) trials and underconfident in easy trials—a finding often referred to as the hard-easy effect (Gigerenzer et al. 1991). Hard trials allow participants to make a correct decision about 50% of the time, but reported confidences were larger. In contrast, easy trials allow for close to 100% confidences but reported confidences were strictly lower. This hard-easy effect, or underextremity (Griffin and Brenner 2004), can be explained by a regressive tendency (Moore and Healy 2008). That is, participants were biased toward their prior belief to observe trials with moderate difficulty. But it can also be explained by a bias introduced through the response format: Olsson (2014) argue that a half scale (50–100%), as it is often used, biases participants to respond closer to the center of the scale.

In contrast, real groups did not tend to be overconfident for hard trials in our setting but real groups exhibited overall underconfidence in a double sense: First, group confidences were lower than ideal responses (see Fig. 3b). Second, group confidences were lower than determined by CWMV simulations based on individual responses (see Fig. 4b).

Interestingly, confidence ratings reflected subjective probabilities rather than consistency in our study. For example, we presented a stimulus sequence that is suited to evoke a low ideal confidence of 54% (see Table 1, Scenario III, Individual B). For this sequence, participants

gave the correct decision in 85.7% of the trials and reported confidences relatively close to the ideal confidence with an average of 62% (see Fig. 4, second black point from left). In other words: Participants consistently determined the correct decision but nevertheless reported in their confidence ratings that the strength of evidence was quite low as intended.

One limitation that our well-controlled setting cannot account for is situations in which individuals consensually reach incorrect decisions with high confidence (see Koriat 2015, 2017; Litvinova et al. 2019). In such situations, confidences toward the incorrect decision are aggregated and can lead to high group confidences toward incorrect decisions. In how far CWMV can adequately reflect real groups in these situations remains to be shown because consensually incorrect decisions were too rare in our setting to allow inferences, see bottom left quadrants in Fig. 4.

Further insight into group processes can be gained by fixing the ideal group confidence and varying the constellation of individual confidences. For example, our Scenario II determined an ideal group confidence of 75% based on one confident individual (76% for biased coin) and two almost uninformative individuals (51% for fair coin). The same ideal confidence of 75% would come from three equally confident members (59% for biased coin). CWMV predicts the same ideal confidence, but real groups may behave differently in these two cases. From our current estimates of the equality bias ($\beta = 0.67$), we predict that real groups are more confident in the latter constellation where each individual contributes an equal confidence as compared to a situation where only one individual is very confident.

Our controlled setting provided optimal conditions for CWMV with independent confidence ratings, but it was rather artificial. This allowed us to verify that groups are indeed able to perform confidence weighting to some extent. However, in real-world tasks, bad calibration of confidences may prevent simulated groups to perform as well as real groups. For example, Klein and Epley (2015) observed that individuals could not report well-calibrated confidence ratings, but real groups still outperformed simulations using MV. One possibility is that individuals failed to rate their confidence in a comparable way when verbal scales were used instead of numeric scales (Windschitl and Wells 1996): Klein and Epley used a 9-point Likert scale from "not at all confident" (1) to "very confident" (9). Nevertheless, participants might have been able to share calibrated confidences in the real group discussions. This could have led to a better performance of real compared to simulated groups.

Another possible reason for real groups outperforming simulated groups is that the assumption of independence is violated. These—arguably more realistic—situations have been investigated under the name of *hidden profiles*, where a hidden profile determines the distribution of information that is either common among or unique to individuals (Stasser and Titus 2003; Stasser and Abele 2020). Distinguishing between evidence that is held by all individuals of a group versus evidence that is uniquely known by few individuals is a crucial aspect of successful real groups (Mercier 2016). Consider again the example of three individuals deciding whether a suspect is guilty or not. Say, individuals have in total five pieces of evidence: two incriminating, I_1 and I_2, and three exonerating, E_1, E_2 and E_3. All individuals know all the incriminating evidence but each individual knows only one unique piece of exonerating evidence. That is, the first individual knows I_1, I_2, and E_1; the second knows I_1, I_2, and E_2; and the third knows I_1, I_2, and E_3. For each individual there is more incriminating evidence and each would decide 'guilty' with some confidence. Incorrectly assuming independence, CWMV would simulate the group decision to be guilty as well. However, a real group might lay out all the evidence, find in total more exonerating evidence, and decide 'not guilty.'

There are some approaches to handle such dependencies formally (Kaniovski and Zaigraev 2011; Shapley and Grofman 1984; Stasser and Titus 1987) each coming with its own set of particular, additional assumptions. To sketch the approach that we find most promising: CWMV could be applied not to the potentially dependent individual responses but to the independent pieces of evidence, with confidences indicating the strength of each piece of evidence. Incorporating CWMV in this way could improve theoretical predictions: Rather than comparing group performance to the best individual (as is often done), CWMV-inspired approaches may provide a more adequate baseline for group performance even when information is distributed in a way that violates the independence assumptions for individual responses.

Conclusion

Confidence ratings of individuals play an important role in real group decisions and can be used to increase simulated group performance. In a controlled setting, real groups have proven to aggregate confidences in a way that is to some extent consistent with the CWMV even though they tend to treat individual responses more equal and with lower confidence than when using CWMV simulations. Developing group simulation methods (for example to account for dependencies) and comparing simulated group decisions using those methods to real group decisions will deepen our understanding of real-world group discussion.

Meyen *et al. Cogn. Research*   (2021) 6:18

Page 12 of 13

Abbreviations
MV: Majority vote; CWMV: Confidence weighted majority vote.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s41235-021-00279-0.

> **Additional file 1:** Summary statistics and parameter estimates for our experiment. For each group, we report the average accuracy (%-correct responses across 12 trials) for real and simulated groups, regression coefficients a and b (real confidence $= a + b \cdot$ ideal confidence $+ \varepsilon$), and fitted parameters from our formal cognitive modeling approach: equality effect β, group confidence effect γ and group precision σ_g. Additionally, we report mean, standard deviation as well as the more robust median and quartiles.

Authors' contributions
SM, DMBS, and VF conceptualized the experiment and the analyses. DMBS and SM performed the experiment and implemented the analyses. SM, VF, DMBS, and UL wrote manuscript. All authors read and approved the final manuscript.

Availability of data and materials
The experimental code and data set supporting the conclusions of this article is available in the Open Science Framework repository (doi: https://doi.org/10.17605/OSF.IO/G69KZ), https://osf.io/g69kz/?view_only=5c068a3d93cd4915ac6e86c53b19ccdc.

Ethics approval and consent to participate
Written informed consent was obtained from all participants. The experiment was conducted in accordance with the 1964 Declaration of Helsinki and with the ethical guidelines of the German Psychological Society (DGPs) and the Professional Association of German Psychologists (BDP) (2005, C.III).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1] Experimental Cognitive Science, Department of Computer Science, University of Tübingen, Tübingen, Germany. [2] Theory of Machine Learning, Department of Computer Science, University of Tübingen, Tübingen, Germany. [3] Max Planck Institute for Intelligent Systems, Tübingen, Germany.

References
Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., & Van Gael, J. (2012). Crowd IQ: Aggregating opinions to boost performance. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems* (Vol. 1, pp. 535–542).

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

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, *4*(8), 170193.

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

Birnbaum, M. H., & Diecidue, E. (2015). Testing a class of models that includes majority rule and regret theories: Transitivity, recycling, and restricted branch independence. *Decision*, *2*(3), 145.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674–681.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge: Cambridge University Press.

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 171–189). New York, NY: Psychology Press.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286.

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279.

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, *5*(1), 1.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3–4), 303–336.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, *199*, 158–177.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435.

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

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508.

Hautz, W. E., Kämmer, J. E., Schauber, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *Journal of the American Medical Association*, *313*, 303–304.

Kaniovski, S., & Zaigraev, A. (2011). Optimal jury design for homogeneous juries with correlated votes. *Theory and Decision*, *71*(4), 439–459.

Klein, N., & Epley, N. (2015). Group discussion improves lie detection. *Proceedings of the National Academy of Sciences*, *112*(24), 7460–7465.

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80.

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362.

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*(5), 934.

Koriat, A. (2017). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077.

Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J., & Graepel, T. (2012). Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th annual ACM web science conference* (pp. 151–160).

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., et al. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782.

Meyen *et al. Cogn. Research*   *(2021) 6:18*

Page 13 of 13

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379.

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision*,. https://doi.org/10.31234/osf.io/ngweh.

Litvinova, A., Kurvers, R. H., Hertwig, R., & Herzog, S. M. (2019). *When experts make inconsistent decisions*. https://doi.org/10.31234/osf.io/dtaz3.

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

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276.

Marshall, J. A., Brown, G., & Radford, A. N. (2017). Individual confidence-weighting and group decision-making. *Trends in Ecology and Evolution*, *32*(9), 636–645.

Martins, A. C. (2006). Probability biases as Bayesian inference. *Judgment and Decision Making*, *1*(2), 108.

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*, 1455.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502.

Neth, H., Sims, C. R., & Gray, W. D. (2016). Rational task analysis: A methodology to benchmark bounded rationality. *Minds and Machines*, *26*(1–2), 125–148.

Nitzan, S., & Paroush, J. (1982). Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, *23*(2), 289–297.

Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research*, *67*(8), 1766–1770.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366.

Puncochar, J. M., & Fox, P. W. (2004). Confidence in individual and group decision making: When "two heads" are worse than one. *Journal of Educational Psychology*, *96*(3), 582.

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., et al. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shapley, L., & Grofman, B. (1984). Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, *43*(3), 329–343.

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

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

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183.

Stasser, G., & Abele, S. (2020). Collective choice, collaboration, and communication. *Annual Review of Psychology*, *71*, 589–612.

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology*, *53*(1), 81.

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, *14*(3–4), 304–313.

van Dijk, F., Sonnemans, J., & Bauw, E. (2014). Judicial error by groups and individuals. *Journal of Economic Behavior and Organization*, *108*, 224–235.

Vercammen, A., Ji, Y., & Burgman, M. (2019). The collective intelligence of random small crowds: A partial replication of Kosinski et al. (2012). *Judgment and Decision Making*, *14*(1), 91.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLoS ONE*, *10*(8), e0134269.

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: What subjective measures of visual consciousness are about. *Attention, Perception, and Psychophysics*, *75*(7), 1406–1426.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.