How wise is the crowd: Can we infer people are accurate and competent merely because they agree with each other?

Abstract

Are people who agree on something more likely to be right and competent? Evidence suggests that people tend to make this inference. However, standard wisdom of crowds approaches only provide limited normative grounds. Using simulations and analytical arguments, we argue that when individuals make independent and unbiased estimates, under a wide range of parameters, individuals whose answers converge with each other tend to have more accurate answers and to be more competent. In 6 experiments (UK participants, total N = 1197), we show that participants infer that informants who agree have more accurate answers and are more competent, even when they have no priors, and that these inferences are weakened when the informants were systematically biased. In conclusion, we speculate that inferences from convergence to accuracy and competence might help explain why people deem scientists competent, even if they have little understanding of science.

Introduction

Imagine that you live in ancient Greece, and a fellow called Eratostenes claims the circumference of the earth is 252000 stades (approximately 40000 kilometers). You know nothing about this man, the circumference of the Earth, or how one could measure such a thing. As a result, you discard Eratostenes' opinion and (mis)take him for a pretentious loon. But what if other scholars had arrived at very similar measurements, independently of Eratosthenes? Or even if they had carefully checked his measurement, with a critical eye? Wouldn't that give you enough ground to believe not only that the estimates might be correct, but also that Eratosthenes and his fellow scholars must be quite bright, to be able to achieve such a feat as measuring the Earth?

In this article, we explore how, under some circumstances, we should, and we do infer that a group of individuals whose answers converge are likely to be correct, and to be competent in the relevant area, even if we had no prior belief about either what the correct answer was, or about these individuals' competence.

We begin by reviewing existing studies showing that people infer that competent in-

formants who converge in their opinions are likely to be accurate. The wisdom of crowds literature provides normative grounds for this inference. We then argue that both the experimental and theoretical literature have paid little attention to extending this inference to cases in which there is no information about the informants' competence, and to inferences about the competence of the informants. We first develop normative models, both analytically and with simulations, to show that inferences from convergence to accuracy and to competence are warranted under a wide range of parameters. Second, we present a series of experiments in which participants evaluate both the accuracy and competence of informants as a function of how much their answers converge on a given problem, in the absence of any priors about these individuals' competence, or what the correct answer is.

Do people infer that individuals whose answers converge tend to be right, and to be competent?

The literature on the wisdom of crowds has treated separately situations with continuous answers, such as the weight of an ox in Galton's famous observation (Galton, 1907), and with categorical answers, as when voters have to choose between two options, in the standard Condorcet Jury Theorem (De Condorcet, 2014). The continuous and the categorical case are typically modeled with different tools, and they have usually been studied in different empirical literatures (see below). Given that they both represent common ways for answers to converge more or less (e.g. when people give numerical estimates vs. vote on one of a limited number of options), we treat them both here, with different simulations and experiments.

In the continuous case, the most relevant evidence comes from the literature on 'advice-taking' (for review, see, Kämmer, Choshen-Hillel, Müller-Trede, Black, & Weibler, 2023). In these experiments, participants are called 'judges' who need to make numerical estimates–sometimes on factual knowledge, e.g. 'What year was the Suez Canal opened first?' (Yaniv, 2004), sometimes on knowledge controlled by the experimenters, e.g. 'How many animals were on the screen you saw briefly?' (Molleman et al., 2020). To help answer these questions, participants are given estimates from others, the 'advisors'.

Most of this literature is irrelevant to the point at hand since participants are presented with single estimates, either from a single advisor (e.g. Bednarik & Schultze, 2015; Harvey & Fischer, 1997; Soll & Larrick, 2009; Yaniv, 2004; Yaniv & Kleinberger, 2000), or as an average coming from a group of advisors (e.g. Jayles et al., 2017; Mannes, 2009), but without any information about the distribution of initial estimates, so that we cannot tell whether participants put more weight on more convergent answers.

Some advice-taking studies provide participants with a set of individual estimates. One subset of these studies manipulates the degree of convergence between groups of advisors, through the variance of estimates (Molleman et al., 2020; Yaniv, Choshen-Hillel, & Milyavsky, 2009), or their range (Budescu & Rantilla, 2000; Budescu, Rantilla, Yu, & Karelitz, 2003; Budescu & Yu, 2007). These studies find that participants are more confident about, or rely more on, estimates from groups of advisors that converge more.

Other studies manipulated the degree of convergence within a group of advisors. These studies present participants with a set of estimates, some of which are close to each

other, while others are outliers (Harries, Yaniv, & Harvey, 2004; Yaniv, 1997, study 3 & 4). These studies find that participants discount outliers when aggregating estimates.

Studies on advice taking thus suggest participants believe that more convergent opinions are more likely to be correct. None of these studies investigated whether participants also believe that those whose opinions converge are also more likely to possess greater underlying competence.

In categorical choice contexts, there is ample and long-standing (e.g. Crutchfield, 1955) evidence from experimental psychology that participants are more likely to be influenced by majority opinions, and that this influence is stronger when the majority is larger, both in absolute and in relative terms (e.g., Morgan, Rendell, Ehn, Hoppitt, & Laland, 2012; for review, see H. Mercier & Morin, 2019). This is true even if normative conformity (when people follow the majority because of social pressure rather than a belief that the majority is correct) is unlikely to play an important role (e.g. because the answers are private, see Asch, 1956). Similar results have been obtained with young children (Bernard, Harris, Terrier, & Clément, 2015; Bernard, Proust, & Clément, 2015; Chen, Corriveau, & Harris, 2013; Corriveau, Fusaro, & Harris, 2009; e.g. Fusaro & Harris, 2008; Herrmann, Legare, Harris, & Whitehouse, 2013; Morgan, Laland, & Harris, 2015).

If many studies have demonstrated that participants tend to infer that more convergent answers are more likely to be correct, few have examined whether participants thought that this convergence was indicative of the informants' competence. One potential exception is a study with preschoolers in which the children were more likely to believe the opinion of an informant who had previously been the member of a majority over that of an informant who had dissented from the majority (Corriveau et al., 2009). However, it is not clear whether the children thought the members of the majority were particularly competent, since their task–naming an object–was one in which children should already expect a high degree of competence from (adult) informants. This result might thus indicate simply that children infer that someone who disagrees with several others on how to call something is likely wrong, and thus likely less competent at least in that domain.

What inferences from convergence should we expect people to draw?

Should we expect that people be able to infer that more convergent answers likely indicate not only more accurate answers, but also that those who gave the answers were competent? In order to make the best of communicated information, humans have to be able to evaluate it, so as to discard inaccurate or harmful information, while accepting accurate and beneficial information (Smith & Harper, 2003). It has been argued that a suite of cognitive mechanisms–mechanisms of epistemic vigilance–evolved to serve this function (Hugo Mercier, 2020; Sperber et al., 2010). Since the opinion of more than one individual is often available to us, there should be mechanisms of epistemic vigilance dedicated to processing such situations. It would be these mechanisms that lead us to put more weight on an opinion that is shared by a larger majority (in relative or absolute terms), and, in some cases at least, to discount majority opinion when the opinions haven't been formed independently of each other (Hugo Mercier & Miton, 2019). Evidence suggests that these mechanisms rely on heuristics which become more refined with age (Morgan et al., 2015),

and which are far from perfect (Hugo Mercier & Miton, 2019; in particular, they ignore many cases of informational dependencies, see, e.g. Yousif, Aboody, & Keil, 2019). As mentioned above, in the experiments evaluating how people process convergent information, the participants had grounds to believe that the information came from competent informants. However, the same mechanisms could lead people to perform the same inference when they do not have such information (especially since, as is shown presently, such an inference is warranted).

Regarding the inference from convergence to competence, other cognitive mechanisms allow us to infer how competent people are, based on a variety of cues, from visual access (Pillow, 1989), to the time it takes to answer a question (Richardson & Keil, 2022). One of the most basic of these mechanisms infers from the fact that someone was right, that they possess some underlying competence (e.g. Koenig, Clément, & Harris, 2004). As a result, if participants infer that convergent opinions are more likely to be accurate, they should also infer that the informants who provided the opinions are competent. Before testing whether participants infer accuracy and competence from convergence, we show that these inferences are normatively warranted, both in the categorical and in the continuous case.

Analytical argument

Regarding the analytical answer, the question can be broken down into two questions. First, can we infer that a population of informants whose answers converge more is, on average, more competent? Second, can we infer that, within this population, individuals who are closer to the consensus or to the average answer are more competent?

In the continuous case, let us imagine a population of informants. Individual opinions are drawn from a normal distribution, centered on the correct answer (i.e., informants have no systematic bias, as for instance in Galton's classic demonstration, Galton, 1907). The variance of this distribution represents the individuals' competence: the larger the variance, the lower the competence. We observe the individual answers. In this setting, it is well known in statistics (see also Electronic Supplementary Materials, ESM) that the sample mean (i.e. the mean of the answers of all the informants) is the best estimator of the correct answer, and the sample variance (i.e. the mean squared distance between the answers and the sample mean) is the best estimator of the population's average competence (best understood here as the least volatile estimator). This means that a population of informants whose answers converge more (lower sample variance) is, on average, more competent (informants tend to answer with a lower variance).

We can extend this argument from populations to individuals: Consider that competence varies within a population, such that each informant's answer is drawn from their own distribution – always centered on the correct answer, but with different variances. In this case, the distance between each informant's answer and the sample mean provides the best estimate of that informant's competence (i.e. individuals whose answers are further away from the sample mean tend to be less competent).

In the categorical case, we define the competence of an informant as the probability of choosing the correct answer. The law of large numbers implies that the relative size of the majority – the share of informants who choose the most chosen answer – is, when

the population is large enough, a good approximation of the average competence of the population (e.g. if the average competence is .66, and there are enough informants, then approximately 66% of informants will select the right answer). For smaller populations, the relationship holds, with some degree of noise. In other words, the more the answers converge, the more the population can be inferred to be competent (and the larger the population, the more reliable the inference).

Now, can we infer that informants belonging to the majority are more competent? To do so, we must assume some distribution of competence in the population. Using Bayes theorem, we can then show (see ESM, section A) that the more informants agree with a focal informant, the more the focal informant can be inferred to be competent (Fig. 1). We also find that, the more competence varies in the population, the more the degree of convergence is indicative of an informant's competence. For example, if competence is roughly uniform in the population, then being part of a minority likely reflects bad luck, rather than incompetence. More generally, the fact that the focal individual is right (their 'accuracy') can be inferred more strongly than their competence, as there is noise in the answer choice (an incompetent individual can always pick the correct answer by chance and vice versa). Figure 1 provides an example with a specific categorical choice scenario, under two different population distributions of competence.

Simulations

In order to better understand the influence of different parameters (e.g. degree of convergence, number of individuals) on the relationship between convergence, accuracy, and competence of informants, we conducted simulations. In all simulations, we assume agents to be unbiased and independent in their answers, but varying in their competence. All code and data regarding the simulations can be found on Open Science Framework project page (https://osf.io/6abqy/?view_only=42632bccea604fd7928dbe58e087d23b).

Continuous case. Groups of agents, with each agents' competence varying, provide numerical answers. We measure how accurate these answers are, and how much they converge (i.e. how low their variance is). We then look at the relationship between convergence and both the accuracy of the answers and the competence of the agents.

More specifically, agents provide an estimate on a scale from 1000 to 2000 (chosen to match the experiments below). Each agent is characterized by a normal distribution of possible answers. All of the agents' distributions are centered around the correct answer, but their standard deviation varies, representing varying degrees of competence. The agents' standard deviation varies from 1 (highest competence) to 1000 (lowest competence). Each agent's competence is drawn from a population competence distribution, expressed by a beta distribution, which can take different shapes. We conducted simulations with a variety of beta distributions which cover a wide range of possible competence populations (see Fig. 2 A).

A population of around 990000 agents (varying slightly as a function of group sizes) with different competence levels is generated. An answer is drawn for each agent, based on their respective competence distribution. The accuracy of this answer is defined as the squared distance to the true answer. Having a competence and an accuracy value for each

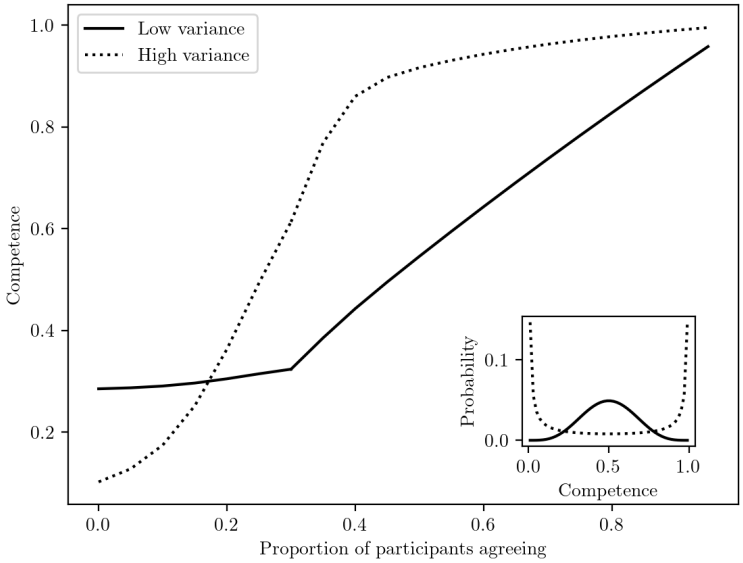


Figure 1. Results of the analytic argument. The figure shows the average estimated competence and accuracy for a focal individual, depending on the proportion of individuals agreeing with them, in a categorical choice scenario. Here, we assume that 20 individuals answer a 5-choice question. For values higher than four, we assume that the focal individual is part of the majority. Below four, we assume that the focal individual is part of a minority, and that the majority size is four. Two situations are represented: one in which there is low variance in the distribution of competence in the population (orange line in the insert), and one in which there is high variance in this distribution (blue line in the insert).

agent, we randomly assign agents to groups of, e.g., three. For each group, we calculate the average of the agents' competence and accuracy. We measure the convergence of a group's answers by calculating the standard deviation of the agents' answers. We repeat this process for different sample sizes for the groups, and different competence distributions. Fig. 2 C displays the resulting relation of convergence with accuracy (left), and competence (right) for different underlying competence distributions and group sizes. We draw broad conclusions from these results after reporting the outcome of the simulations with categorical answers.

Categorical case. In the case of categorical answers, convergence can be measured as the relative size of informants picking the same option. The Condorcet jury theorem (De Condorcet, 2014; for a recent treatment, see Dietrich & Spiekermann, 2013) and its extensions to situations with more than two options (e.g., Hastie & Kameda, 2005) already shows that the answer defended by the majority/plurality is more likely to be correct. Regarding competence, Romeijn and Atkinson (2011) have shown that when individuals are more competent, the share of the majority vote tends to increase. However, they have only studied a case with a uniform distribution of competence. By contrast, here, we investigate a wide range of distributions of competence, and we do not assume that all individuals are equally competent, meaning that, from an observed set of answers, we can

attribute an average competence to individuals whose answers form a majority/plurality vs. individuals whose answers form a minority.

We simulate agents whose competence varies and who have to decide between a number of options, one of which is correct. Competence is defined as a value p which corresponds to the probability of selecting the right answer (the agents then have a probability (1-p)/(m-1), with m being the number of options, of selecting any other option). Competence values range from chance level ($p = 1/m$) to always selecting the correct option ($p = 1$). Individual competence levels are drawn from the same population competence beta distributions as in the numerical case (see Fig. 2 A). Based on their competence level, we draw an answer for each agent. We measure an agent's accuracy as a binary outcome, namely whether they selected the correct option or not. In each simulation around 99900 agents (varying slightly as a function of the group size) are generated, and then randomly assigned to groups (of varying size in different simulations). Within these groups, we first calculate the share of individuals voting for each answer, allowing us to measure convergence. For example, in a scenario with three choice options and three individuals, two might vote for option A, and one for option C, resulting in two levels of convergence, 2/3 for A and 1/3 for C. For each level of convergence occurring within a group, we then compute (i) the accuracy (either 1 if the correct option or 0 else)„ (ii) the average competence of agents. Across all groups, we then compute the averages of these values, for each level of convergence.

We repeat this procedure varying population competence distributions and the size of informant groups, holding the number of choice options constant at $n = 3$ (for simulations with varying choice options, see ESM section B). Fig. 2 B shows the average accuracy (left), and the average competence (right) value as a function of convergence, across different underlying competence distributions and group sizes.

The simulations for the numerical and categorical case demonstrate a similar pattern, which can be summarized as follows:

1. Irrespective of group size (and number of choice options) and of the competence distribution, there is a very strong relation between convergence and accuracy: more convergent answers tend to be more accurate.
2. For any group size and any competence distribution, there is a relation between convergence and the competence of the agents: more convergent answers tend to stem from more competent agents. The strength of this relation is not much affected by the number of agents whose answers are converging, but, although it is always positive, it ranges from very weak to very strong depending on the population's competence distribution.
3. The relation between convergence and accuracy is always much stronger than the relation between convergence and competence of the agents.

Overview of the experiments

Our models indicate that groups of informants are more likely to have given accurate answers, and to be competent, when their answers converge. In a series of experiments, we test whether people draw these inferences both in numerical tasks (Experiments 1, 2, 3), and in categorical tasks (Experiments 4, 5, 6). By contrast with previous studies,

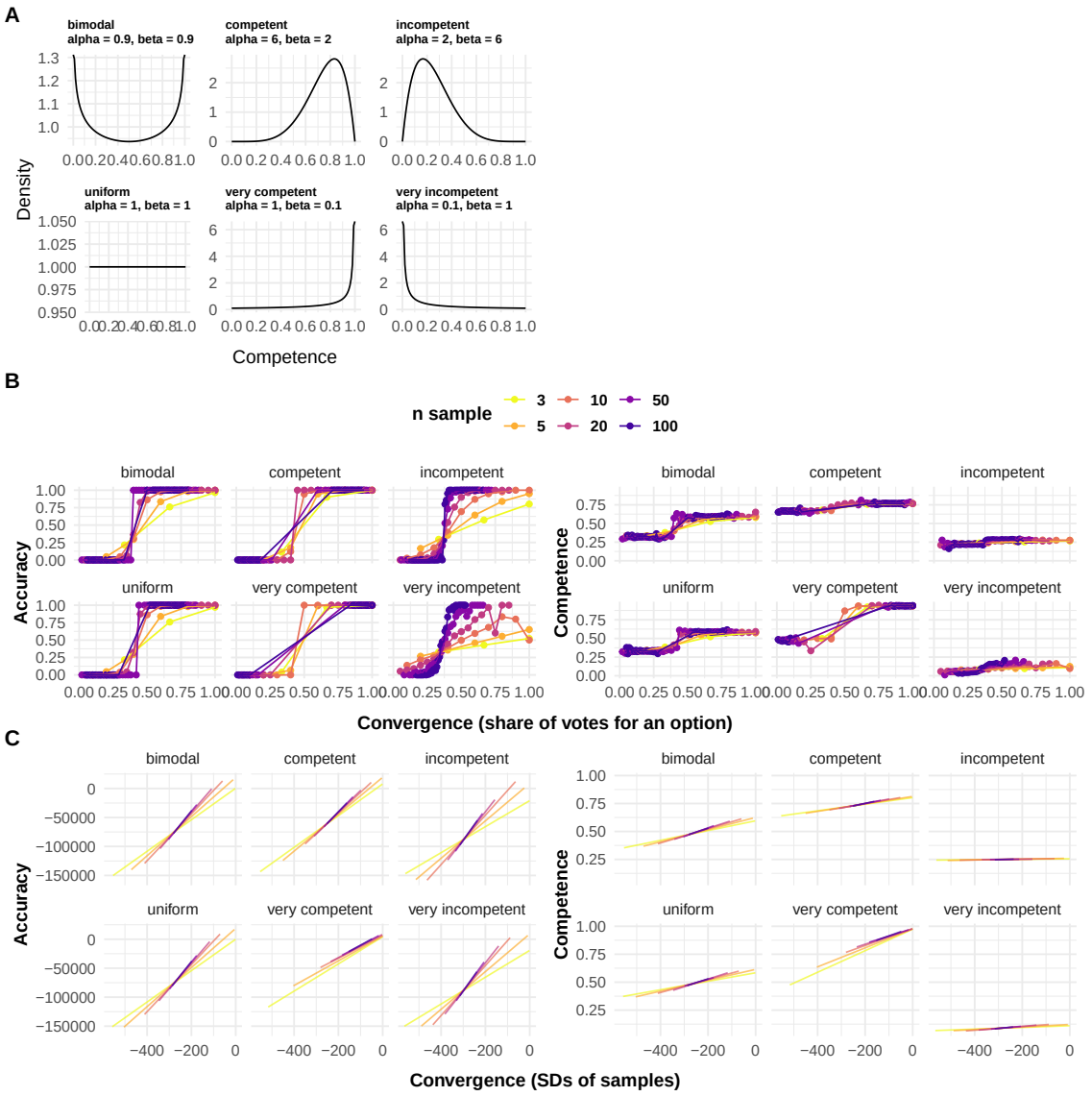


Figure 2. Results of the simulations. **A** Shows the different population competence distributions we considered in our simulations. In the continuous simulations, competence values of 0 correspond to a very large standard deviation (1000, with a mean of 1500, on a scale from 1000 to 2000), thereby practically taking the form of a uniform distribution, while competence of 1 corresponds to a very small standard deviation (1, on the same scale). In the categorical simulations, a competence value of 0 corresponds to chance (e.g. in a 3-choice-options scenario, an individual picking the correct answer with a probability of 1/3), while a competence value of 1 corresponds to definitely picking the correct answer. **B** Shows the results of simulations in a categorical setting with three choice options. Points represent average accuracy (left)/competence (right) values by degree of convergence (measured by the share of votes for an option), for different population competence distributions (panels) and sample sizes (colors). **C** Shows the results in a continuous setting. Regression lines represent the correlation between accuracy (left; measured by squared distance to true mean and reversed such that greater accuracy corresponds to being further up on the y-axis) or competence (right), respectively, and convergence (reversed such that greater convergence corresponds to being more right on the x-axis).

participants were not given any information about the tasks–how difficult they were–and the informants–how competent they might be. There has recently been much interest in investigating whether participants are able to take informational dependencies into account when evaluating convergent information (Desai, Xie, & Hayes, 2022; Hugo Mercier & Miton, 2019; Xie & Hayes, 2022; Yin, Xu, Lin, Zhou, & Guo, 2024; Yousif et al., 2019). To test whether participants took into account dependencies between the informants in the present context, this factor was manipulated in two different ways (in Experiment 2, some informants had discussed with each other; in Experiments 3 and 5, some informants had an incentive to provide the same answer). We predicted that dependencies between the informants' answers should reduce participants' reliance on the convergence of the answer as a cue to accuracy and to competence.

All experiments were preregistered. All documents, data and code can be found on Open Science Framework project page (https://osf.io/6abqy/?view_only=42632bccea604fd7928dbe58e087d23b). All analyses were conducted in R (version 4.2.2) using R Studio. For most statistical models, we relied on the `lme4` package and its `lmer()` function. Unless mentioned otherwise, we report unstandardized model coefficients that can be interpreted in units of the scales we use for our dependent variables.

Experiment 1

In Experiment 1, participants were provided with a set of numerical estimates which were more or less convergent, and asked whether they thought the estimates were accurate, and whether the informants making the estimates were competent. Perceptions of accuracy were measured as the confidence in what the participants thought the correct answer was, on the basis of the numerical estimates provided: the more participants think they can confidently infer the correct answer, the more they must think the estimates accurate, on average (the results replicate with a more direct measure of accuracy, see Experiment 3). Our hypotheses were:

H1: When making a guess based on the estimates of (independent) informants, participants will be more confident about their guess when these estimates converge compared to when they diverge.

H2: Participants perceive (independent) informants whose estimates converge more as more competent than informants whose estimates diverge.

We had three research questions regarding the number of informants which report in the ESM (section C).

Methods

Participants. We recruited 200 participants from the UK via Prolific (100 female, 100 male; age_{mean}: 39.73, age_{sd}: 15.39, age_{median}: 35.50). Not a single participant failed our attention check. The sample size was determined on the basis of a power analysis for a t-test to detect the difference between two dependent means ("matched pairs") run on G*Power3. The analysis suggested that a combined sample of 199 would provide us with 80% power to detect a true effect size of Cohen's d $\geq$ 0.2 (alpha = .05, two-tailed).

Procedure. After providing their consent to participate in the study and passing an attention check, participants read the following introduction: "Some people are playing games in which they have to estimate various quantities. Each game is different. You have no idea how competent the people are: they might be completely at chance, or be very good at the task. It's also possible that some are really good while others are really bad. Some tasks might be hard while others are easy. Across various games, we will give you the answers of several players, and ask you questions about how good they are. As it so happens, for all the games, the estimates have to be between 1000 and 2000, but all the games are completely different otherwise, and might require different skills, be of different difficulties, etc. Each player in the game makes their own estimate, completely independent of the others". After being presented with the results of a game (Fig. 3), participants had to (i) make a guess about the correct answer based on the estimates they see, "What would you guess is the correct answer, if there is one?", (ii) estimate their confidence in this guess, "How confident are you that your answer is at least approximately correct?" on a 7-point Likert scale ("not confident at all" to "extremely confident"), (iii) estimate the competence of the group of players whose estimates they saw, "On average, how good do you think these players are at the game?", also on a 7-point Likert scale (from "not good at all" to "extremely good").

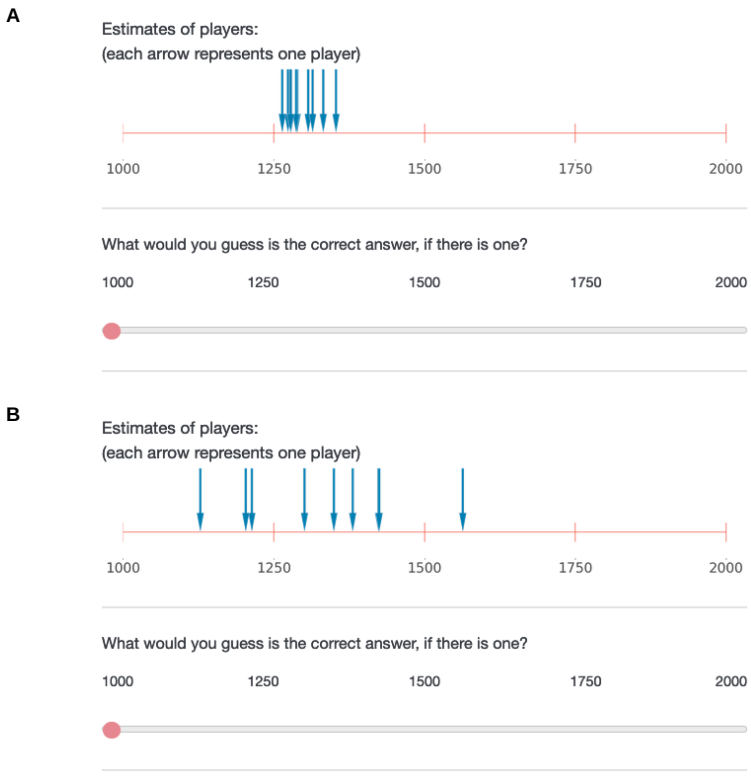


Figure 3. Example of two stimuli from Experiment 1, both in the 10 players condition, **A** corresponding to the convergent, **B** to the divergent condition. Similar stimuli are used in Experiments 2 and 3.

Design. We manipulated two experimental factors, with two levels each: the convergence of the estimates (how close they were to each other; levels: divergent/convergent), and the number of estimates (how many players there were; levels: three/ten). This latter factor was chiefly included to make our results more robust, and was not attached to specific hypotheses. We used a 2 (convergence: divergent/convergent) x 2 (number: three/ten) within-participant design, with each participant seeing all the conditions. Participants saw two different sets of estimates per condition, for a total of eight sets of estimates per participant.

Materials. We generated sets of estimates with random draws from normal distributions. First, we varied the standard deviation of these distributions to simulate the degree of convergence (150 for divergence, 20 for convergence; estimate scale ranged from 1000 to 2000). Second, we varied the number of draws (either three or ten) from these distributions. For each of the four possible resulting conditions, we generated two random draws. We repeated this process for three different sets of estimates, and participants were randomly assigned to one of these sets. More information on how the stimuli were created can be found in the ESM, section C.

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept and a random slopes of convergence for participants. In the models for our hypotheses, we control for the number of estimates provided to the participants (three or ten). Visualizations and descriptive statistics can be found in ESM, section C. We find a positive effect of convergence on accuracy: Participants were more confident about their estimate in convergent scenarios (mean = 4.56, sd = 1.45) than in divergent ones (mean = 3.19, sd = 1.39; $\hat{b}_{\text{Accuracy}} = 1.37$ [1.225, 1.51], p = $< .001$). We also find a positive effect of convergence on competence: participants rated players as more competent in convergent scenarios (mean = 4.75, sd = 1.24) than in divergent ones (mean = 3.52, sd = 1.27; $\hat{b}_{\text{Competence}} = 1.23$ [1.065, 1.4], p = $< .001$).

In an exploratory, non-preregistered analysis, we tested whether the effect of convergence is larger on accuracy than on competence. To this end, we regressed the outcome score on convergence and its interaction with a binary variable indicating which outcome was asked for (accuracy or competence), while controlling for the number of informants. We do not find a statistically significant interaction that would indicate a difference of the effect of convergence ($\hat{b} = 0.14$ [-0.001, 0.271], 0.052). Pooled across divergent and convergent conditions, we find that participants reported lower perceived accuracy than competence ($\hat{b}$ = -0.26 [-0.359, -0.156], $< .001$)

In summary, as predicted, when the informants' answers were more convergent, participants were more confident that their answers were correct, and they believed the informants to be more competent. This was true both when there were three informants and when there were ten informants.

Experiment 2

We have shown that it is rational to infer that convergent estimates are more likely to be accurate, and to have been made by competent individuals, only if these individuals were independent and unbiased. However, convergence could come about differently. If the individuals do not make their estimates independently of each other, a single individual might exert a strong influence on the others, making their convergence a poor cue to their accuracy. Alternatively, all individuals might have an incentive to provide a similar, but not accurate answer. In Experiment 2, we investigate the first possibility, and the second in Experiment 3. In particular, for Experiment 2 we rely on past results showing that participants, under some circumstances, put less weight on opinions that have been formed through discussion, by contrast with more independent opinions (Harkins & Petty, 1987; Einav, 2018; Hess & Hagen, 2006; see also Lopes, Vala, & Garcia-Marques, 2007). We sought to replicate this finding in the context of convergent estimates, formulating the following hypotheses:

H1: When making a guess based on convergent estimates of informants, participants will be more confident about their guess when informants were independent compared to when they weren't (i.e. they could discuss before).

H2: Participants perceive informants whose estimates converge as more competent when they are independent, compared to when they weren't (i.e. they could discuss before).

Note that these predictions only stem from past empirical results, and are not necessarily normatively justified (modeling the effects of discussion would be beyond the scope of this paper, but see Dietrich & Spiekermann, 2024).

Methods

Participants. We recruited 200 participants from the UK via Prolific (100 female, 99 male, 1 not-identified; age_{mean}: 40.54, age_{sd}: 13.56, age_{median}: 38.50). Not a single participant failed our attention check. As for experiment 1, the sample size was determined on the basis of a power analysis for a t-test to detect the difference between two dependent means ("matched pairs") run on G*Power3. The analysis suggested that a combined sample of 199 would provide us with 80% power to detect a true effect size of Cohen's $d \geq 0.2$ (alpha = .05, two-tailed).

Design. In a within-participants design, participants saw both an independence condition, in which they were told "Players are asked to make completely independent decisions – they cannot see each other's estimates, or talk with each other before giving their estimates," and a dependence condition, in which they were told "Players are asked to talk with each other about the game at length before giving their estimates."

Materials. We used the materials generated for the convergent condition of Experiment 1. By contrast to Experiment 1, participants saw only two stimuli in total (one set of estimates per condition), and we only used stimuli involving groups of three informants. Otherwise, we proceeded just as in Experiment 1: we randomly assigned individual participants to one of the three series of stimuli, and for each participant, we randomized the order of appearance of conditions.

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept for participants. Visualizations and descriptive statistics can be found in ESM, section D. The data does not support our hypotheses. Participants were slightly less confident about their estimates when the converging informants were independent (mean = 3.78, sd = 1.50), compared to when they discussed (mean = 4.03, sd = 1.39; $\hat{b}_{\text{Accuracy}}$ = -0.26 [-0.462, -0.048], p = 0.016). The effect is small, but in the opposite direction of what we had predicted. We do not find an effect regarding competence ($\hat{b}_{\text{Competence}}$ = -0.12 [-0.272, 0.032], p = 0.120).

Contrary to the hypotheses, participants did not deem convergent estimates made after a discussion, compared to independently made estimates, to be less accurate, or produced by less competent individuals. This might stem from the fact that participants, in various situations, neglect informational dependencies (Yousif et al., 2019), or from the fact that discussing groups actually perform better than non-discussing groups in a range of tasks (for review, see, e.g., H. Mercier, 2016), including numerical estimates (e.g. H. Mercier & Claidière, 2022). As a result, the participants in the current experiment might have been behaving rationally when they did not discount the estimates made after discussion.

Experiment 3

Experiment 3 tests whether participants are sensitive to another potential source of dependency between convergent estimates: when the individuals making the estimate share an incentive to bias their estimates and disregard accuracy. Even though Experiment 3 is formally similar to Experiment 1, the setting is different, as participants were told that they would be looking at (fictional) predictions of experts for stock values, instead of the answers of individuals in abstract games. In the conflict of interest condition, the experts had an incentive to value the stock in a given way, while they had no such conflict of interest in the independence condition. We tested for an interaction, namely whether the positive effect of convergence is reduced when informants are systematically biased, compared to when they are not. On this basis, we formulate four hypotheses, two of which are identical to those of Experiment 1, and only apply in the independent condition, and two that bear on the interaction:

H1a: Participants perceive predictions of independent informants as more accurate when they converge compared to when they diverge.

H1b: Participants perceive independent informants as more competent when their predictions converge compared to when they diverge.

H2a: The effect of convergence on accuracy (H1a) is more positive in a context where informants are independent compared to when they are in a conflict of interest.

H2b: The effect of convergence on competence (H1b) is more positive in a context where informants are independent compared to when they are in a conflict of interest.

We have not conducted simulations to validate these predictions. However, given the operationalization chosen, in which convergence should provide little evidence of either accuracy or competence, we believe the predictions regarding the superiority of the independent informants stem naturally from the model of independent informants presented above.

Methods

Participants. The interaction design of our third experiment made the power analysis more complex and less standard than for experiments one and two. Because we could build upon data from the first experiment, we ran a power analysis by simulation. The simulation code is available on the OSF, and the procedure is described in the preregistration document. The simulation suggested that 100 participants provide a significant interaction term between 95% and 97% of the time, given an alpha threshold for significance of 0.05. Due to uncertainty about our effect size assumptions and because we had resources for a larger sample, we recruited 199 participants for this study – again, from the UK and via Prolific (99 female, 100 male; age_{mean}: 40.30, age_{sd}: 12.72, age_{median}: 38).

Procedure. After providing their consent to participate in the study and passing an attention check, participants read the following introduction: "You will see four scenarios in which several experts predict the future value of a stock. You have no idea how competent the experts are. It's also possible that some are really good while others are really bad. As it so happens, in all scenarios, the predictions for the value of the stock have to lie between 1000 and 2000. Other than that, the scenarios are completely unrelated: it is different experts predicting the values of different stocks every time." Participants then saw the four scenarios, each introduced by a text according to the condition the participant was assigned to. To remove any potential ambiguity about participants' inferences on the accuracy of the estimates, we replaced the question about confidence to one bearing directly on accuracy: "On average, how accurate do you think these three predictions are?" on a 7-point Likert scale ("not accurate at all" to "extremely accurate"). The question about competence read: "On average, how good do you think these three experts are at predicting the value of stocks?", also assessed on a 7-point Likert scale (from "not good at all" to "extremely good").

Design. We manipulated two factors: informational dependency (two levels, independence and conflict of interest; between participants) and convergence (two levels, convergence and divergence; within participants). In the independence condition, the participants read "Experts are independent of each other, and have no conflict of interest in predicting the stock value - they do not personally profit in any way from any future valuation of the stock." In the conflict of interest condition, the participants read "All three experts have invested in the specific stock whose value they are predicting, and they benefit if other people believe that the stock will be valued at [mean of respective distribution] in the future."

Materials. The distributions presented were similar to those of Experiment 1, although generated in a slightly different manner (see ESM, section E). Each participant rated four scenarios, two for each level of convergence. By contrast to experiment one, all scenarios only involved groups of three informants.

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept and a random slope for convergence for participants. We find evidence for all four hypotheses. As for the first set of hypotheses, to match the setting of experiment one, we reduced the sample of Experiment 3 to half of the participants, namely those who were assigned to the independence condition. On this reduced sample, we ran the exact same analyses as in Experiment 1 and replicated the results. As for accuracy, participants rated informants in convergent scenarios (mean = 5.28, sd = 1.05) as more accurate than in divergent ones (mean = 3.40, sd = 1.08; $\hat{b}_{\text{Accuracy}}$ = 1.88 [1.658, 2.102], p = < .001). As for competence, participants rated informants in convergent scenarios (mean = 5.24, sd = 0.99) as more competent than in divergent ones (mean = 3.61, sd = 1.11; $\hat{b}_{\text{Competence}}$ = 1.62 [1.411, 1.839], p = < .001).

The second set of hypotheses targeted the interaction of informational dependency and convergence (Fig. 4). In the independence condition, the effect of convergence on accuracy was more positive ($\hat{b}_{\text{interaction, Accuracy}}$ = 0.99 [0.634, 1.348], p = < .001) than in the conflict of interest condition ($\hat{b}_{\text{baseline}}$ = 0.89 [0.636, 1.142], p = < .001). Likewise the effect of convergence on competence is more positive ($\hat{b}_{\text{interaction, Competence}}$ = 0.80 [0.474, 1.13], p = < .001) than in the conflict of interest condition ($\hat{b}_{\text{baseline}}$ = 0.82 [0.591, 1.056], p = < .001).

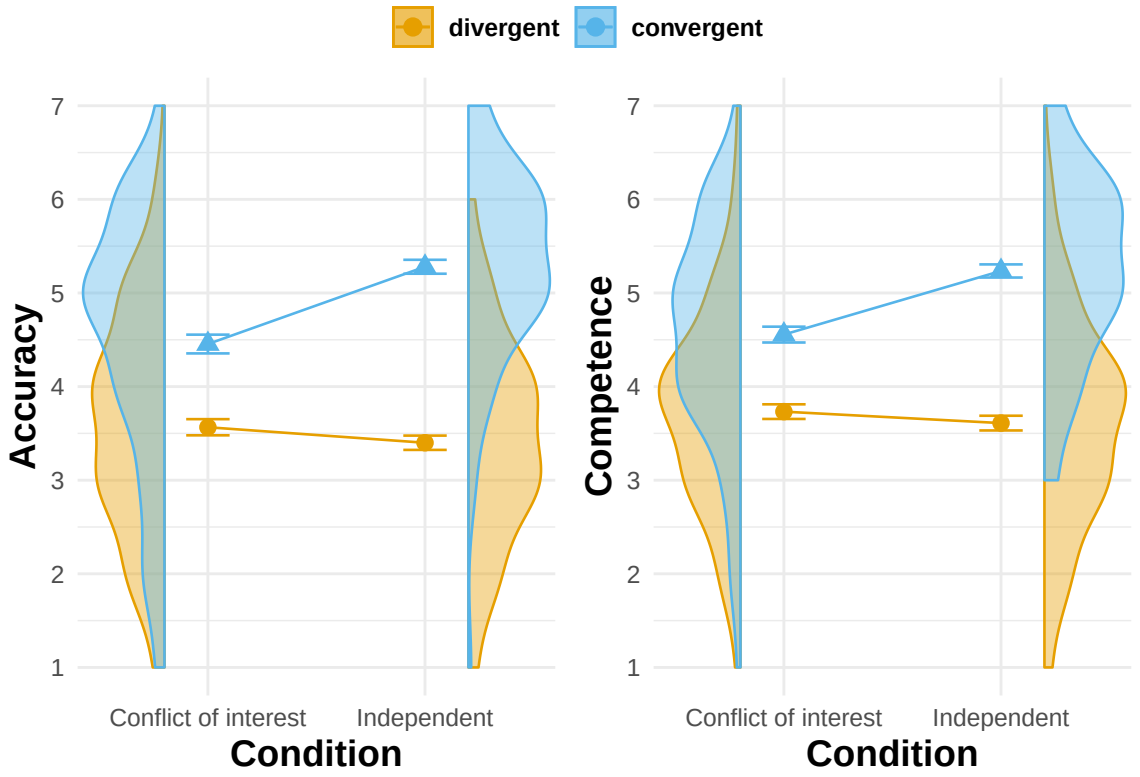


Figure 4. Results of Experiment 3, showing the distributions of accuracy and competence ratings by convergence and informational dependency.

In an exploratory, non-preregistered analysis, we tested whether the effect of convergence is larger on accuracy than on competence. To this end, we regressed the outcome score on convergence and its interaction with a binary variable indicating which outcome was asked for (accuracy or competence), while controlling for informational dependency. We find a negative interaction effect, indicating that pooled across independent and conflict of interest conditions, the effect of convergence had as smaller effect on competence than on accuracy ($\hat{b}$ = -0.16 [-0.294, -0.028], 0.018). Pooled across all conditions, participants reported higher perceived competence than accuracy ($\hat{b}$ = 0.11 [0.025, 0.191], 0.011).

Experiment 3 shows that, when the individuals making the estimates are systematically biased, participants put less weight on the convergence of their estimates to infer that the estimates are accurate, and that the individuals making them are competent.

Experiment 4

In a second series of experiments, we test similar predictions to those of the previous experiments, but in a categorical choice context. The set-up is similar to that of Experiment 1, except that the outcomes seen by the participants are not numerical estimates, but choices made between a few options. An additional difference is that participants rate a focal informant, and not a group of informants. There were two reasons for this choice: First, accuracy is not on a continuum as in the first three experiments (an option was either correct or not), so forming an average across informants who chose different options was less sensible. Second, rating a focal individual allowed us to have a minority condition, which would not have been possible when providing an average rating for a group. Experiment 4 tests hypotheses that are analogous to those of Experiment 1:

H1: Participants perceive an estimate of an independent informant as more accurate the more it converges with the estimates of other informants.

H2: Participants perceive an independent informant as more competent the more their estimate converges with the estimates of other informants.

Methods

Participants. We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found on the OSF. We ran two different power analyses, one for each outcome variable. We set the power threshold for our experiment to 90%. The power simulation for accuracy suggested that even for as few as 10 participants (the minimum sample size we simulated data for), we would have a power of close to 100%. The simulation for competence suggested that we achieve statistical power of at least 90% with a sample size of 30. Due to uncertainty about our assumptions and because it was within our budget, we recruited 100 participants, from the UK and via Prolific (50 female, 50; age_{mean}: 37.32, age_{sd}: 11.53, age_{median}: 36).

Procedure. After providing their consent to participate in the study and passing an attention check, participants read the following introduction: “To be able to understand the task, please read the following instructions carefully: Some people are playing games in which they have to select the correct answer among three answers. You will see the results of several of these games. Each game is different, with different solutions and involving

different players. All players answer independently of each other. At first, you have no idea how competent each individual player is: they might be completely at chance, or be very good at the task. It's also possible that some players are really good while others are really bad. Some games might be difficult while others are easy. Your task will be to evaluate the performance of one of the players based on what everyone's answers are." They were then presented to the results of eight such games (Fig. 5). To assess perceived accuracy, we asked: "What do you think is the probability of player 1 being correct?". Participants answered with a slider on a scale from 0 to 100. To assess perceived competence, we asked participants: "How competent do you think player 1 is in games like these?" Participants answered on a 7-point Likert scale (from (1)"not competent at all" to (2)"extremely competent").

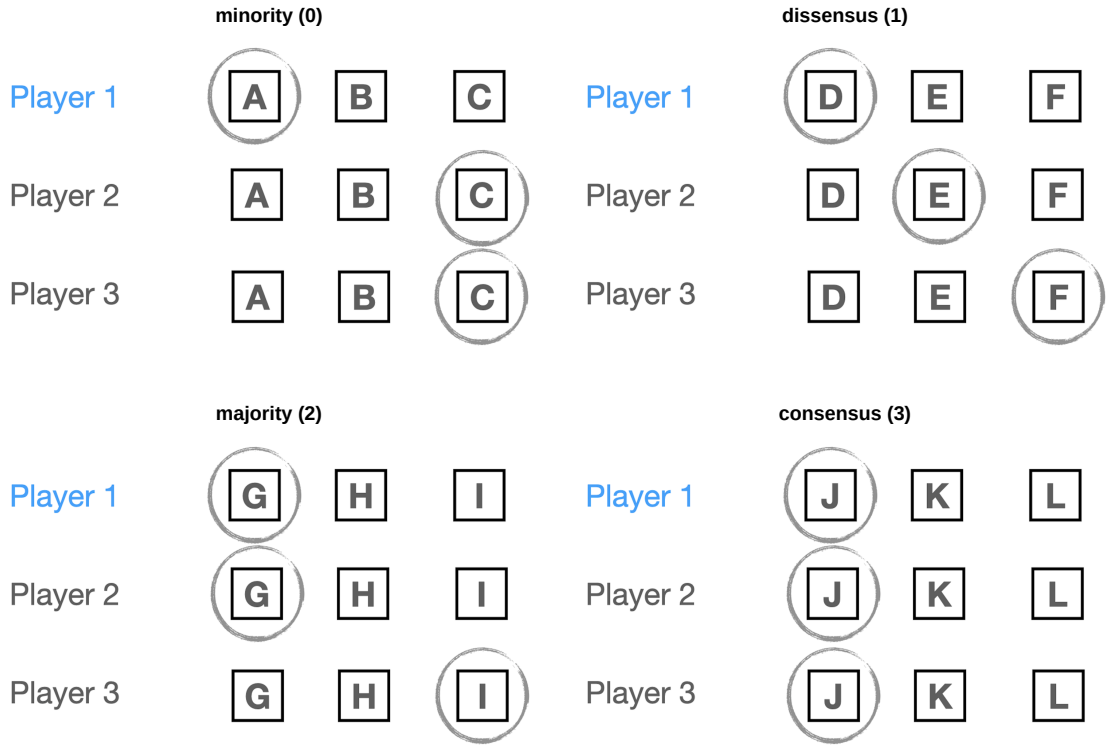


Figure 5. One set of stimuli by level of convergence, in Experiment 4 (similar stimuli are used in Experiments 5 and 6). A full set of stimuli can be found in the ESM, section F.

Design. We manipulated convergence within participants, by varying the ratio of players choosing the same response as a focal player (i.e. the one that participants evaluate). The levels of convergence are: (i) consensus, where all three players pick the same option [coded value = 3]; (ii) majority, where either the third or second player picks the same option as the first player [coded value = 2]; (iii) dissensus, where all three players pick different options [coded value = 1]; (iv) minority, where the second and third player pick the same option, but one that is different from the first player's choice [coded value = 0]. In our analysis, we treat convergence as a continuous variable, assigning the coded values in squared parenthesis here.

Materials. All participants saw all four conditions, with two stimuli per condition. Each participant therefore saw eight stimuli in total (4 convergence levels x 2 stimuli).

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept and a random slope for participants.

As in the numerical setting, we found a positive effect of convergence on both accuracy ($\hat{b}_{\text{Accuracy}}$ = 16.84 [15.009, 18.668], p = < .001; on a scale from 0 to 100) and competence ($\hat{b}_{\text{Competence}}$ = 0.68 [0.578, 0.788], p = < .001; on a scale from 1 to 7).

In the ESM (section F), we show that compared to what the normative models would predict, participants underestimate the effect of convergence on both accuracy and competence, but especially on accuracy.

In an exploratory, non-preregistered analysis, we tested whether the effect of convergence is larger on accuracy than on competence. To do so, we first standardized both outcome scores to account for the different scales. We then regressed the outcome score on convergence and its interaction with a binary variable indicating which outcome was asked for (accuracy or competence). We find a negative interaction, indicating that convergence had a smaller effect on competence than on accuracy ($\hat{b}$ = -0.10 [-0.136, -0.053], < .001; units in standard deviations).

Experiment 5

Experiment 5 is a conceptual replication of Experiment 3 in a categorical instead of a numerical case: are participants less likely to infer that more convergent estimates are more accurate, and the individuals who made them more competent, when the estimates are made by individuals with a conflict of interest pushing them to all provide a given answer, compared to when they are made by independent individuals? The independence condition of Experiment 5 also serves as a replication of Experiment 4, leading to the following hypotheses:

H1a: Participants perceive an estimate of an independent informant as more accurate the more it converges with the estimates of other informants.

H1b: Participants perceive an independent informant as more competent the more their estimate converges with the estimates of other informants.

H2a: The effect of convergence on accuracy (H1a) is more positive in a context where informants are independent compared to when they are biased (i.e. share a conflict of interest to pick a given answer).

H2b: The effect of convergence on competence (H1b) is more positive in a context where informants are independent compared to when they are biased (i.e. share a conflict of interest to pick a given answer).

Methods

Participants. We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found on the OSF. We ran two different power analyses, one for each outcome variable. We set the power threshold for both to 90%.

The power simulation for accuracy suggested that for 80 participants, we would have a power of at least 90% for the interaction effect. The simulation for competence suggested that with already 40 participants, we would detect an interaction, but only with 60 participants would we also detect an effect of convergence. Due to uncertainty about our assumptions and because resources were available for a larger sample, we recruited 200 participants, in the UK and via Prolific (99 female, 100, 1 non-identified; age_{mean}: 41.88, age_{sd}: 13.94, age_{median}: 39).

Procedure. After providing their consent to participate in the study and passing an attention check, participants read the following introduction: "We will show you three financial advisors who are giving recommendations on investment decisions. They can choose between three investment options. Their task is to recommend one. You will see several such situations. They are completely unrelated: it is different advisors evaluating different investments every time. At first you have no idea how competent the advisors are: they might be completely at chance, or be very good at the task. It's also possible that some are really good while others are really bad. Some tasks might be difficult while others are easy. Your task will be to evaluate the performance of one of the advisors based on what everyone's answers are." To assess perceptions of accuracy, we asked: "What do you think is the probability of advisor 1 making the best investment recommendation?". Participants answered with a slider on a scale from 0 to 100. To assess perceptions of competence, we asked: "How competent do you think advisor 1 is regarding such investment recommendations?" Participants answered on a 7-point Likert scale (from (1)"not competent at all" to (7)"extremely competent").

Design. We manipulated convergence within participants, and conflict of interest between participants. In the conflict of interest condition, experts were introduced this way: "The three advisors have already invested in one of the three options, the same option for all three. As a result, they have an incentive to push that option in their recommendations." Participants assigned to the independence condition read: "The three advisors are independent of each other, and have no conflict of interest in making investment recommendations."

Materials. We used the same stimuli as in Experiment 4. Identical to Experiment 4, participants saw all four convergence conditions, with two stimuli (i.e. expert predictions) per condition. Each participant therefore saw eight stimuli in total (4 convergence levels x 2 stimuli).

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept and a random slope for convergence for participants.

We find evidence for all four hypotheses (see Fig. 6). To test H1a and H1b, we

use the same analyses as in Experiment 4, restricted on the independence condition, and replicate the results. We find a positive effect of convergence on both accuracy ($\hat{b}_{\text{Accuracy}}$ = 12.34 [10.362, 14.311], p = < .001) and competence ($\hat{b}_{\text{Competence}}$ = 0.56 [0.459, 0.665], p = < .001).

The second set of hypotheses targeted the interaction of informational dependency and convergence (Fig. 6). In the independence condition, the effect of convergence on accuracy was more positive ($\hat{b}_{\text{interaction, Accuracy}}$ = 3.01 [0.027, 5.988], p = 0.048) than in the conflict of interest condition ($\hat{b}_{\text{baseline}}$ 9.33 [7.232, 11.426], p = < .001). Likewise, the effect of convergence on competence was more positive ($\hat{b}_{\text{interaction, Competence}}$ = 0.16 [0.014, 0.316], p = 0.032) than in the conflict of interest condition ($\hat{b}_{\text{baseline}}$ = 0.40 [0.291, 0.503], p = < .001).

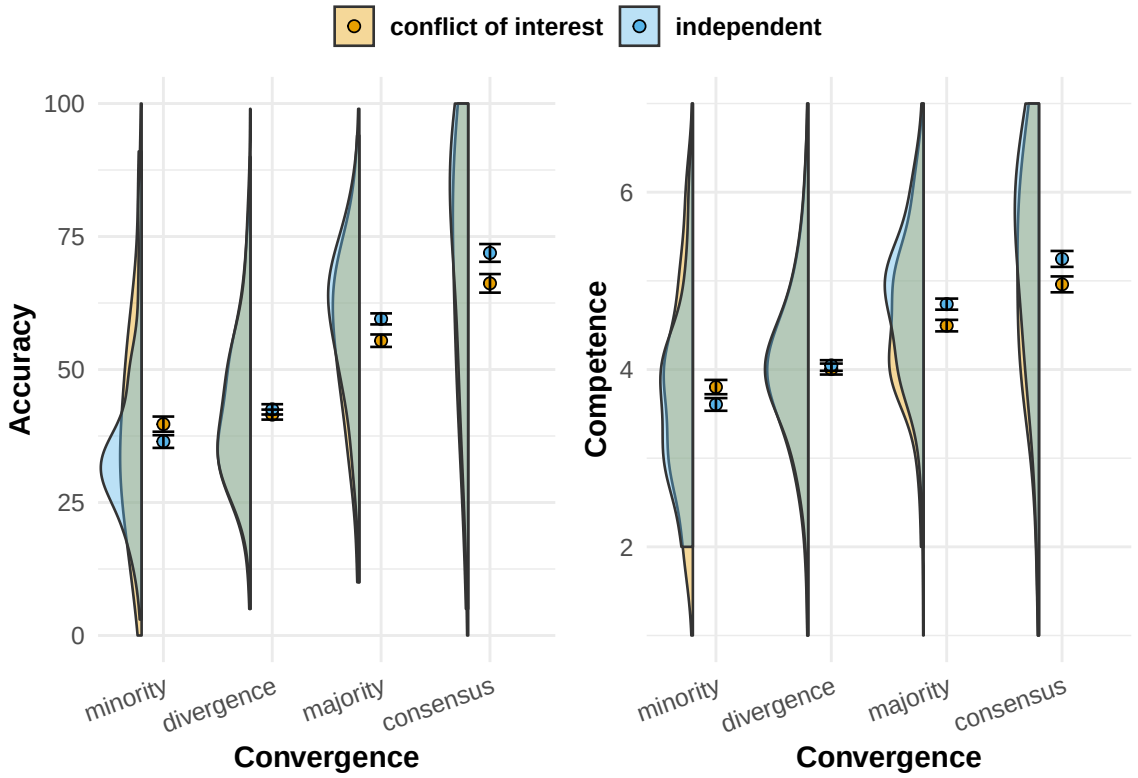


Figure 6. Interaction of convergence and informational dependency.

In an exploratory, non-preregistered analysis, we tested whether the effect of convergence is larger on accuracy than on competence. To do so, we first standardized both outcome scores to account for the different scales. We then regressed the outcome score on convergence and its interaction with a binary variable indicating which outcome was asked for (accuracy or competence), while controlling for informational dependency. We find a negative interaction, indicating that convergence had a smaller effect on competence than on accuracy ($\hat{b}$ = -0.08 [-0.113, -0.044], < .001; units in standard deviations).

Experiment 6

Experiment 6 is a replication and extension of Experiment 4 in which we test the effect of the number of choice options (three and ten, instead of only three). Our simulations suggested that, at least for some underlying population competence distributions, consensus should be more indicative of competence when there are more choice options, compared to fewer (see ESM, section B and H).

First, considering only the three options condition, we ran a direct replication of experiment 4. Second, following the results from our model, we predict that

H1: The effect of convergence on accuracy (H1a) is more positive in a context when informants can choose among ten response options compared to when they can choose among only three.

H2: The effect of convergence on competence (H1b) is more positive in a context when informants can choose among ten response options compared to when they can choose among only three.

Methods

Participants. We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found on the OSF. We used previous experiments and estimates of our models to inform our choice of parameter values. We ran two different power analyses, one for each outcome variable. We set the power threshold for our experiment to 90%. The power simulation for accuracy suggested that for 140 participants we would cross the power threshold of 90% for the interaction effect (power = 0.928). The simulation for competence suggested that with 300 participants, we would detect an interaction with a power of 87%. Due to budget constraints, we considered aiming for a sample of 300 participants as good enough, although slightly below our threshold. Due to two failed attention checks, our final sample consisted of 298 subjects, recruited, as in all experiments, in the UK and via Prolific (149 female, 149, 1 non-identified; age_{mean}: 42.09, age_{sd}: 13.06, age_{median}: 40).

Procedure. We used the same procedure as in Experiment 4, with the addition of one condition described below.

Design. The number of choice options was manipulated between participants. Participants were randomly assigned to either see stimuli with three options (as in Experiment 4), or stimuli with ten options. Participants assigned to the ten options condition were divided into one of two distinct sub-conditions: one in which the range of the answers corresponds to the range of the three options condition, and another with increased range (see ESM, section H). We found no differences between the two sub-conditions and collapsed them into a single ten options condition.

Materials. For the three options condition, we used the same stimuli as in Experiments 4 and 5. For the ten options condition, we created new sets of stimuli (see ESM, section H). Identical to Experiments 4 and 5, participants saw all four convergence conditions, with two stimuli per condition. Each participant therefore saw eight stimuli in total.

Results and discussion

To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept and a random slope for convergence for participants.

We replicate the results of experiment 4, but do not find evidence for an interaction between convergence and the number of choice options. To match the setting of experiment one, we reduced the sample to half of the participants, namely those who were assigned to the three options condition. On this reduced sample, we ran the exact same analyses as in experiment 4 and replicated the results. We find a positive effect of convergence on both accuracy ($\hat{b}_{\text{Accuracy}} = 15.68$ [14.112, 17.246], p = $< .001$) and competence ($\hat{b}_{\text{Competence}} = 0.65$ [0.564, 0.736], p = $< .001$). This finding also holds across the entire sample, pooling three and ten choice option observations ($\hat{b}_{\text{Accuracy}} = 16.30$ [15.124, 17.485], $< .001$; $\hat{b}_{\text{Competence}} = 0.68$ [0.611, 0.739], $< .001$).

We do not find evidence of an interaction, i.e. evidence that the number of choice options changes the effect of convergence ($\hat{b}_{\text{interaction, Accuracy}} = 1.25$ [-1.11, 3.613], 0.298; $\hat{b}_{\text{interaction, Competence}} = 0.05$ [-0.078, 0.178], 0.442).

We tested whether the effect of convergence is larger on accuracy than on competence in an exploratory, non-preregistered analysis. To this end, we first standardized both outcome scores to account for the different scales. We then regressed the outcome score on convergence and its interaction with a binary variable indicating which outcome was asked for (accuracy or competence), while controlling for the number of choice options. In line with all previous experiments, we find a negative interaction, indicating that convergence had a smaller effect on competence than on accuracy ($\hat{b} = -0.05$ [-0.075, -0.023], $< .001$).

Experiment 6 replicates Experiments 3 and 4 (independence condition), but suggests that the number of options the informants choose from does not powerfully affect participants' estimates of the informants' accuracy or competence. In the ESM, section H, we show that, compared to our model, participants underestimate the effect of the number of choice options.

General discussion

Using both analytical arguments and simulations, we have shown that, under a wide range of parameters, more convergent answers tend to be more accurate, and to have been made by more competent informants.

In two experiments (Experiment 1, and independence condition of Experiment 3), we find that participants presented with a set of more (rather than less) convergent numerical estimates find the estimates more accurate, and the individuals making the estimates more competent, thus drawing normatively justified inferences. Experiment 2 suggests that participants do not think that a discussion between the individuals makes their convergence less indicative of accuracy or their competence. By contrast, Experiment 3 reveals that, when the individuals making the estimates are systematically biased by a conflict of interest, participants put less weight on the convergence of their estimates.

Similar results are obtained in a categorical choice context, in which participants see the answers of individuals made within a limited set of options. Experiments 4, 5 (independence condition), and 6 show that, the more the answers converge, the more participants believe them to be accurate and the individuals who made to be competent, again drawing normatively justified inferences. Experiment 5 shows that these inferences are weakened when the convergence can be explained by a conflict of interest (as in Experiment 3). Experiment 6 fails to find an effect of the number of options.

We also observe that, in line with our simulations, participants draw stronger inferences from convergence to accuracy than to competence in five of the six experiments.

On the whole, participants thus appear to draw normatively justified inferences: 1. They infer that more convergent answers are more accurate; 2. They infer that more convergent answers are coming from more competent informants; 3. The inference on accuracy tends to be stronger than the inference on competence ; 4. Both the inference on accuracy and on the inference on competence are weaker when the informants have a conflict of interest, but not when they are merely discussing with each other. The only exception appears to be that participants do not take into account the number of options in categorical choices scenarios.

Conclusion

When people see that others agree with each other, they tend to believe that they are right. This inference has been evidenced in several experiments, both for numerical estimates (Budescu & Rantilla, 2000; Budescu et al., 2003; Budescu & Yu, 2007; e.g. Molleman et al., 2020; Yaniv et al., 2009), and for categorical choices (e.g., Morgan et al., 2012; for review, see H. Mercier & Morin, 2019). However, these experiments do not test whether this inference of accuracy of the information extends to an inference of competence of the informants. Moreover, by their design, participants arguably assumed a degree of competence among the informants. For instance, when children are confronted with several individuals who agree on how to name a novel object (e.g. Corriveau et al., 2009), they can assume that these (adult) individuals tend to know what the names of objects are. If a certain competence of the informants is assumed, then well-known results from the literature on judgment aggregation–the wisdom of crowds–show that the average opinion of a set of individuals is, in a wide range of circumstances, more likely to be accurate than that of a single individual (e.g. Larrick & Soll, 2006).

Here, we assumed no prior knowledge of individual competence, asking the question: if we see informants, whose competence is unknown, converge on an answer, is it rational to infer that this answer is more likely to be correct, and that the informants are likely to be competent? We have shown that the answer is yes on both counts–assuming there is no systematic bias among the informants. An analytical argument and a series of simulations revealed that, for both the numerical choice context and the categorical choice context, the more individuals agree on an answer, the more likely the answer is correct, and the more likely the individuals are competent, with the former effect being stronger than the latter. Moreover, this is true for a wide range of assumed population distributions of competence. In a series of experiments, we have shown that participants (UK) draw these inferences, but

that they do so less when there is reason to assume a bias among informants.

The results–both simulations and experiments–are a novel contribution to the wisdom of crowds literature. In this literature–in particular that relying on the Condorcet Jury Theorem–a degree of competence is assumed in the individuals providing some answers. From that competence, it can be inferred that the individuals will tend to agree, and that their answers will tend to be accurate. Here, we have shown that the reverse inference–from agreement to competence–is also warranted, and that it is warranted under a wide range of circumstances: If one does not suspect any systematic bias, convergence alone can be a valid cue when determining who tends to be an expert. This finding qualifies work suggesting that people need a certain degree of expertise themselves, in order to figure out who is an expert (Hahn, Merdes, & Sydow, 2018; Nguyen, 2020).

The present experiments show that people draw inferences from convergence to accuracy and to competence, but then do not precisely show how they do it. As suggested in the introduction, it is plausible that participants use the combination of two heuristics: one leading them from convergence to accuracy, and one from accuracy to competence. This two step process is coherent with the observation that the inference from convergence to accuracy (one step) is stronger than that from convergence to competence (two steps). However, our results are also compatible with participants performing two inferences from convergence, one to accuracy and one to competence, the latter being weaker because the inferences roughly follow the normative model. Our results are coherent with work on other mechanisms of epistemic vigilance that process the combination of several informants' opinions, showing that these mechanisms rely on sensible cues, but also systematically fail to take some subtle cues into account (H. Mercier & Morin, 2019).

A most prominent context in which the inferences uncovered here might play a role is that of science. Much of science is counterintuitive, and most people do not have the background knowledge to evaluate most scientific evidence. However, science is, arguably, the institution in which individuals end up converging the most in their opinions (on consensus being the defining trait of science by contrast with other intellectual enterprises, see Collins, 2002). For instance, scientists within many disciplines agree on things ranging from the distance between the solar system and the center of the galaxy to the atomic structure of DNA. This represents an incredible degree of convergence. When people hear that scientists have measured the distance between the solar system and the center of the galaxy, if they assume that there is a broad agreement within the relevant experts, this should lead them to infer that this measure is accurate, and that the scientists who made it are competent. Experiments have already shown that increasing the degree of perceived consensus among scientists tends to increase acceptance of the consensual belief (Van Stekelenburg, Schaap, Veling, Van 'T Riet, & Buijzen, 2022), but it hasn't been shown yet that the degree of consensus also affects the perceived competence of scientists.

In the case of science, the relationship between convergence and accuracy is broadly justified. However, at some points of history, there has been broad agreement on misbeliefs, such as when Christian theologians had calculated that the Earth was approximately six thousand years old. To the extent that people were aware of this broad agreement, and believed the theologians to have reached it independently of each other, this might have not

only fostered acceptance of this estimate of the age of the Earth, but also a perception of the theologians as competent.

The current study has a number of limitations. In our simulations, we assume agents to be independent and unbiased. Following previous work generalizing the Condorcet Jury Theorem to cases of informational dependency (Ladha, 1992), more robust simulations would show that–while still assuming no systematic bias–our results hold even when agents influence each others' answers. Regarding our experiments, if the very abstract materials allow us to remove most of the priors that participants might have, they might also reduce the ecological validity of the results. Although the main results replicate well across our experiments, and we can thus be reasonably certain of their robustness, it's not clear how much they can be generalized. Experimental results with convenience samples can usually be generalized at least to the broader population the samples were drawn from–here, UK citizens (Coppock, 2019). However, we do not know whether they would generalize to other cultures.

These limitations could be overcome by replicating the present results in different cultures, using more ecologically valid stimuli. For instance, it would be interesting to test whether the inference described here, from convergence to competence, might be partly responsible for the fact that people tend to believe scientists to be competent (Cologna et al., 2024). Finally, future studies could also attempt to systematically model cases of informational dependencies more subtle than those used in Experiments 3 and 5, to derive and test normative predictions.

Data availability. Data for all experiments and the simulations is available on the OSF project page (https://osf.io/6abqy/?view_only=42632bccea604fd7928dbe58e087d23b).

Code availability. The code used to create all results (including tables and figures) of this manuscript is also available on the OSF project page (https://osf.io/6abqy/?view_only=42632bccea604fd7928dbe58e087d23b).

Competing interest. The authors declare having no competing interests.

References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making*, *10*(3), 265–276. https://doi.org/10.1017/S1930297500004666

Bernard, S., Harris, P., Terrier, N., & Clément, F. (2015). Children weigh the number of informants and perceptual uncertainty when identifying objects. *Journal of Experimental Child Psychology*, *136*, 70–81. https://doi.org/10.1016/j.jecp.2015.03.009

Bernard, S., Proust, J., & Clément, F. (2015). Four- to Six-Year-Old Children's Sensitivity to Reliability Versus Consensus in the Endorsement of Object Labels. *Child Development*, *86*(4), 1112–1124. https://doi.org/10.1111/cdev.12366

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. https://doi.org/10.1016/S0001-6918(00)00037-8

Budescu, D. V., Rantilla, A. K., Yu, H.-T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Budescu, D. V., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177. https://doi.org/10.1002/bdm.547

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2013). Children Trust a Consensus Composed of Outgroup Members-But Do Not Retain That Trust. *Child Development*, *84*(1), 269–282. https://doi.org/10.1111/j.1467-8624.2012.01850.x

Collins, R. (2002). *The sociology of philosophies: a global theory of intellectual change* (4. print., 1. Harvard Univ. Pr. paperback ed., 2000). Cambridge, Mass. London: Belknap Press of Harvard Univ. Press.

Cologna, V., Mede, N. G., Berger, S., Besley, J., Brick, C., Joubert, M., … Linden, D. S. van der. (2024). *Trust in scientists and their role in society across 67 countries*. https://doi.org/10.31219/osf.io/6ay7s

Coppock, A. (2019). Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods*, *7*(3), 613–628. https://doi.org/10.1017/psrm.2018.10

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going With the Flow: Preschoolers Prefer Nondissenters as Informants. *Psychological Science*, *20*(3), 372–377. https://doi.org/10.1111/j.1467-9280.2009.02291.x

Crutchfield, R. S. (1955). Conformity and character. *American Psychologist*, *10*(5), 191–198. https://doi.org/10.1037/h0040237

De Condorcet, N. (2014). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press.

Desai, S. C., Xie, B., & Hayes, B. K. (2022). Getting to the source of the illusion of consensus. *Cognition*, *223*, 105023. https://doi.org/10.1016/j.cognition.2022.105023

Dietrich, F., & Spiekermann, K. (2013). Epistemic Democracy with Defensible Premises. *Economics and Philosophy*, *29*(1), 87–120. https://doi.org/10.1017/S0266267113000096

Dietrich, F., & Spiekermann, K. (2024). Deliberation and the wisdom of crowds. *Economic Theory*. https://doi.org/10.1007/s00199-024-01595-4

Einav, S. (2018). Thinking for themselves? The effect of informant independence on children's endorsement of testimony from a consensus. *Social Development*, *27*(1), 73–86. https://doi.org/10.1111/sode.12264

Fusaro, M., & Harris, P. L. (2008). Children assess informant reliability using bystanders' non-verbal cues. *Developmental Science*, *11*(5), 771–777. https://doi.org/10.1111/j.1467-7687.2008.00728.x

Galton, F. (1907). Vox Populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Hahn, U., Merdes, C., & Sydow, M. von. (2018). How Good Is Your Evidence and How Would You Know? *Topics in Cognitive Science*, *10*(4), 660–678. https://doi.org/10.1111/tops.12374

Harkins, S. G., & Petty, R. E. (1987). Information utility and the multiple source effect. *Journal of Personality and Social Psychology*, *52*(2), 260.

Harries, C., Yaniv, I., & Harvey, N. (2004). Combining advice: the weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, *17*(5), 333–348. https://doi.org/10.1002/bdm.474

Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review*, *112*(2), 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Herrmann, P. A., Legare, C. H., Harris, P. L., & Whitehouse, H. (2013). Stick to the script: The effect of witnessing multiple actors on children's imitation. *Cognition*, *129*(3), 536–543. https://doi.org/10.1016/j.cognition.2013.08.010

Hess, N. H., & Hagen, E. H. (2006). Psychological adaptations for assessing gossip veracity. *Human Nature*, *17*(3), 337–354. https://doi.org/10.1007/s12110-006-1013-z

Jayles, B., Kim, H., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., ... Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, *114*(47), 12620–12625. https://doi.org/10.1073/pnas.1703695114

Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*, *10*(2), 107–137. https://doi.org/10.1037/dec0000199

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in Testimony: Children's Use of True and False Statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Ladha, K. K. (1992). The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, *36*(3), 617. https://doi.org/10.2307/2111584

Larrick, R. P., & Soll, J. B. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

Lopes, D., Vala, J., & Garcia-Marques, L. (2007). Social validation of everyday knowledge: Heterogeneity and consensus functionality. *Group Dynamics: Theory, Research, and Practice*, *11*(3), 223–239. https://doi.org/10.1037/1089-2699.11.3.223

Mannes, A. E. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science*, *55*(8), 1267–1279. https://doi.org/10.1287/mnsc.1090.1031

Mercier, H. (2016). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Mercier, Hugo. (2020). *Not born yesterday: the science of who we trust and what we believe*. Retrieved from https://doi.org/10.1515/9780691198842

Mercier, H., & Claidière, N. (2022). Does discussion make crowds any wiser? *Cognition*, *222*, 104912. https://doi.org/10.1016/j.cognition.2021.104912

Mercier, Hugo, & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314. https://doi.org/10.1016/j.evolhumbehav.2019.01.001

Mercier, H., & Morin, O. (2019). Majority rules: how good are we at aggregating convergent

opinions? *Evolutionary Human Sciences*, *1*, e6. https://doi.org/10.1017/ehs.2019.6

Molleman, L., Tump, A. N., Gradassi, A., Herzog, S., Jayles, B., Kurvers, R. H. J. M., & Bos, W. van den. (2020). Strategies for integrating disparate social information. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1939), 20202413. https://doi.org/10.1098/rspb.2020.2413

Morgan, T. J. H., Laland, K. N., & Harris, P. L. (2015). The development of adaptive conformity in young children: effects of uncertainty and consensus. *Developmental Science*, *18*(4), 511–524. https://doi.org/10.1111/desc.12231

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 653–662. https://doi.org/10.1098/rspb.2011.1172

Nguyen, C. T. (2020). Cognitive islands and runaway echo chambers: problems for epistemic dependence on experts. *Synthese*, *197*(7), 2803–2821. https://doi.org/10.1007/s11229-018-1692-0

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology*, *47*(1), 116–129. https://doi.org/10.1016/0022-0965(89)90066-0

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Romeijn, J.-W., & Atkinson, D. (2011). Learning juror competence: a generalized Condorcet Jury Theorem. *Politics, Philosophy & Economics*, *10*(3), 237–262. https://doi.org/10.1177/1470594X10372317

Smith, J. M., & Harper, D. (2003). *Animal signals*. Oxford University Press.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. https://doi.org/10.1037/a0015145

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Van Stekelenburg, A., Schaap, G., Veling, H., Van 'T Riet, J., & Buijzen, M. (2022). Scientific-Consensus Communication About Contested Science: A Preregistered Meta-Analysis. *Psychological Science*, *33*(12), 1989–2008. https://doi.org/10.1177/09567976221083219

Xie, B., & Hayes, B. (2022). Sensitivity to Evidential Dependencies in Judgments Under Uncertainty. *Cognitive Science*, *46*(5). https://doi.org/10.1111/cogs.13144

Yaniv, I. (1997). *Weighting and Trimming: Heuristics for Aggregating Judgments under Uncertainty*. 13.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 558–563. https://doi.org/10.1037/a0014589

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Dis-

counting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Yin, J., Xu, Z., Lin, J., Zhou, W., & Guo, X. (2024). Smartly following others: Majority influence depends on how the majority behavior is formed. *Journal of Experimental Social Psychology*, *115*, 104644. https://doi.org/10.1016/j.jesp.2024.104644

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. https://doi.org/10.1177/0956797619856844

Appendix A
Analytical argument

Categorical choice

Inferring population average competence. There are $n+1$ individuals, who choose among $m+1$ answers. There is one good answer and m wrong ones. We denote $\bar{p}$ the probability that a random individual is right. The probability to be wrong is $1-\bar{p}$, and the probability to choose a particular wrong answer is $\frac{1-\bar{p}}{m}$. We denote this last probability $\bar{q}$ (it is common in probabilities to use $q = 1-p$, here note that it is not the case).

We assume that the population is quite large, so that it is very likely to be the biggest group of answers. We want to estimate the average competence $\bar{p}$ from the size of the majority K. To simplify, we concentrate on situations where either $\bar{p}$ is quite large compared to $\bar{q}$, or the population quite large. In this situation, it is almost certain that the biggest group is made of individuals who chose the right answer and a proportion $\bar{p}$ of the population should, on average, choose this answer. The size $\frac{K}{n+1}$ should thus be a good estimator of $\bar{p}$.

Figure A1 illustrates this logic. When the population is competent enough or large enough (e.g. when $n = 100$), the average size of the biggest group closely aligns with the average competence. When $n = 10$ and average competence is low, it is frequent that some individuals converge by chance on another answer, nd the size of the majority is is therefore on average larger than $\bar{p}$, hence an overestimate of its members competence.

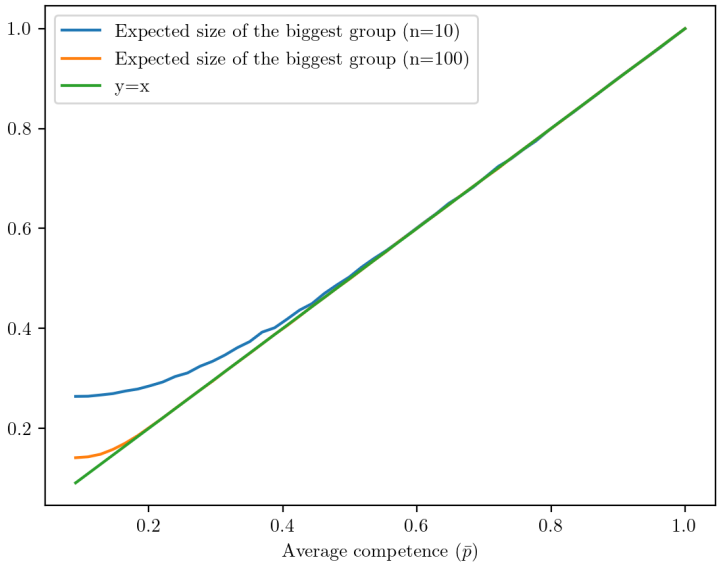


Figure A1. Expected size of the majority depending on average competence, with $n = 10$ (blue) and $n = 100$ (orange).

Inferring within population competence. Here, we focus on one individual and we aim to infer his personal level of competence p. We observe the number K of individuals

who agree with him. Either the focus individual is right (which happens with probability p) and K follows a binomial distribution of parameters n and $\bar{p}$; or the individual is wrong (which happens with probability $1-p$) and then K follows a binomial distribution of parameters n and $\bar{q}$. Thus, K follows a mixture distribution of different parameters, with weights p and $1-p$.

Using the mixture distributions property, we can compute the mean of K:

$$E(K) = n(p\bar{p} + (1-p)\bar{q}) = n(p(\bar{p} - \bar{q}) + \bar{q})$$

Clearly, $E(K)$ increases with p if and only if $\bar{p} - \bar{q} > 0$. In other words, if the population is better than chance, then we expect that the more competent an individual is, the more people agree with him.

But now, what exactly can we infer from p by observing K? To answer that, we need to assume a distribution of competence levels in the population, that we will also use as a prior. We assume that in the competence levels are distributed as a beta distribution, of parameters α and β. Their density is $f(x) = x^{\alpha}(1-x)^{\beta}$. Then, $\bar{p} = \frac{\alpha}{\alpha+\beta}$, the mean of our beta distribution.

Using Bayes formula, we can write the posterior distribution of p:

$$\Pi(x|K=k) \propto f(x)P(K=k|p=x) = f(x)(x\bar{p}^k(1-\bar{p})^{n-k} + (1-x)\bar{q}^k(1-\bar{q})^{n-k})$$

$$= x^{\alpha+1}(1-x)^{\beta}[\bar{p}^k(1-\bar{p})^{n-k}] + x^{\alpha}(1-x)^{\beta+1}[\bar{q}^k(1-\bar{q})^{n-k}]$$

We recognise a mixture distribution between two beta distributions, with parameters $\alpha+1$ and β for the first one, α and $\beta+1$ for the second. The first one has a bigger mean than our original Beta(α,β) distribution, the second a lower one (see Fig. A2). In a way, the bayesian updating shifts the distribution to the right or to the left, depending on K.

The weightings $(\bar{p}^k(1-\bar{p})^{n-k}$ and $\bar{q}^k(1-\bar{q})^{n-k}$ correspond to the probability that k individuals choose the right answer, and a particular wrong answer, respectively (we took out the binomial coefficients, who are equal for both answer types). Since $\bar{p} > \bar{q}$, the former is much stronger for large k, and the latter much stronger for low k. For $k = n$ (i.e. when the focus individual is part of a consensus), the ratio of weighting if $(p/q)^n$, which is very large when n is large or when $\bar{p} \gg \bar{q}$. In this case, for high k, the posterior distribution is well approximated by a Beta($\alpha+1$,β). Conversely, when k is close to zero (and when the other conditions are present), the posterior distribution is well approximated by a Beta(α,$\beta+1$).

We can plot the mean of the posterior distribution, which can be interpreted as the average estimation we can make of the individual competence, depending on K (Fig. A3). For low n and m (the main experiment used $n = m = 2$, the function is not too steep. For higher values, it gets sigmoidal, or even a jump function.

How much we learn from the observation depends crucially on α and β, that is, on the prior distribution of competence in the population. The range of possible mean estimates is $\frac{1}{\alpha+\beta+1}$, it is larger when α and β are low.

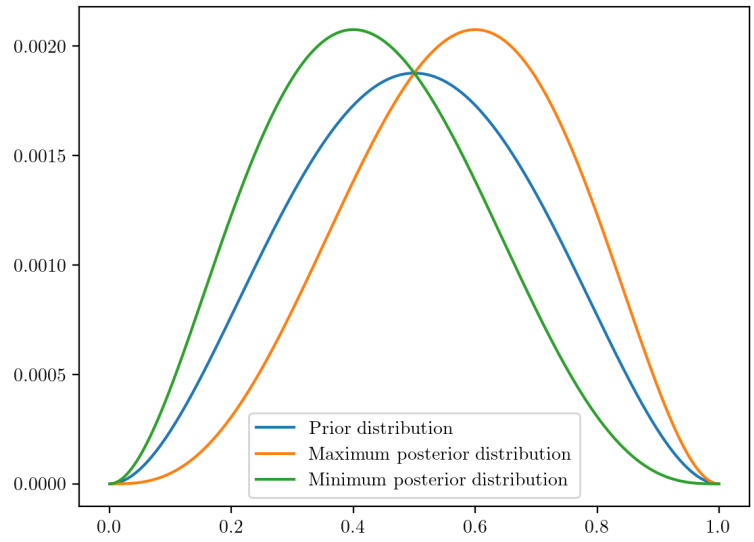


Figure A2. Prior and boundary posterior distributions. These distributions are the limit distributions when the individual is either part of a consensus or a dissensus, and when $\bar{p} \gg \bar{q}$ or n is large. Here, $\alpha = \beta = 2$ and we do not need to define the other parameters. As we explain below, the prior distribution determines the range between the possible posterior distributions.

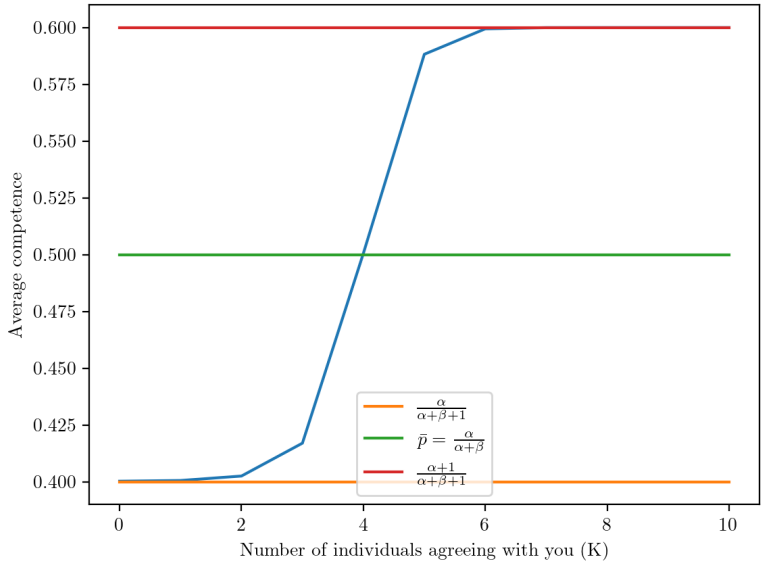


Figure A3. Mean of the posterior competence distribution, depending on the number of individuals agreeing with the focus individual. Here, $n = m = 10$.

For instance, let us set $\alpha = \beta$ (which implies $\bar{p} = \frac{1}{2}$). Now, we vary the two shape parameters from close to 0 (beta distributions need positive shape parameters) to a large number. This shifts the distribution from a bimodal one (with $\alpha = \beta \approx 0$, the distribution is perfectly bimodal, with half of the population always right and the other half always right) to a sharp unimodal distribution (when α and β are large, the whole population has basically a competence of .5). The more bimodal the distribution, the larger the range of possible estimates (Fig. A4). In the limit, when α and β are close to 0, we can infer that an individual is either perfectly competent or perfectly incompetent, depending on whether they are part of a consensus or alone in their opinion. On the contrary, when α and β are large, we do not change our estimate at all as the distribution is concentrated around .5. Intuitively, this makes sense: If we assume everyone is equally (in)competent, with no exception, we wouldn't infer anything from observing a convergence.

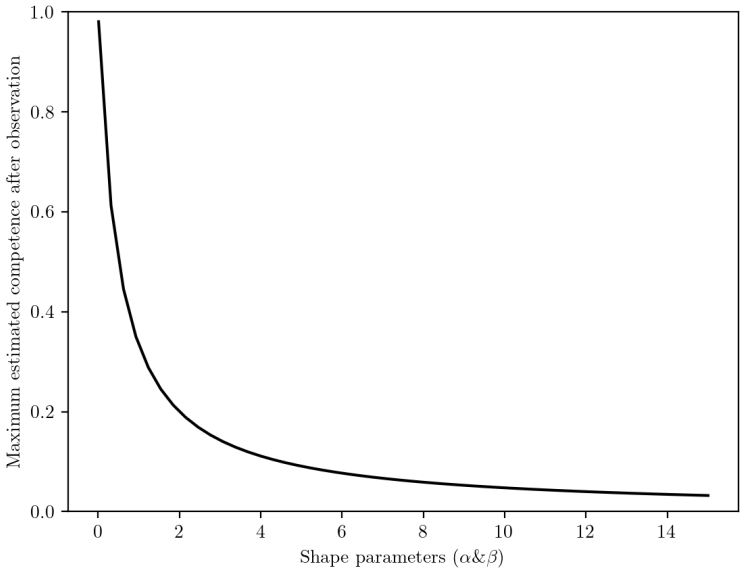


Figure A4. Range of possible estimates (difference between the minimum and the maximum estimate), depending on the shape of the distribution (here, we assume that $\alpha = \beta$).

Inferring accuracy. Now, instead of inferring competence, we want to infer accuracy, i.e. the probability that an individual answer is correct.

Again, using Bayes formula, we have:

$$p(right|K = k) = \frac{p(K = k|right)p(right)}{p(K = k)} = \frac{\bar{p}(\bar{p}^{k-1}(1-\bar{p})^{n-k})}{((\bar{p}^k(1-\bar{p})^{n-k}) + (\bar{q}^k(1-\bar{q})^{n-k}))}$$

$$= \frac{1}{1 + (\bar{p}/\bar{q})^k \left(\frac{1-\bar{q}}{1-\bar{p}}\right)^{n-k}}$$

Compared to competence, it is clear (Figure A5) that accuracy can be inferred much more strongly.

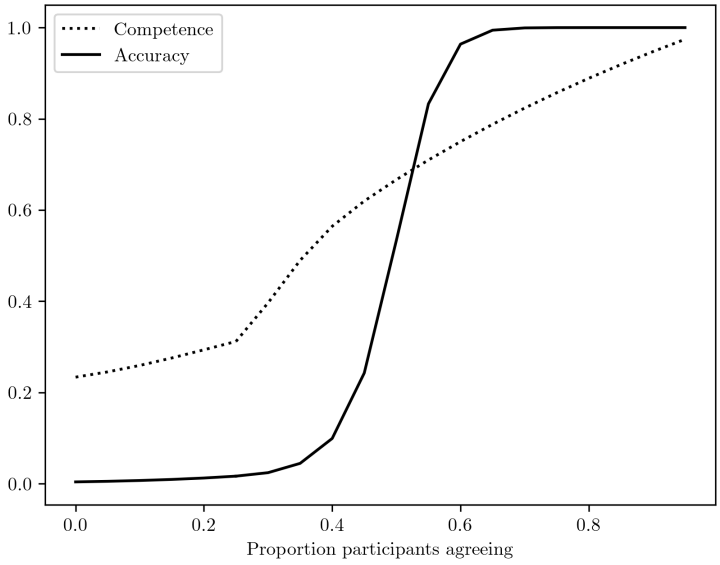


Figure A5. Estimated competence and accuracy depending on the degree of consensus. Here, we assume that $\alpha = 1$, $n = 20$ and $m = 4$. β is inferred from α and the majority size.

Continuous choices

A group of N individuals makes a prediction. The right answer is θ, we assume their answers $x_i, i \in 1, \ldots, N$ are normally distributed around θ, with a variance σ^2 reflecting the group incompetence (so the lower σ^2, the more the group can be said to be competent).

We estimate the degree of divergence through the unbiased sample variance $\hat{\sigma}^2$, defined as the average squared distance to the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

It is well known that $\hat{\sigma}^2$ (when coupled with the sample mean) is a "sufficient" estimator of σ^2, that is, it contains all the available information about σ^2. It follows, from the Lehmann-Scheffé theorem, that it is also the minimum-variance unbiased estimator, that is, the most precise way to estimate σ^2 while staying unbiased.

Appendix B
Simulations

Are we justified in inferring competence and accuracy from convergence? To complement our analytical argument, we ran simulations, intended to mirror our experimental setup. We find that–under certain conditions–more convergent groups indeed tend to be more competent and accurate. In this appendix, we describe these simulations in detail.

Numerical choice context

When several people estimate a quantity (numeric scenario), their convergence can be measured for example by the empirical variance. The closer the estimates, i.e. the smaller the empirical variance, the greater convergence. This measure is at the group level.

To provide a normative answer, we ran simulations for a scenario in which individuals provide an estimate on a scale from 1000 to 2000. In our simulations, we suppose that an individual's answer is drawn from a normal distribution. Each individual has their own normal distribution. All normal distributions are centered around the true answer - but they differ in their standard deviations. The value of that standard deviation is what we define as an individual's competence. The lower the standard deviation, the higher the competence, i.e. the more likely a guess drawn from the normal distribution will be close to the true answer. We (arbitrarily) define a range of competence: we set the lowest competence equal to the range of possible values, i.e.the largest standard deviation (2000 - 1000) = 1000. We set the highest competence to 0.1% of the range of possible values, i.e.the smallest standard deviation (0.001 x 1000) = 1 (see Fig. B1).

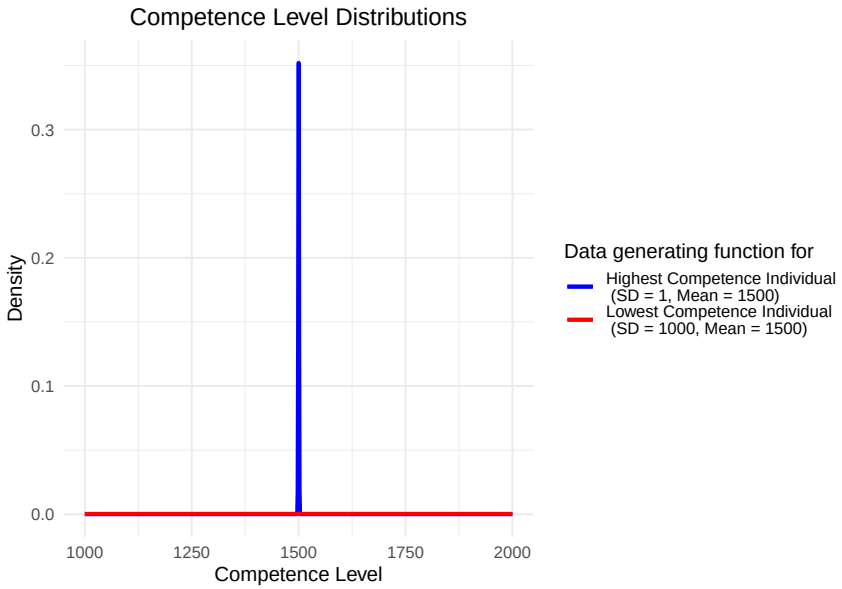


Figure B1. Range of possible data generating functions for individuals.

We suppose that individual competence levels are drawn from a competence distribution, which can be expressed by a beta distribution. This competence distribution can

take vary different shapes, depending on the alpha and beta parameters that describe the distribution (see Fig. 2).

We draw an estimate for each individual based on their respective competence distribution. For each individual, we then measure accuracy as the (squared) distance between their estimate and the true answer. Having a competence and an accuracy value for each individual, we randomly assign individuals to groups of three. For each group, we calculate the average of the three individuals' competence and accuracy. We measure the convergence of a group by calculating the standard deviation of the estimates. We run this simulation on a sample size of roughly 99900 (varying slightly as a function of sample size). We repeat this simulation process for various sample sizes and competence distributions. The results are displayed in Fig. B7 for accuracy, and Fig. B8 for competence. Across all underlying competence distributions, we find a positive correlation between convergence and accuracy, which tends towards 1 as sample size increases (see Fig. B2). As for competence, we find a positive correlation between convergence and competence across all underlying competence distributions. However, these correlations are weaker than for accuracy, and do not increase with sample size (see Fig. B2).

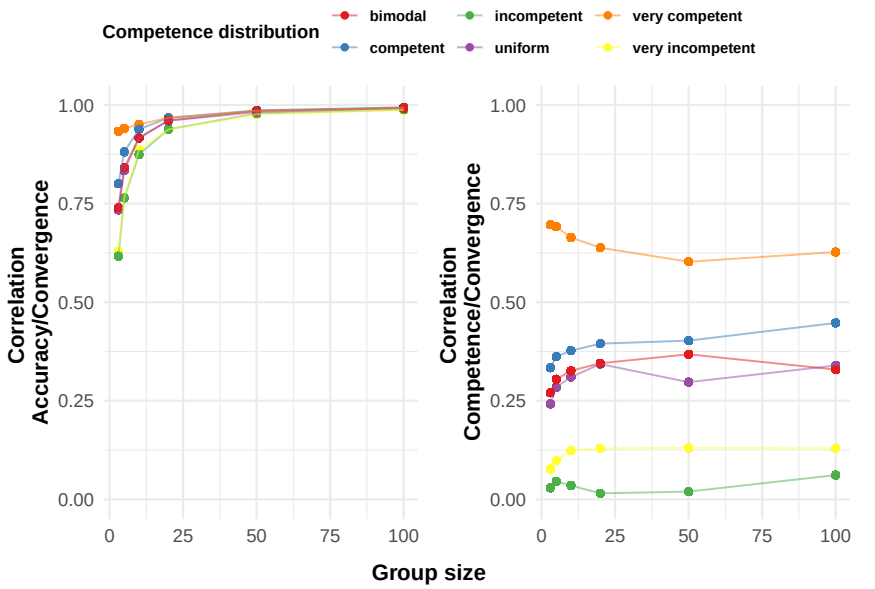


Figure B2. Correlation between the convergence of the answers of a group of agents and (left) the average accuracy of their answers; (right) the average competence of the agents, as a function of how many agents are in the group, and of the competence distribution of the agents (see Fig. 2).

Categorical choice context

When people make a choice based on several categories, their answers cannot be ranked by their nature (i.e. they are nominal, not ordinal), and that there are fewer of them (e.g. one of three possible products to choose, instead of an estimate between one and two thousand). In this case, convergence can be measured by the share of people agreeing on an

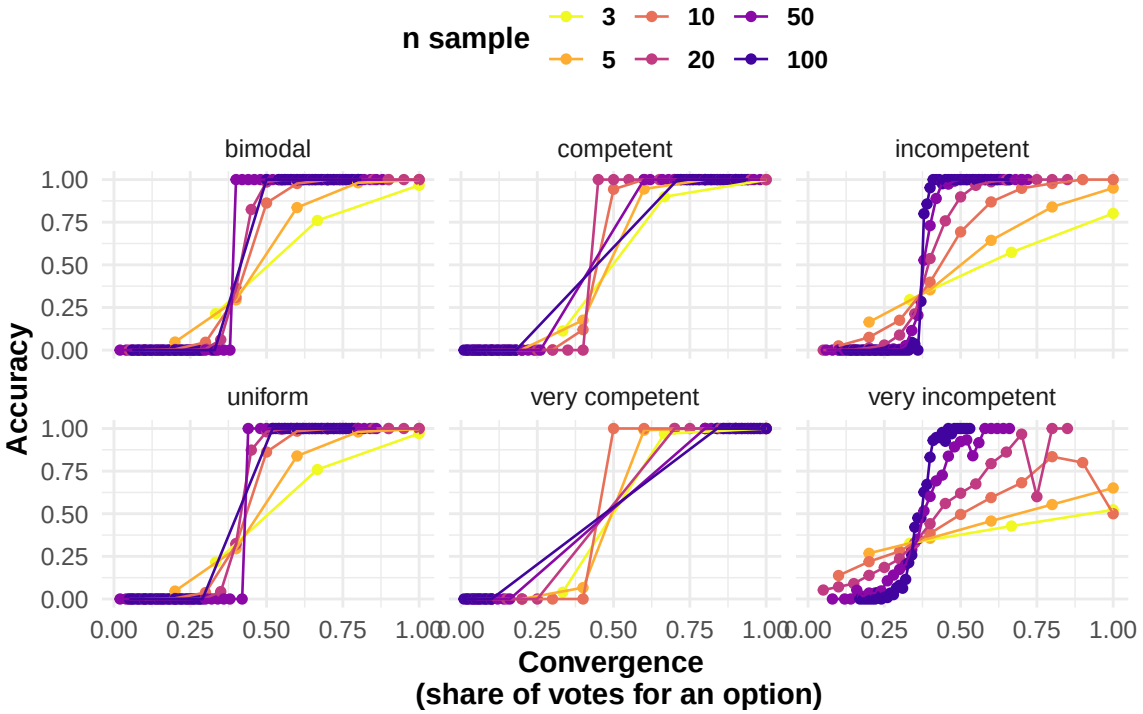


Figure B3. Accuracy as a function of convergence, for different population competence distributions and sample sizes. The number of choice options is three.

answer. The larger the share of informants agreeing on an answer, the greater convergence. This measure is at the response level, nested within the group level.

As for the numeric scenario, we ran simulations to provide a normative answer as to whether it is justified to infer accuracy and competence from greater convergence. We, again, suppose that an individual's answer is drawn from an internal distribution - in this case, a multinomial distribution, that describes how likely the individual is to choose each available option. If there are m choice options, an individual has the probability p of picking the right one, and the probability of (1-p)/(m-1) to pick any other option. Each individual has their own multinomial distribution. We define competence as the probability of making the correct choice. The higher the competence, the greater the probability that an individual will choose the correct option. Competence values range from being at chance for selecting the correct option ($p = 1/m$) to certainly selecting the correct option ($p = 1$). As before in the numeric case, suppose that individual competence levels are drawn from a competence distribution, which can be expressed by a beta distribution (see Fig. 2). Based on their competence level, we draw an estimate for each individual. We measure an individual's accuracy as a binary outcome, namely whether they picked the correct option, or not. We then randomly assign individuals to groups of informants (we vary the sample size from one simulation to another). Within these groups, we calculate the share of individuals voting for an answer option. For example, in a scenario in which three individuals pick

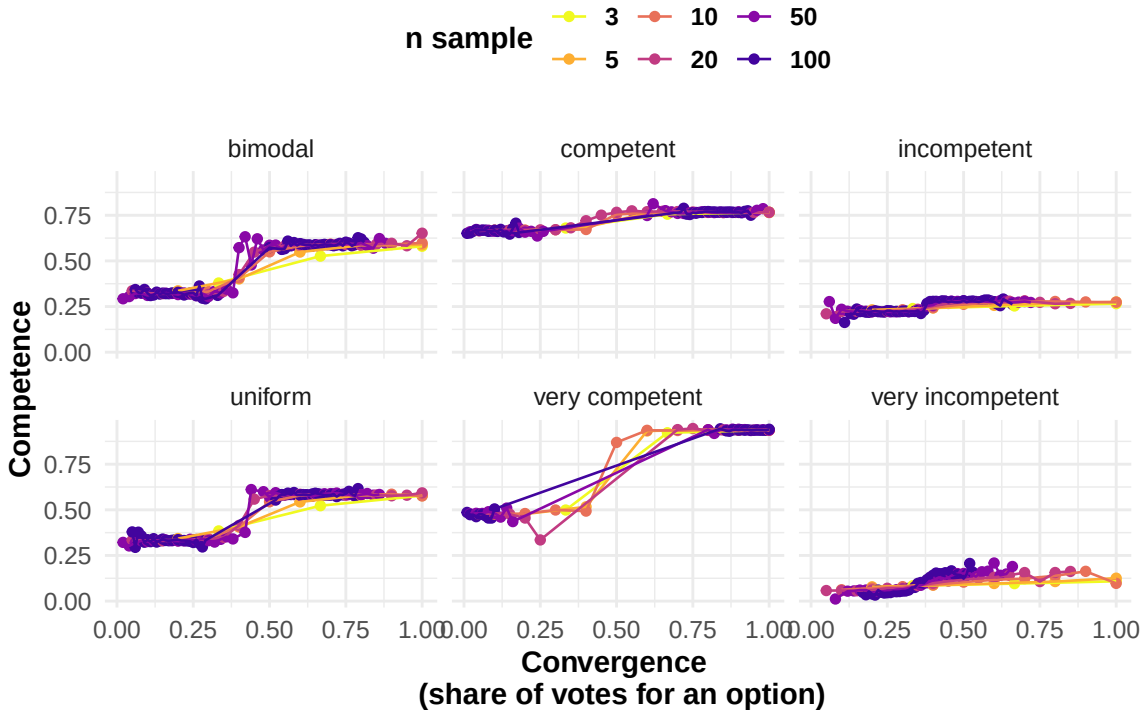
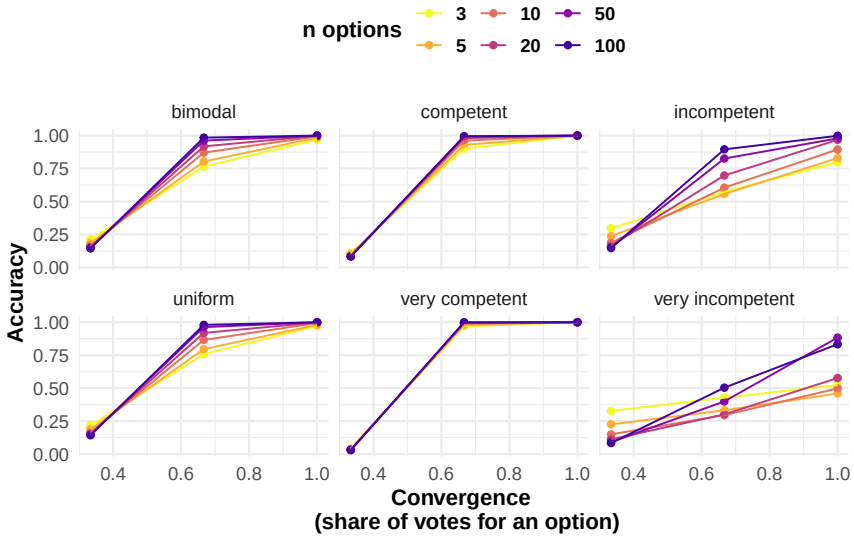


Figure B4. Competence as a function of convergence (i.e. vote share for an option), for different population competence distributions and sample sizes. The number of choice options is three.

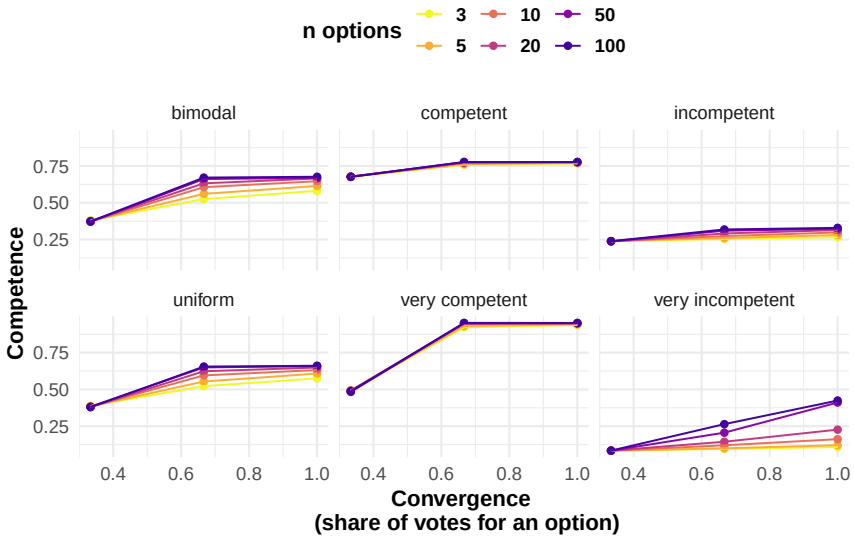
among three options (A, B and C), two individuals might vote C and one B. In this case we obtain an average accuracy and an average competence value for a share of 2/3 (option C) and for a share of 1/3 (option B). We simulate this on a population of 99900 individuals. We repeat this procedure varying the underlying population competence distributions, and additionally varying either (a) the sample size of informants, or (b) the number of choice options. If we vary the sample size, we hold the number of choice options constant at n = 3, and vice versa when varying the number of choice options. Fig. B3 shows the average accuracy, and Fig. B4 the average competence value for each share of votes, for different competence levels and varying the sample size. Fig. B5 and Fig. B6 display the same relationship, but varying the number of choice options instead. The figures display that, across all sample sizes and competence levels, the larger the share of votes for an option, the more accurate the option is on average. That relationship appears to follow some sigmoid curve which switches from an average accuracy of 0 to an average accuracy of 1 before a share of 0.5 is attained, and which is steeper for larger sample sizes. For competence, we observe a similar sigmoid-like relationship, but of lesser amplitude and varying considerably as a function of the underlying population competence distributions.

In sum, given the set of specific assumptions we made, our simulations suggest that people are indeed justified in inferring accuracy and competence from convergence in both



Sample size of informants: 3; Iterations per sample size: 100; Population size per iteration: 999

Figure B5. Accuracy as a function of vote share for an option, for different population competence distributions and number of choice options. Points represent averages across all simulations within the respective segment. The sample size is three.



Sample size of informants: 3; Iterations per sample size: 100; Population size per iteration: 999

Figure B6. Competence as a function of vote share for an option, for different population competence distributions and number of choice options. Points represent averages across all simulations within the respective segment. The sample size is three.

numeric and categorical choice settings.

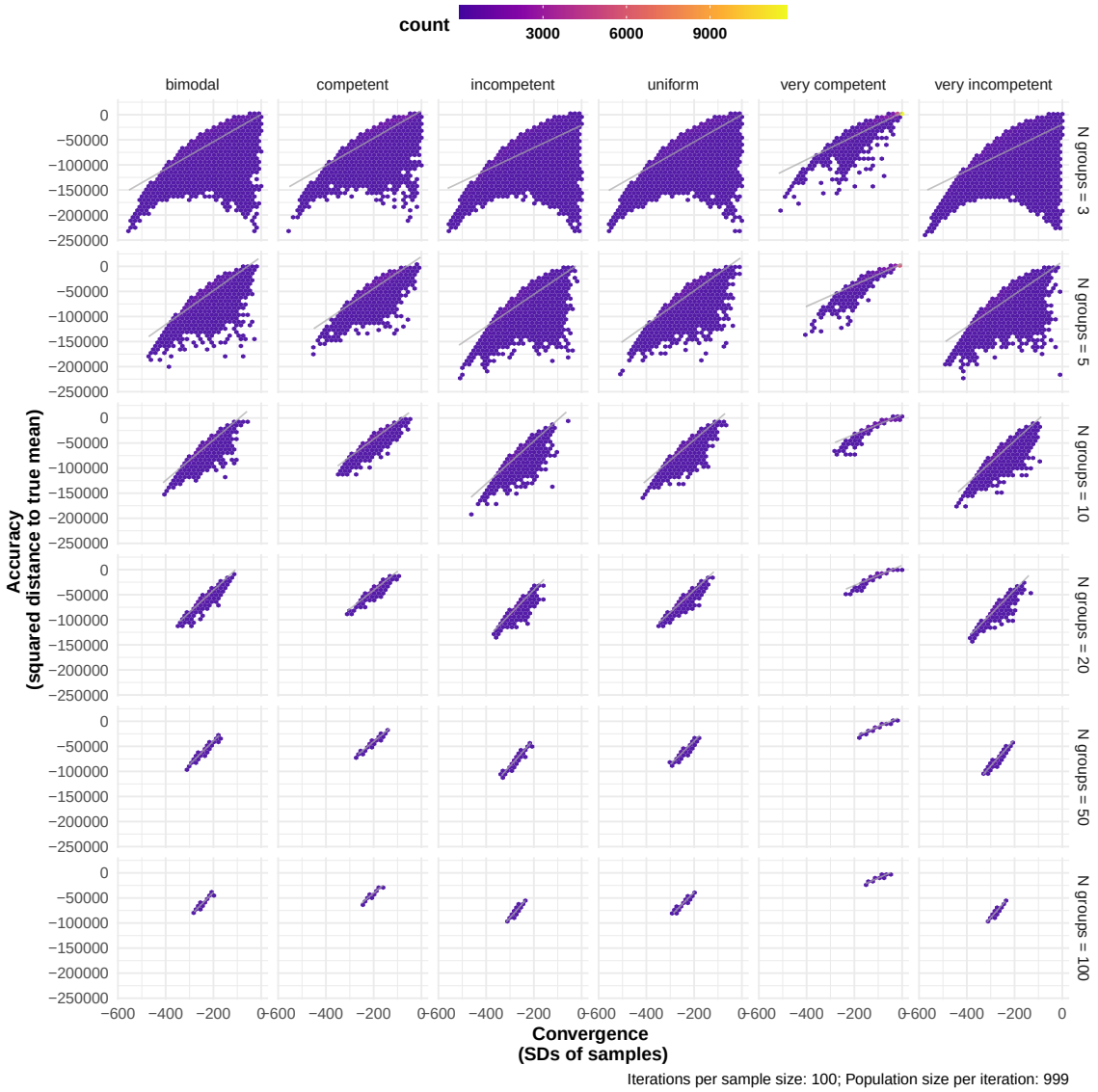


Figure B7. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.

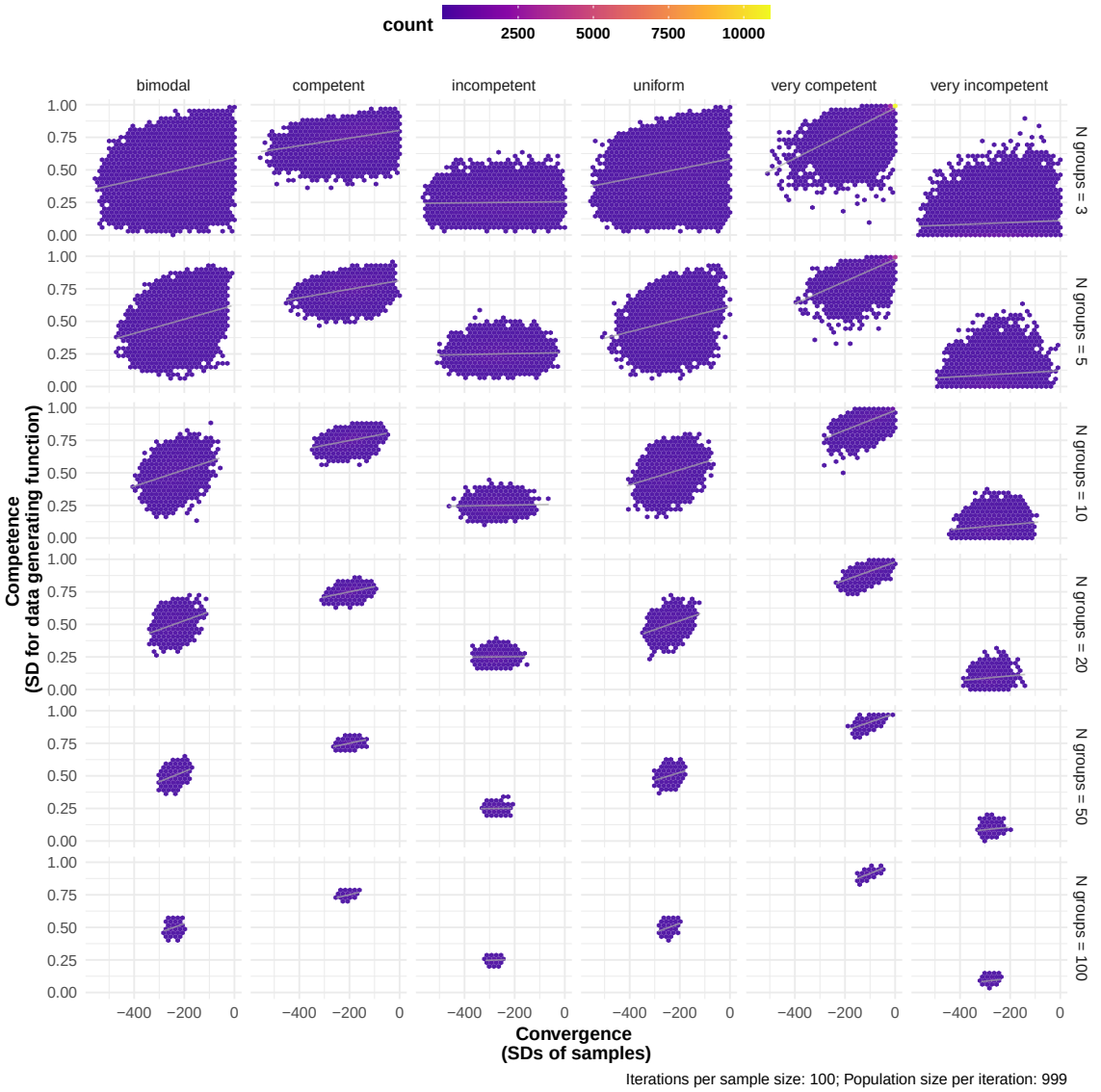


Figure B8. Simulation results showing the relationship between convergence and competence for different population competence distributions.

Appendix C
Experiment 1

Stimuli

The means of the normal distributions that we draw our estimates from were distinct between sets of estimates. Considering our within-participant design, we wanted to ensure that participants understood each set of estimates as being the result of a different, unrelated game. In order to assure that random draws from the distributions will (most likely) appear on the response scale (1000 to 2000), we constrained the means of all normal distributions to lie between the first and third quartile of the response scale (i.e. smallest possible mean was 1225 and largest 1775). We define a set of eight means–one for each set of estimates– that cover the range from first to third quartile of the predefined scale with an equal interval (1250, 1325, 1400, 1475, 1550, 1625, 1700, 1775). We randomly paired means with conditions when generating the stimuli. We then drew the set of estimates from the respective normal distributions given the assigned means and the condition constraints. We repeated this three times, resulting in three different series of eight sets of estimates. We randomly assign participants to one of these series. Additionally, for each participant, we randomize the order of appearance of the sets of estimates within the respective series. Images of all sets of estimates can be found on the OSF.

Attention check

Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

Do you agree with your friend? (Yes/No)

Results

Figure C1 visualizes the results and table C1 contains descriptive results. Figure C2 visualizes score differences between the two outcomes, accuracy and competence.

Research questions

We had three additional research questions regarding the number of informants:

RQ1: Do H1 and H2 hold for both a small [3] and a large [10] number of estimates?

Table C1

Convergence	Number	Accuracy	Competence
divergent	small	3.152 (sd = 1.495)	3.57 (sd = 1.341)
divergent	large	3.232 (sd = 1.282)	3.465 (sd = 1.184)
convergent	small	4.425 (sd = 1.461)	4.695 (sd = 1.221)
convergent	large	4.695 (sd = 1.424)	4.805 (sd = 1.251)

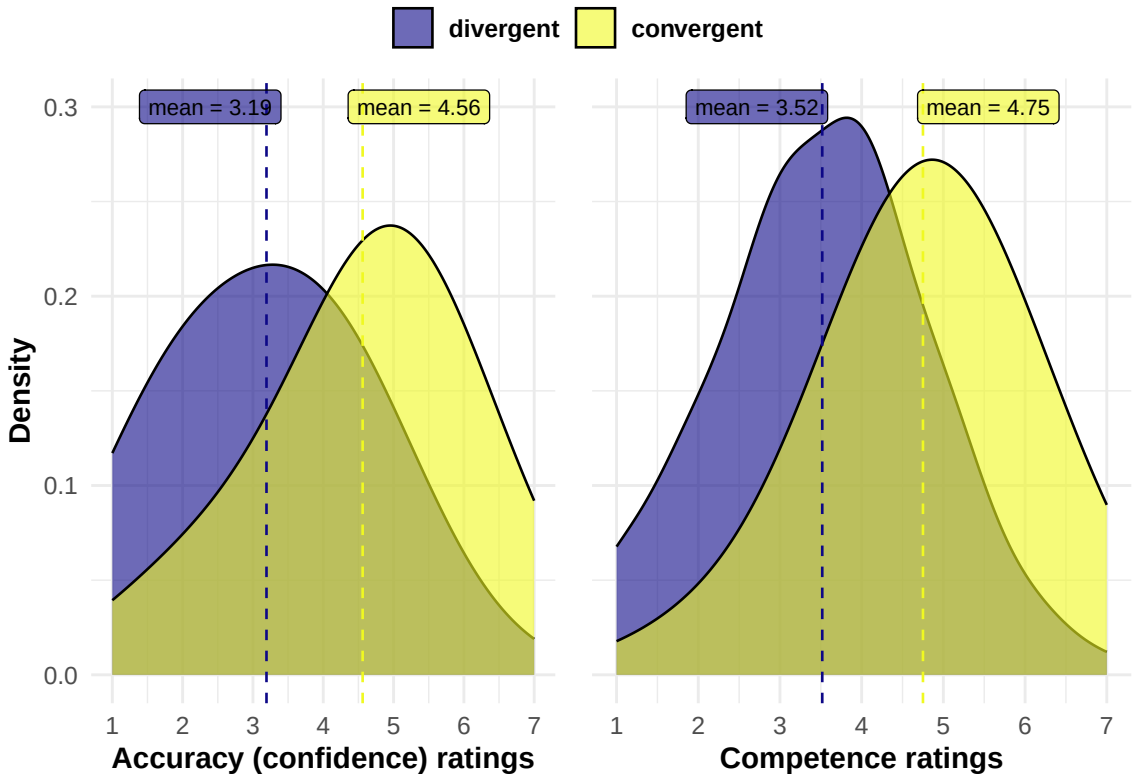


Figure C1. Distributions of accuracy and competence by level of convergence.

RQ2: When making a guess based on the opinions of (independent) informants, will participants be more confident about their guess when there is a larger number of estimates compared to when this number is smaller?

RQ3: Is there an interaction effect between the number of estimates and convergence on perceived competence of informants?

Regarding RQ1, we ran the same mixed models as for the two hypotheses, but without number as a covariate. Instead, we ran separate models for the three and the ten informants conditions. We find that for both sub-samples, there is a positive effect of convergence on both accuracy ($\hat{b}_{\text{three informants}} = 1.27$ [1.112, 1.433], p = < .001; $\hat{b}_{\text{ten informants}} = 1.46$ [1.285, 1.64], p = < .001) and competence ($\hat{b}_{\text{three informants}} = 1.12$ [0.955, 1.295], p = < .001; $\hat{b}_{\text{ten informants}} = 1.34$ [1.142, 1.538], p = < .001).

To test the other two research questions, we ran the same mixed-models as for the hypotheses, but this time including an interaction between number of informants and convergence. We use deviation-coded versions of our convergent variable (divergent = -0.5, convergent = 0.5), allowing us to have a coefficient measuring the main effect of the number of informants in our model. Regarding RQ2, pooling across convergent and divergent conditions, we find a main effect of number, such that participants had more confidence in their estimate when they relied on ten informants compared to three informants ($\hat{b}_{\text{Accuracy}}$ = 0.18 [0.076, 0.274], p = < .001). Regarding RQ3, we find an interaction between num-

ber of informants and convergence on competence: the positive effect of convergence was stronger in scenarios involving ten informants compared to three informants ($\hat{b}_{\text{Competence}}$ = 0.22 [0.028, 0.402], p = 0.024). We do not find a statistically significant interaction on accuracy ($\hat{b}_{\text{Accuracy}}$ = 0.19 [-0.007, 0.387], p = 0.059).

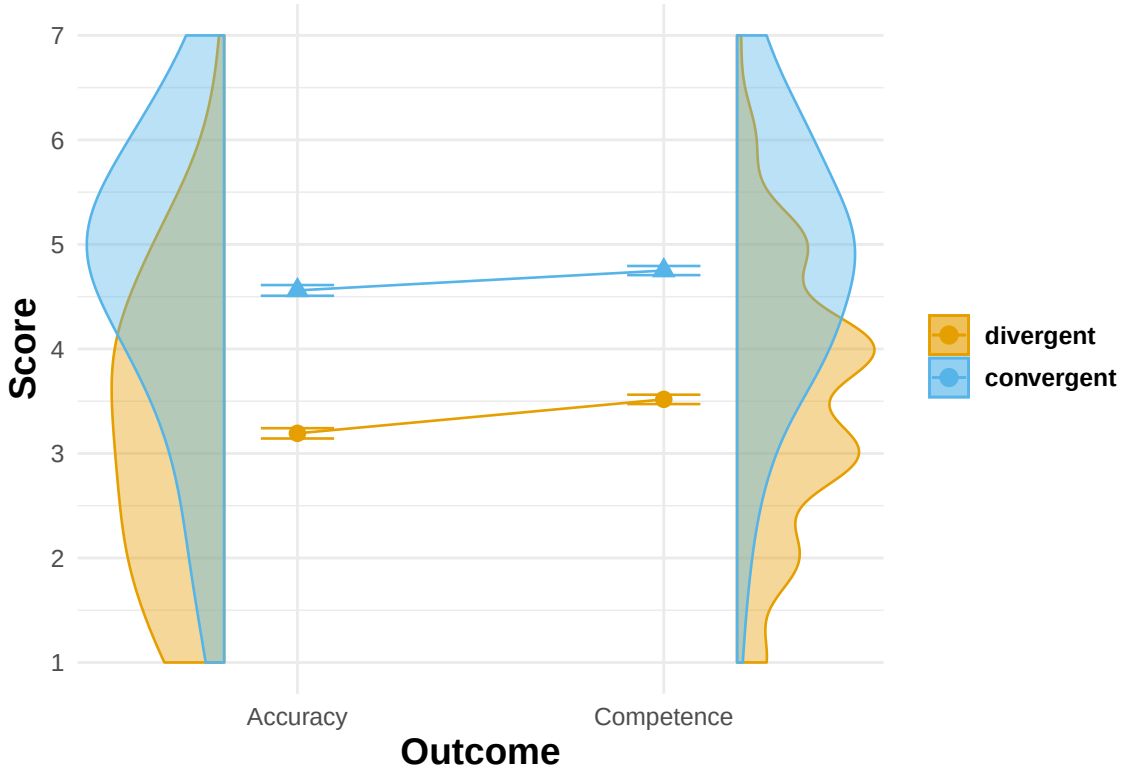


Figure C2. Differences between accuracy and competence ratings, by level of convergence.

Minor deviation from preregistration

In our preregistration, we stated we would run paired t-tests and only conduct mixed models as a robustness check. However, it became clear to us later that mixed models are more appropriate in light of our data structure. In the main article, we therefore only report the results of the mixed models. Note, however, that using paired t-tests, we confirm our results (Accuracy: -1.37, t-statistic = -26.14, p < .001; Competence: -1.23, t-statistic = -21.82, p < .001)

Sensitivity simulation

We had based our power calculations on a paired t-test (using the widely used tool G*Power), which does not correspond to the mixed-models we eventually ran. To address this point, here, we seek to identify the minimum effect size that our analysis with mixed models could have detected, given a 90% power, an N of 200 participants, and other parameters derived from the estimates observed in the sample of Experiment 1 (see C2). By

Table C2
Model Results Experiment 1

	Accuracy	Competence
(Intercept)	3.105	3.516
	(0.077)	(0.068)
convergenceconvergent	1.368	1.232
	(0.072)	(0.085)
numberlarge	0.175	0.002
	(0.050)	(0.048)
SD (Intercept ID)	0.905	0.764
SD (convergenceconvergent ID)	0.732	0.993
Cor (Intercept~convergenceconvergent ID)	-0.128	-0.554
SD (Observations)	1.006	0.955
Num.Obs.	1600	1600
R2 Marg.	0.191	0.195
R2 Cond.	0.594	0.532
AIC	5131.0	4950.5
BIC	5168.7	4988.1
ICC	0.5	0.4
RMSE	0.91	0.86

contrast to a standard power simulation (which varies sample size), a sensitivity simulation varies the effect size, while holding the sample size and desired power constant.

The results of this can be seen in Fig. C3. In our sample, we found a statistically significant effect size of 1.37 for accuracy and an effect size of 1.23 for competence (both on their original scales from 1 to 7). The simulations show that, given our parameter assumptions (based on the observations in our sample) and a sample size of N = 200, we would have detected an effect as small as 0.4 with a power of at least 90% (in fact nearly 100%) for both accuracy and competence.

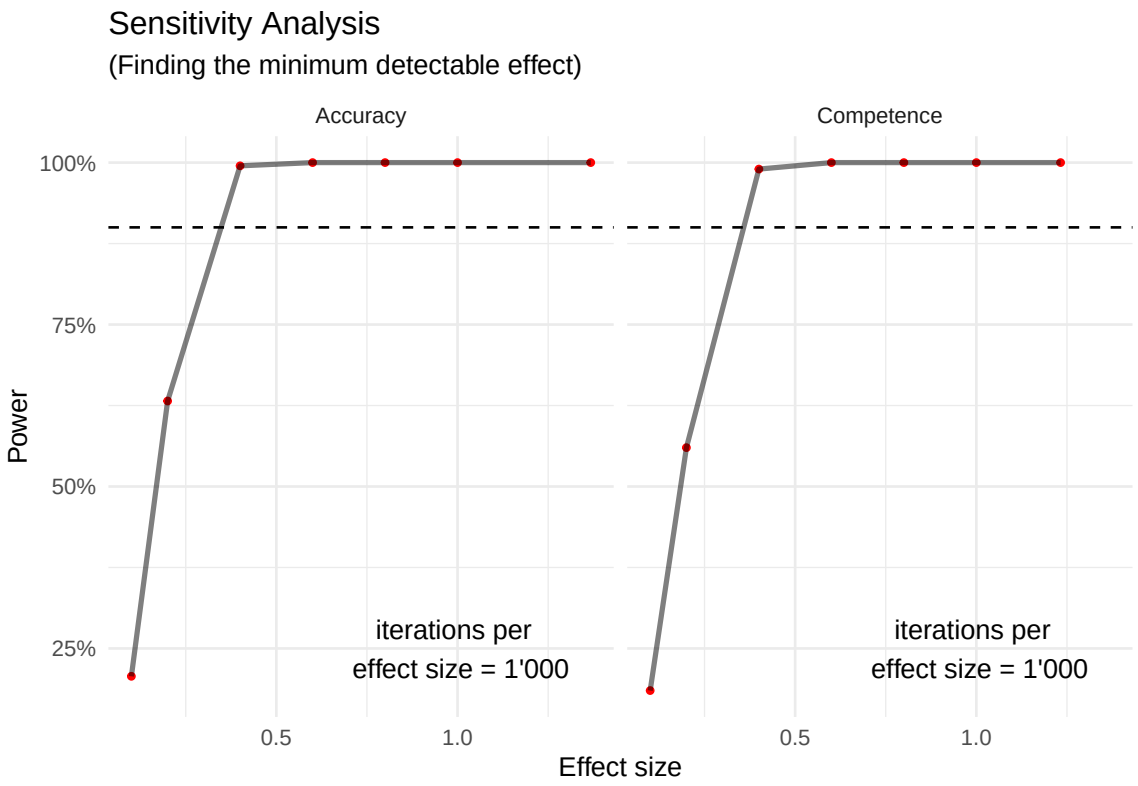


Figure C3. Results of the sensitivity analysis.

Appendix D
Experiment 2

Design

We manipulated informational dependency. According to the condition, players read a different introduction before seeing a set of estimates (see Table D1).

Attention check

Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

Do you agree with your friend? (Yes/No)

Results

Figure D1 visualizes the results and table D2 contains descriptive results.

Table D1

Condition	Description
Independence	Players are asked to make completely independent decisions – they cannot see each other's estimates, or talk with each other before giving their estimates.
Discussion	Players are asked to talk with each other about the game at length before giving their estimates.

Table D2

Independence	Accuracy	Competence
dependent	4.03 (sd = 1.389)	4.48 (sd = 1.022)
independent	3.775 (sd = 1.502)	4.36 (sd = 0.967)

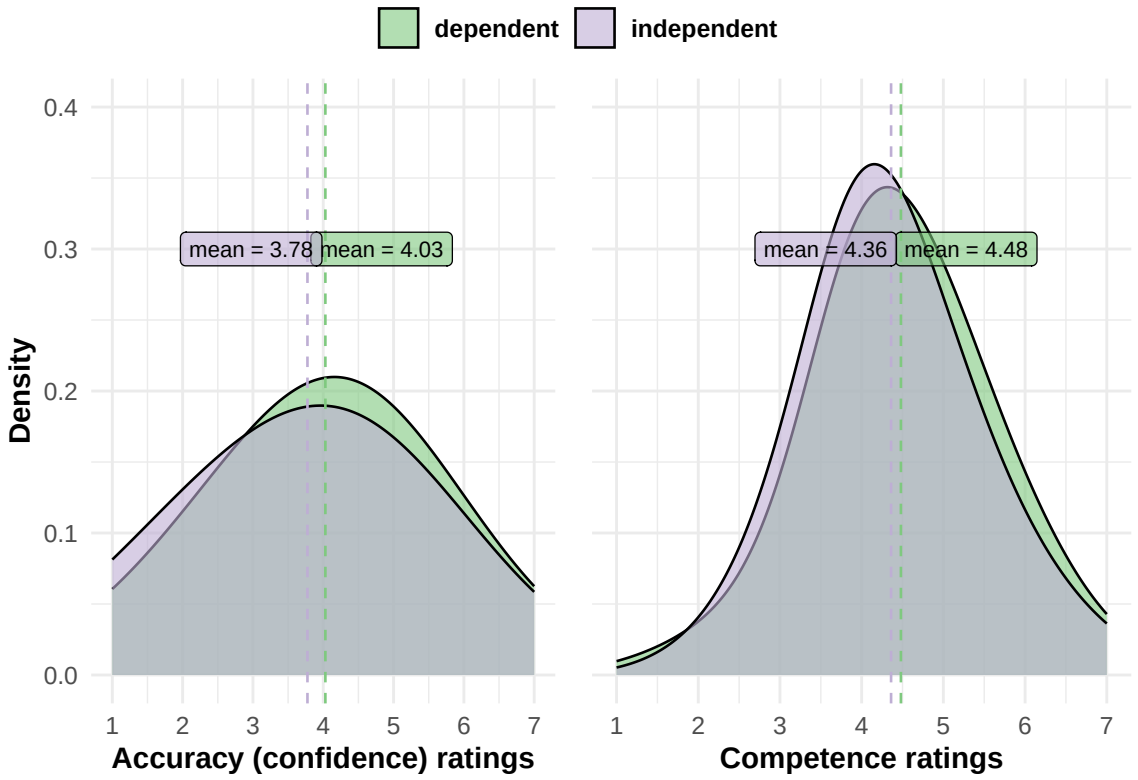


Figure D1. Distributions of accuracy and competence by level of informational dependency.

Appendix E
Experiment 3

Design

We manipulated two factors: informational dependency (two levels, independence and conflict of interest, see Table E1; between participants) and convergence (two levels, convergence and divergence; within participants). Participants saw four scenarios, one for each combination of the two factors convergence and informational dependency.

Attention check

Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

Do you agree with your friend? (Yes/No)

Stimuli

Our design required that each participant sees four different sets of predictions–two convergent and two divergent ones. We generated new stimuli: As in experiment 1, estimates would appear on a scale from 1000 to 2000. By contrast with experiment 1, we generated the sets of estimates with random draws from uniform distributions (instead of normal distributions). We varied the range of these distributions according to convergence (60 for convergence, 600 for divergence). Switching from normal distributions to uniform distributions made 'unlucky' draws in which conditions are visually not distinguishable less likely. And additional difference to experiment one was that all scenarios only involved groups of three informants (never ten).

Similar to the previous experiments, we also vary the value on the prediction scale (from 1000 to 2000) across which the range is centered. Considering our within-participant design, this, again, made it seem more likely that participants understand each set of predictions as being the result of a different stock, with a different true value. In order to assure all random draws from the distributions would appear on the response scale, we constrained the center of the uniform distributions to lie between 1300 and 1700. We

Table E1

Condition	Description
Independence	Experts are independent of each other, and have no conflict of interest in predicting the stock value - they do not personally profit in any way from any future valuation of the stock.
Conflict of interest	All three experts have invested in the specific stock whose value they are predicting, and they benefit if other people believe that the stock will be valued at [mean of respective distribution] in the future.

Table E2

	Accuracy		Competence	
	Divergent	Convergent	Divergent	Convergent
Conflict of interest	3.566 (sd = 1.215)	4.455 (sd = 1.409)	3.732 (sd = 1.101)	4.556 (sd = 1.19)
Independent	3.4 (sd = 1.08)	5.28 (sd = 1.052)	3.61 (sd = 1.111)	5.235 (sd = 0.992)

define four center values – one per set of predictions – that divide this interval in (rounded) quartiles (1300, 1430, 1570, 1700). Given a center and a range, we then draw the predictions from uniform distributions. For example, in a draw for a divergent set of estimates with a center of 1700, each value within a range of 1400 and 2000 is equally likely to get selected. To avoid that single draws overlap too much within the same set, we defined a minimum space of 5 between the three predictions of a set.

To minimize confounding of convergence with a certain placement of the center value, we paired center values with conditions such that each condition appears once in each half of the scale and each condition gets one of the extreme values. For example, in one set the convergence condition would get assigned center values of 1300 and 1570, the divergent condition center values of 1430 and 1700). We generated four such series (all possible combinations) and randomly assigned participants to one of them. Additionally, for each participant, we randomized the order of appearance of the sets of predictions within the respective series.

For each set of predictions, we calculated the empirical mean based on the randomly drawn estimates. In the conflict of interest condition, this mean was inserted as the value that participants were be told experts would gain from. Consequently, the convergent predictions converge around what is said to be the incentivized value for the experts to choose.

Results

Table E2 contains descriptive results.

Appendix F
Experiment 4

Attention check

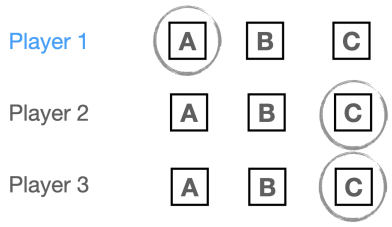
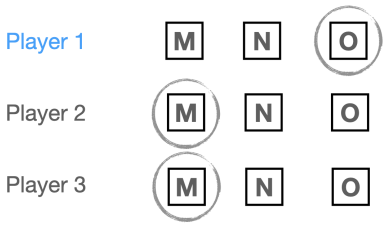
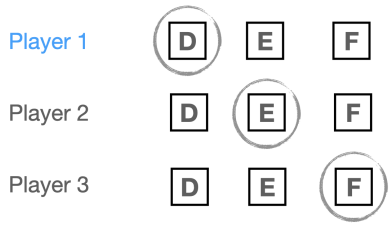
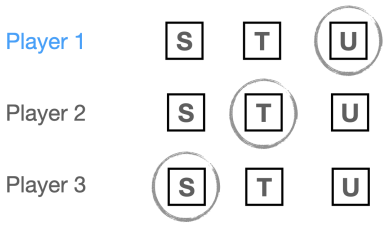
Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

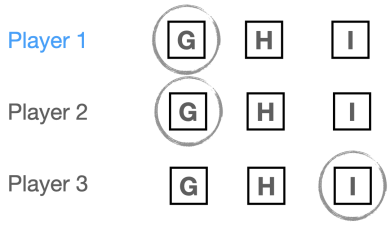
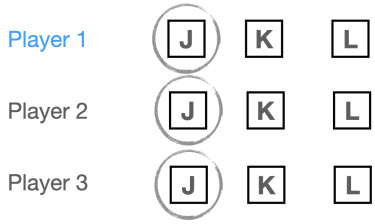
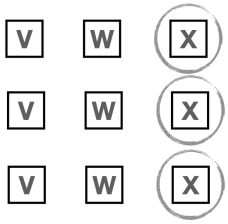
Do you agree with your friend? (Yes/No)

Stimuli

We manipulate convergence within participants. All participants see all four conditions, with two stimuli (i.e. game results) per condition (see Table F1). Each participant therefore sees eight stimuli in total (4 convergence levels x 2 stimuli)

Table F1
All stimuli by levels of convergence

Level	Version a)	Version b)
minority (0)		
dissensus (1)		

Level	Version a)	Version b)
majority (2)		
consensus (3)		

Results

Table F2 contains descriptive results and Figure F1 visualizes the results.

Comparison with simulation

We compared the accuracy and competence ratings of the participants to actual accuracy and competence values from simulated data. The data was generated using the model described in the main paper, and assuming a uniform competence distribution. Compared to this data generating model, participants underestimate the effect of convergence for both

Table F2

Convergence	Accuracy	Competence
minority (0)	33.13 (sd = 18.267)	3.49 (sd = 1.156)
divergence (1)	38.525 (sd = 13.708)	3.93 (sd = 0.726)
majority (2)	64.065 (sd = 13.86)	4.88 (sd = 0.713)
consensus (3)	80.745 (sd = 18.633)	5.45 (sd = 0.981)

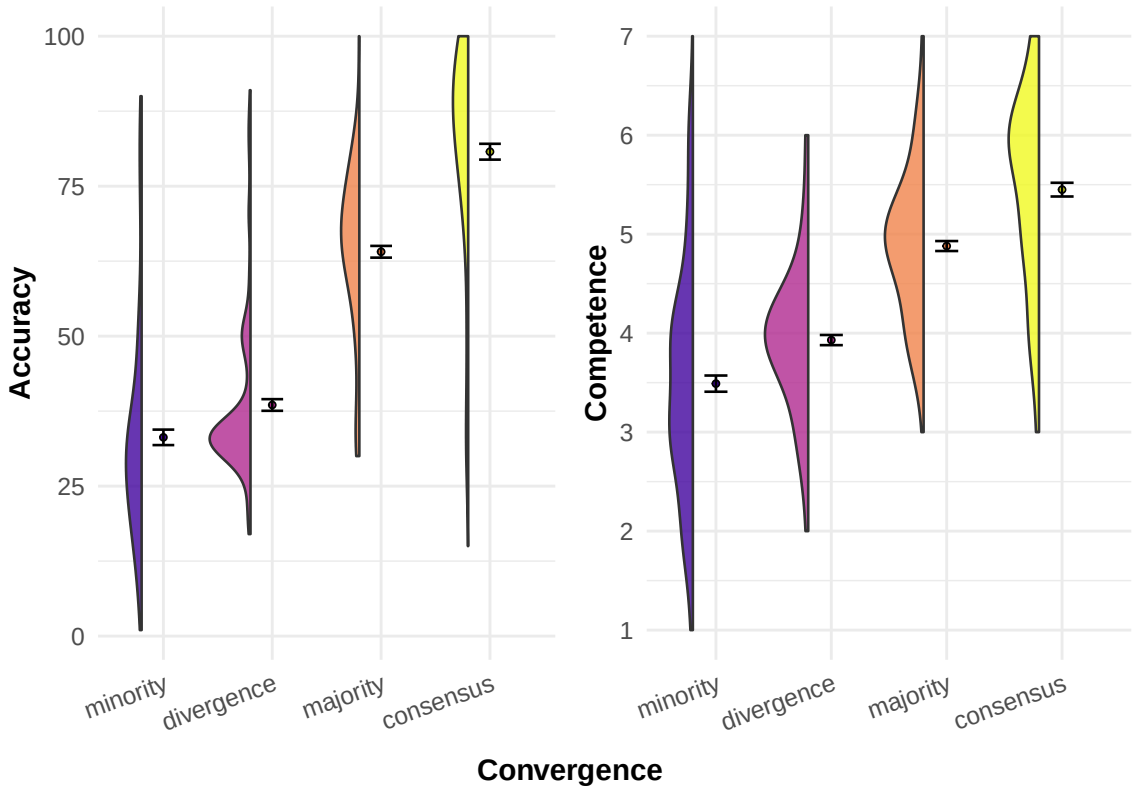


Figure F1. Results of Experiment 4, showing the distributions of accuracy and competence by level of convergence.

accuracy and competence, but more for accuracy. Table F3 compares the results of regressions on participant data and regressions on simulated data. Fig. F2 visualizes the results.

Table F3
Participants vs. Model (Exp. 4)

	Participants		Model	
	Accuracy	Competence	Accuracy	Competence
(Intercept)	28.859***	3.413***	15.211***	3.233***
	(1.702)	(0.098)	(0.221)	(0.008)
convergence	16.838***	0.683***	28.471***	0.425***
	(0.922)	(0.053)	(0.097)	(0.003)
SD (Intercept id)	15.926	0.929		
SD (convergence id)	8.640	0.497		
Cor (Intercept~convergence id)	-0.808	-0.857		
SD (Observations)	10.159	0.553		
Num.Obs.	800	800	333000	333000
R2			0.205	0.047
R2 Adj.			0.205	0.047
R2 Marg.	0.555	0.408		
R2 Cond.	0.839	0.786		
AIC	6402.6	1756.0	3301803.1	1051474.5
BIC	6430.7	1784.1	3301835.2	1051506.7
ICC	0.6	0.6		
Log.Lik.			-1650898.534	-525734.261
F			85844.746	16470.902
RMSE	8.98	0.49	34.42	1.17

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

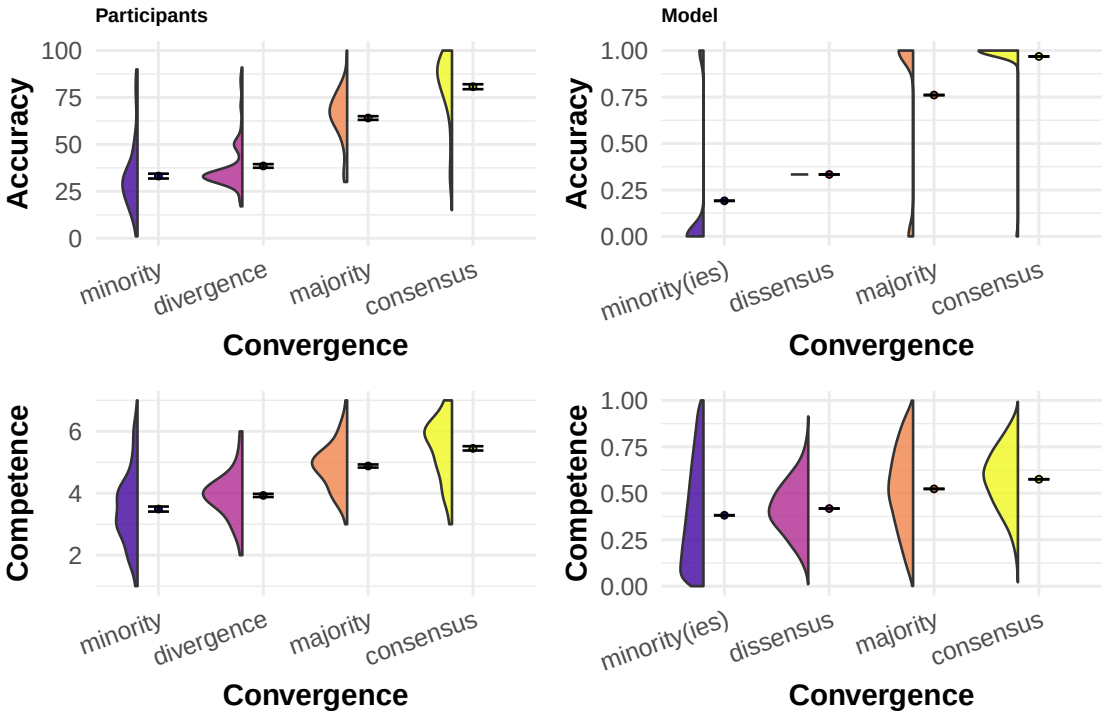


Figure F2. Comparison of accuracy (top) and competence (bottom) ratings in participant data (left) to observed accuracy in simulated data (right), by level of convergence.

Appendix G
Experiment 5

Design

We manipulated convergence within participants in the same way we did in experiment 4. In addition, between participants, we manipulated informational dependence, akin to experiment 3 (Table G1). In the biased condition, experts were described to gain personally from recommending a certain investment option - but without specifying what that option is. In the independent condition, there was no such conflict of interest and experts were described as independent.

Attention check

Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

Do you agree with your friend? (Yes/No)

Results

Table G2 contains descriptive results.

Table G1

Condition	Description
Independence condition	The three advisors are independent of each other, and have no conflict of interest in making investment recommendations.
Conflict of interest condition	The three advisors have already invested in one of the three options, the same option for all three. As a result, they have an incentive to push that option in their recommendations.

Table G2

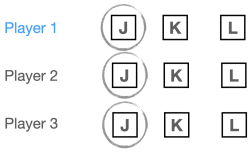
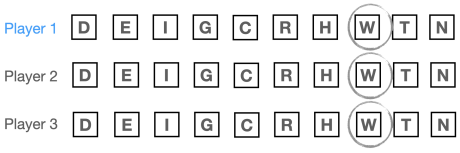
	Accuracy		Competence	
	Conflict of interest	Independent	Conflict of interest	Independent
minority (0)	39.733 (sd = 20.238)	36.46 (sd = 16.793)	3.802 (sd = 1.146)	3.606 (sd = 1.006)
divergence (1)	41.52 (sd = 13.344)	42.51 (sd = 13.6)	4.005 (sd = 0.878)	4.045 (sd = 0.839)
majority (2)	55.416 (sd = 16.637)	59.515 (sd = 14.575)	4.495 (sd = 0.921)	4.737 (sd = 0.879)
consensus (3)	66.198 (sd = 24.825)	71.914 (sd = 23.457)	4.96 (sd = 1.273)	5.247 (sd = 1.264)

Appendix H
Experiment 6

Design

The main focus of experiment 6 was on comparing 3-options conditions to 10-options conditions (see Table H1). Additionally, we varied between two versions of the 10-options conditions: one in which the range of the answers corresponded to the range of the three options condition (see Table H2), and another with increased range (see Table H3).

Table H1
Example of a consensus stimulus for the two 'Number of option' conditions

Number of options: 3	Number of options: 10
	

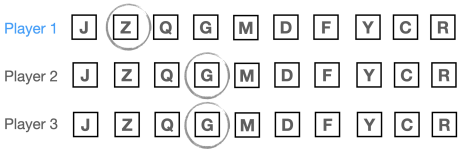
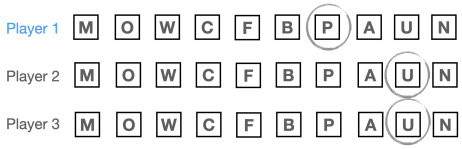
Attention check

Imagine you are playing video games with a friend and at some point your friend says: "I don't want to play this game anymore! To make sure that you read the instructions, please write the three following words"I pay attention" in the box below. I really dislike this game, it's the most overrated game ever."

Do you agree with your friend? (Yes/No)

Stimuli

Table H2
Stimuli for 10 options condition by levels of convergence

Level	Version a)	Version b)
opposing		
majority (0)		

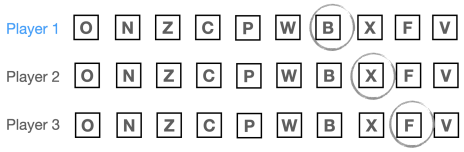
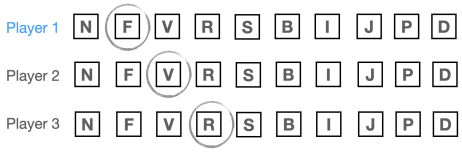
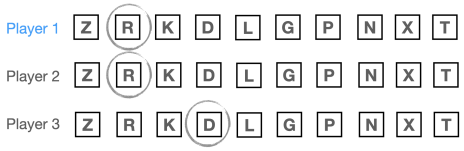
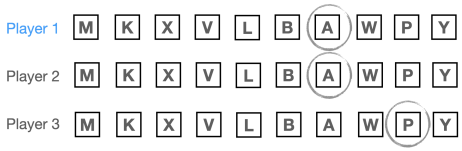
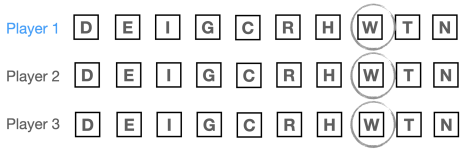
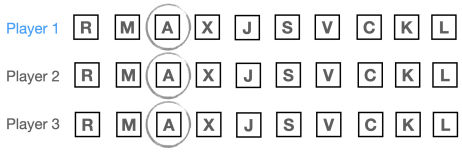
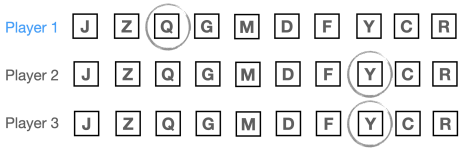
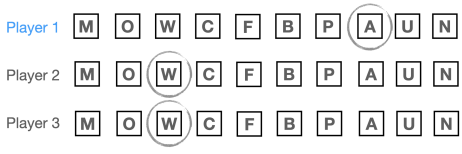
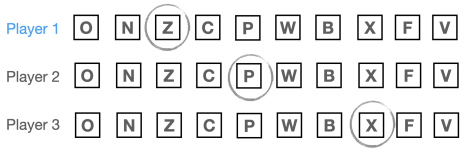
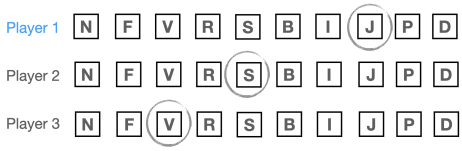
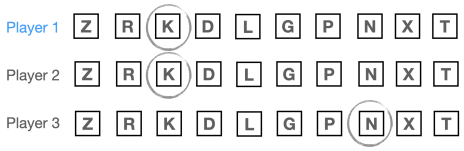
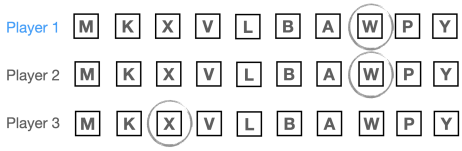
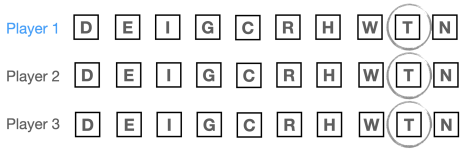
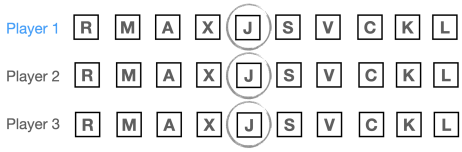
Level	Version a)	Version b)
dissensus (1)		
majority (2)		
consensus (10)		

Table H3

Alternative stimuli for 10 options condition by levels of convergence

Level	Version a)	Version b)
opposing majority (0)		

Level	Version a)	Version b)
dissensus (1)		
majority (2)		
consensus (10)		

Results

Figure 6 visualizes the results and table G2 contains descriptive results.

Comparison with simulation

Since we didn't find a difference between three and ten choice option scenarios in the participant data, we wanted to have a normative point of reference of what should

Table H4

	Accuracy		Competence	
	10 options	3 options	10 options	3 options
minority (0)	32.333 (sd = 22.987)	35.703 (sd = 20.829)	3.61 (sd = 1.245)	3.574 (sd = 1.227)
divergence (1)	37.287 (sd = 19.569)	40.861 (sd = 18.759)	4.067 (sd = 0.831)	4.01 (sd = 0.881)
majority (2)	63.69 (sd = 21.826)	64.301 (sd = 17.355)	4.957 (sd = 0.885)	4.834 (sd = 0.881)
consensus (3)	79.967 (sd = 24.15)	80.152 (sd = 20.774)	5.647 (sd = 1.035)	5.466 (sd = 1.051)

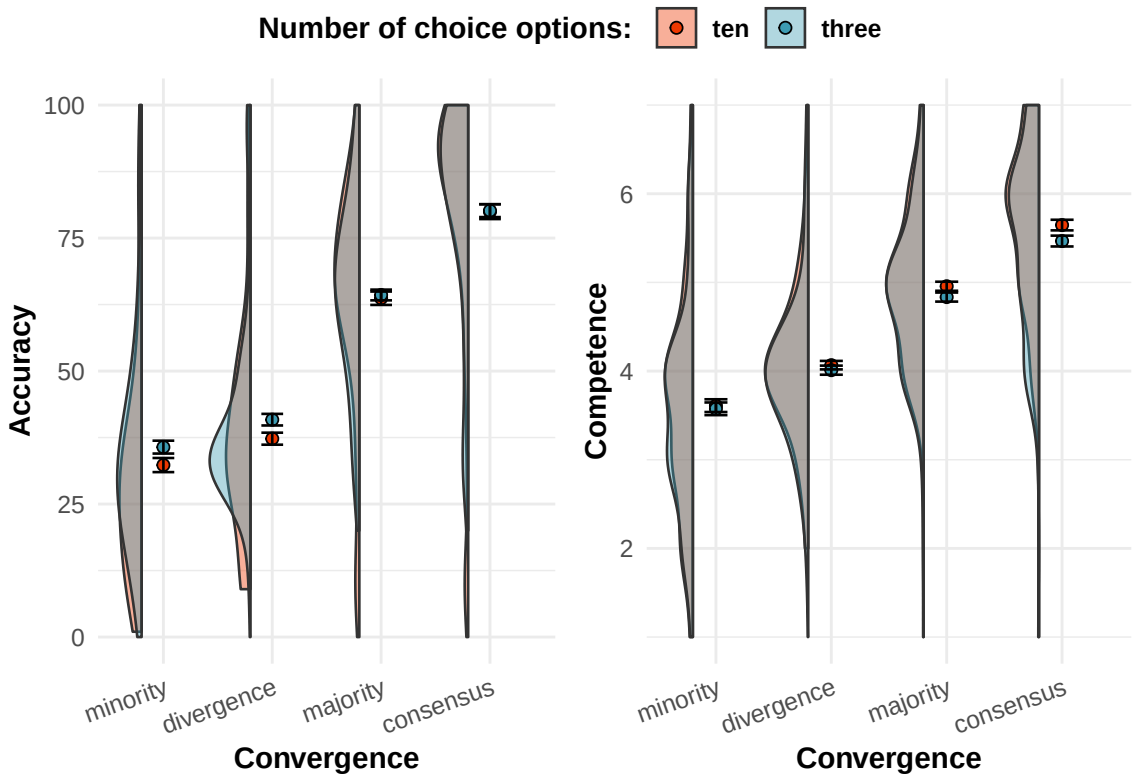


Figure H1. Interaction of convergence and informational dependency.

have been the expected difference, given our model. We therefor compared participant ratings to observed values from simulated data. The data was generated using the model described in the main paper, and assuming a uniform competence distribution. Compared to this data generating model, participants considerably underestimated how the number of choice options alters the effect of convergence for both accuracy and competence. Table F3 compares the results of regressions on participant data and regressions on simulated data. Fig. H2 visualizes the results.

Table H5
Participants vs. Model (Exp. 6)

	Participants		Model	
	Accuracy	Competence	Accuracy	Competence
(Intercept)	29.830***	3.508***	11.407***	3.189***
	(1.197)	(0.063)	(0.220)	(0.009)
convergence	16.305***	0.675***	31.596***	0.526***
	(0.600)	(0.032)	(0.130)	(0.005)
number_options_effect_code	-3.812	0.024	-13.794***	-0.207***
	(2.394)	(0.126)	(0.439)	(0.018)
convergence × number_options_effect_code	1.252	0.050	8.498***	0.251***
	(1.200)	(0.065)	(0.260)	(0.011)
SD (Intercept id)	19.383	1.031		
SD (convergence id)	9.625	0.527		
Cor (Intercept~convergence id)	-0.699	-0.829		
SD (Observations)	12.103	0.598		
Num.Obs.	2384	2384	51003	51003
R2			0.539	0.165
R2 Adj.			0.539	0.165
R2 Marg.	0.423	0.355		
R2 Cond.	0.814	0.779		
AIC	19980.3	5624.5	497742.9	170192.2
BIC	20026.5	5670.7	497787.1	170236.4
ICC	0.7	0.7		
Log.Lik.			-248866.472	-85091.109
F			19902.699	3370.232
RMSE	10.67	0.53	31.83	1.28

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

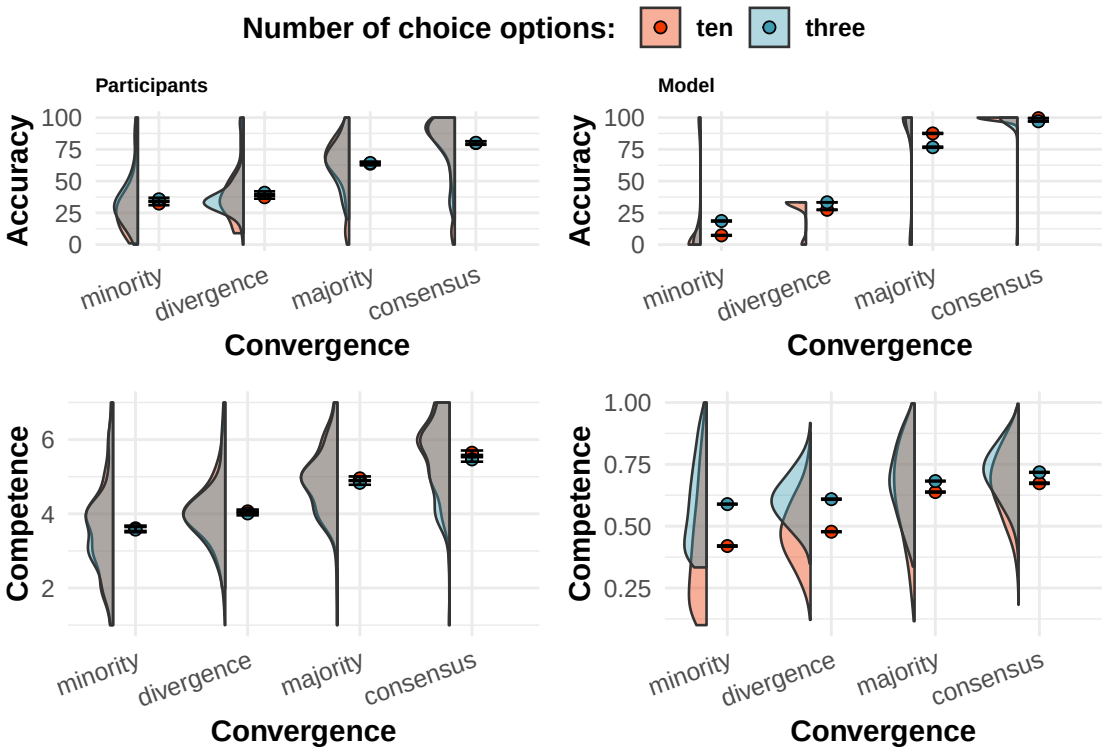


Figure H2. Comparison of accuracy (top) and competence (bottom) ratings in participant data (left) to observations in simulated data (right), by level of convergence.