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

Jan Pfänder[1], Benoît De Courson[2], & Hugo Mercier[1]

[1] Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, France
[2] Max Planck Institute for the Study of Crime, Security and Law, Freiburg im Breisgau, Germany

Abstract

Are people who agree on something more likely to be right? Evidence suggests that people tend to make this inference. However, standard wisdom of crowds approaches do not provide normative grounds for this behavior. Using simulations and analytical arguments, we argue that in stylized scenarios where individuals make independent and unbiased estimates, more convergent groups individuals indeed tend to be more competent and accurate. Mirroring the stylized setting of the simulations, we then show that people make these inferences in a series of experiments. Importantly, in our experiments people do not have any other cues available than the social information (i.e. no priors). These inferences from convergence might help explain why people respect scientists' competence, even if they do not understand much about how scientific results are reached.

Introduction

Imagine that you live in ancient Greece, and a fellow called Eratosthenes claims the circumference of the earth is 252000 stades (approximately 40000 kilometers). You know nothing about this man, and you don't understand trigonometry. Moreover, you know

little about the circumference of the Earth, and even less of how one could measure such a thing. As a result, you discard his opinion and 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?

In this article, we explore how, under some circumstances, we infer that a group of people who provided the same, or very similar answers to a question are likely to be correct, and to be competent in the relevant area, even if we had no prior about either what the correct answer was, or about their competence.

We first review experiments suggesting that people, including young children, appear to trust groups with more convergent opinions to be more accurate, although the evidence is more conclusive in numerical than in categorical choice contexts. We continue by explaining that two well-studied forms of the wisdom of crowds–averaging and the Condorcet Jury Theorem–do not provide normative grounds for this behavior. We then present the results of simulations, suggesting that under certain conditions, inferences from convergence to accuracy and to underlying competence are sound. Finally, we show that participants draw these inferences in a series of six experiments.

Numerical choices

Informants can converge in numerical (e.g. how many calories are there in an apple?) and categorical (e.g. which option is correct - A or B?) choice contexts. In numerical contexts, convergence varies by the proximity of estimates; in categorical contexts, convergence varies by the ratio of people agreeing on a choice. We start by reviewing numerical choices.

For numerical choice contexts, evidence that people take convergence into account can be found in advice-taking experiments. In these experiments, participants are called judges, and they need to make numerical estimates–sometimes on factual knowledge, e.g. 'What year was the Suez Canal opened first?' (Yaniv, 2004a), 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–sometimes presented as fellow participants, sometimes as experts. The overarching question in those studies is how people integrate social information from the advisors when making estimates.

A substantial share of studies using this judge-advisor paradigm is irrelevant for studying inferences from convergence, because they present participants only with single estimates. In some studies, that single estimate is coming from one advisor (e.g. Bednarik & Schultze, 2015; Harvey & Fischer, 1997; Soll & Larrick, 2009; Yaniv, 2004a; Yaniv & Kleinberger, 2000). In other studies, that estimate is presented as an average coming from a group of advisors (e.g. Jayles et al., 2017; Mannes, 2009), but no information about the distribution of initial estimates is given. By contrast, other advice-taking studies provide participants with a set of estimates. One subset of these studies manipulate the degree

of convergence between groups of advisors, e.g. via variance of estimates (Molleman et al., 2020; Yaniv, Choshen-Hillel, & Milyavsky, 2009), or their range (David V. Budescu & Rantilla, 2000; David V. Budescu, Rantilla, Yu, & Karelitz, 2003; David V. Budescu & Yu, 2007). These studies find that participants are more confident about, or rely more on, estimates from groups of advisors who converge more. More indirect evidence comes from another subset of studies varied the degree of convergence within groups of advisors. They present participants with sets of estimates where some estimates are close to each other (or overlapping, in cases where estimates were intervals), while others were clearly outliers (Yaniv, 1997, study 3 & 4; Harries, Yaniv, & Harvey, 2004). These studies find that participants heavily discount outliers when aggregating estimates. These results are thus consistent with the idea that people rely more on convergent opinions, although testing extreme cases of singular dissident opinions. Another study varied convergence between estimates from single advisors (Yaniv, 1997, study 1 & 2). In this study, participants were shown ranges of estimates from two advisors - one with a small range, the other with a big range. Participants relied more on the precise, small range advice. These results can be interpreted as a case of inferring accuracy from (internal) convergence of single informants, perceived by participants perhaps as consistency or for example confidence. However, in this study, the small-range advisor (e.g. 3 to 5) was always within the range of the big-range adivsor (e.g. 1 to 10). If people relied more on the small-range advisor, then it was likely not only because of greater convergence, but also because of the overlap between the two advisors.

Categorical choices

For categorical choice contexts, experimental evidence on people inferring accuracy from convergence is scarce. There is an extensive body of literature suggesting that people are susceptible to adopting majority behavior and opinions. But it is is difficult to tear apart whether they do so in a strive for normative conformity or for accuracy (Mercier & Morin, 2019). However, some aspects of the experimental paradigm should favor motivations for accuracy, for example private answers and individual rewards for accuracy. As for human adults, we are aware of only one article that assessing majority effects using such accuracy motivations: In a series of experiments using such accuracy motivations, T. J. H. Morgan, Rendell, Ehn, Hoppitt, and Laland (2012) systematically varied the proportion of informants agreeing on a given answer, from 50 to 100%. They found that the larger the (relative) size of the majority, the more likely participants were to adopt the majority answer. As for children, relevant evidence comes from studies on social learning in the field of developmental psychology. Several such studies rely on a similar paradigm revolving around a naming tasks (Bernard, Harris, Terrier, & Clément, 2015; Bernard, Proust, & Clément, 2015; Chen, Corriveau, & Harris, 2013; Corriveau, Fusaro, & Harris, 2009; Fusaro & Harris, 2008). In Corriveau et al. (2009), for example, children had to find out which of three unfamiliar objects corresponds to the made up name "modi". Three adult informants pointed to one object, but a fourth adult pointed to a different one. The children were then asked for their opinion. All these studies consistently find that children adopt the majority opinion. It is, however, not clear how the naming paradigm taps into intuitions for inferring accuracy - after all, assigning made-up names to objects has no objectively correct solution and,

accordingly, these studies are often interpreted to reveal conformity in children (see e.g. Leeuwen, Kendal, Tennie, & Haun, 2015). Other experimental paradigms yield similar results. Thomas J. H. Morgan, Laland, and Harris (2015) had children (aged between 3 and 7) count which of two pictures displayed more dots. Among 10 adult informants, they varied the proportion of those agreeing on one picture. They find that children relied more on advice from larger (relative) majorities, although younger children (aged 3 to 4) were only affected in cases where the majority would stand against a single dissident. In Haun, Rekers, and Tomasello (2012), 2-year-olds tried to use an unfamiliar box to deliver a reward. The children were more likely to copy an action demonstrated by three informants rather than a different action demonstrated three times by the same informant. In Herrmann, Legare, Harris, and Whitehouse (2013), children (aged 3 to 6) were required to copy a necklace-making demonstration from video. Children showed more fidelity to the demonstrations after watching two identical demonstrations by two different adults than after watching the same individual twice. In Leeuwen et al. (2018), children could use a choice box with three different pipes to insert a boll, to then receive a toy. Before using the box, they were shown a video of four other children using the box. Three of these children used one particular pipe, each throwing in one ball and receiving one toy per insertion, sequentially. One additional demonstrator used one of the two other pipes three times in a row, also receiving one toy per insertion. Children were then given one ball and had to chose which pipe to use. Children generally followed the majority choice.

Although the evidence is not conclusive, especially for categorical choice scenarios, the current literature suggests that people believe more convergent opinions to be more accurate. But is this inference sound?

Standard wisdom of crowds approaches

Literature on the wisdom of crowds provides normative ground for inferring accuracy from others' aggregated opinions (Surowiecki, 2005). What role does this literature assign to convergence? Here we look at two of the main phenomena that can generate wisdom of crowds effects: averaging, and the CJT.

Averaging. In the beginning of the 20th century, Francis Galton famously demonstrated accuracy gains from averaging: in his "vox populi" experiment, he asked 787 participants to estimate the weight of an ox, and found the average estimate of an ox weight to be just one pound from reality (Galton, 1907). Since then, it has been shown both theoretically and empirically that the mean or median answer of a group is typically closer to the truth than the mean individual is (Mannes, Soll, & Larrick, 2014; see e.g. Yaniv, 2004b). For instance, when considering a range of numerical estimates that deviate more or less from a correct answer, the error of the mean answer will always be either lower than the mean error (if the correct answer is within the range of all the answers provided), or the same as the mean error (otherwise) (Larrick & Soll, 2006). In fact, the error of the mean is often uncannily small compared to the mean error, a phenomenon which has allowed averaging to considerably improve performance on a variety of problems ranging from political predictions to medical diagnoses (Surowiecki, 2005).

Averaging works independently of the degree of convergence of a set of estimates. It is not clear how it leads to more accurate results, without making assumptions about the data generating process of these estimates.

[One might ask why talk about averaging at all - good question. I guess just because it happens in a numeric context and is used to justify aggregating opinions. But the link is not straightforward.]

Condorcet Jury Theorem. In the late 18th Century, Condorcet established a formal argument in favor of large scale majority voting, which would later become known as the Condorcet Jury Theorem (CJT) (De Condorcet, 2014). He demonstrated that, for a binary decision (e.g. correct vs. false), the probability that the majority vote is correct is larger than the probability of each individual being correct. Moreover, as the number of votes increases, the probability that the majority is correct converges towards one. Crucially, the original CJT assumes that (i) each individual has the same probability p to select the correct answer, and (ii) that this probability is at least slightly better than chance (p > 0.5) (there are also assumptions about the independence and lack of strategic motives, for reviews, see, e.g. Ladha (1992), Austen-Smith and Banks (1996), Dietrich and Spiekermann (2013)).

At least in its basic version, the CJT glosses over variation in convergence, i.e. the relative size of the majority. It allows to make statements such as 'given an individual-level competence p, the probability that a (any) majority vote is correct is Y'. This latter probability comprises votes with 99% majorities as well 51% majorities. To justify inferences of accuracy and competence from convergence, CJT would need to allow making statements such as 'given a majority of X the probability that the majority vote is correct is Y, and the individual-level competence is Z.' In an extension of the CJT, Romeijn and Atkinson (2011) provide normative grounds for making such statements. They show that with increasing relative majority, group members are more likely to be competent and the majority decision to be accurate (??). Their framework is limited, however, to binary decisions.

In sum, the wisdom of crowd literature provides at best insufficient normative grounds for inferring accuracy and competence from convergence. Its main purpose is to justify that aggregating across many opinions is a better strategy than picking any individual's opinion. The proposed aggregation strategies - whether averaging or following the majority - are about extracting wisdom from any crowd. But not all crowds are equally wise. Competence should play a crucial role in the confidence we have in any wisdom of crowd inferences. In the absence of more informative cues, our literature review shows that people place more trust in more convergent crowds. The wisdom of crowd literature does not provide generalizable justification for this behavior.

Inferring competence from the convergence of opinions

As Eratosthenes' example above illustrates, intuitively it seems there are cases in which we can infer that the members of a group are more likely to be correct, or, equivalently, more competent, when their answers converge. Here, we explore two types of settings in

which opinions can converge: a numerical and categorical one. Using simulations, we show that in both settings, under certain circumstances, inferences from convergence are sound.

When several people estimate a quantity, 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. When are we justified to infer that people are right (and thus competent) when they converge?

Overview experiments

We test whether people infer that the members of a group are correct, and thus likely competent, when their answers converge in two series of experiments. In the first series (experiment 1 to 3), the scenario is about numerical estimates, as in averaging. In the second series (experiment 4 to 6), the scenario is about categorical choices, as in the CJT. In all experimental, participants did not have any priors about the correct estimates.

In experiment 1 (numerical) and 4 (categorical), participants evaluated several groups of informants, with each informant providing an answer independently. We manipulated convergence, i.e how close estimates were (Exp. 1, numerical setting), or how consensual an answer was (Exp.4, categorical setting). The more convergence, the more participants inferred accuracy and competence.

In experiments 3 (numerical) and 5 (categorical), we test whether cues of informational dependency inferences hamper inferences from convergence. Besides convergence, we manipulated whether informants were independent (mimicking experiments 1 and 4) or sharing a conflict of interest. We find that (i) people infer from convergence even in the conflict of interest situation, but (ii) that there is enhanced inference when informants are independent. We replicate results experiments 1 and 4 in a slightly different experimental setting.

In experiment 2, we operationalized informational dependency differently: informants were either independent informants or "had discussed at great length with each other" before making their estimate. We find no difference.

In experiment 5, we tested whether more choice options (ten vs. three) enhance inferences from convergence. We reasoned that with more options, consensus is less likely to occur by chance, and thus people might find it more impressive. We did not find that the number of choice options altered effects of convergence.

All experiments were pre-registered. Pre-registration documents, data and code can be found on Open Science Framework (OSF, https://osf.io/6abqy/). All analyses were conducted in R (v.4.1.1) using R Studio.

Numerical estimates: Experiments 1 to 3

Experiment 1

The first experiment was designed to test the effect of convergence on the (average) perceived accuracy of estimates (H1) and the perceived competence of a group of informants

(H2). We decided to measure accuracy as confidence in one's estimate based on a group of informants estimates. Our two hypotheses therefore read:

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.

Participants. We recruited 200 participants from the UK via Prolific. 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". They were then presented to the results of eight such games and had to answer questions (Fig.1).

Design. We manipulated two experimental factors: First, the convergence of the estimates (how close they were); second, the number of estimates (how many players there were). We did not make explicit hypotheses on the latter, but included some research questions (cf. pre-registration document). Each factor had two levels: convergent vs. divergent and low (three) vs. high (ten). We used a 2(convergence: divergent/convergent) x 2(number: low/high) within-participant design, with each participant seeing all of the conditions. Per condition, participants saw two sets of estimates (game results). Thus, each participant saw eight sets of estimates in total.

Stimuli. 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 from these distributions per set (three for low number, 10 for high number). For each of the four possible resulting condition, we generated two random draws from the respective normal distribution.

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
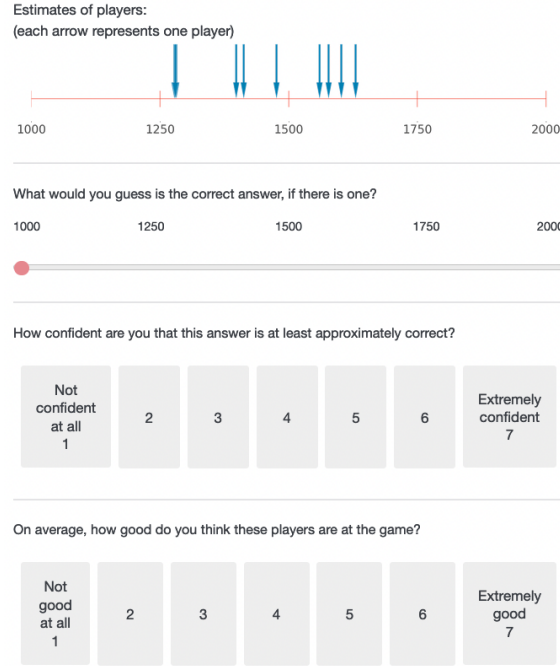
Figure 1. Results of one of eight games that participants have to rate. The stimulus corresponds to a divergent, 10 players condition

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.

Dependent variables. For each set of estimates participants responded to several questions. We first asked participants to 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?"). Participants indicated their numeric guess using a slider on a line identical with the one they saw the estimates on. We did not analyze those guesses. They merely served as a basis for the next question, intended to measure perceived accuracy: "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")[1]. Finally, participants were asked about the competence

---

[1] We changed this measure in study 3. Participants did not make a guess and were asked "How accurate do you think...". The results are highly similar.

Table 1

Convergence	Accuracy	Competence
divergent	3.192 (sd = 1.392)	3.518 (sd = 1.266)
convergent	4.56 (sd = 1.448)	4.75 (sd = 1.237)

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").

 Results. To account from dependencies of observations due to our within-participant design, we ran mixed models, with random intercept and random slope for participants, using the `lme4` package and its `lmer()` function in R. In the models, we control for the number of estimates (our second experimental factor). Figure 2 visualizes the results and table 1 contains descriptive results. We find a positive effect of convergence on accuracy: Participants were more confident about their estimate in convergent scenarios (Δ Accuracy = 1.37 [1.225, 1.51]) than in divergent ones (Accuracy = 3.10 [2.953, 3.257]). We also find a positive effect of convergence on competence: participants rated players as more competent in convergent scenarios (Δ Competence = 1.37 [1.225, 1.51]) than in divergent ones (Competence = 3.10 [2.953, 3.257]).
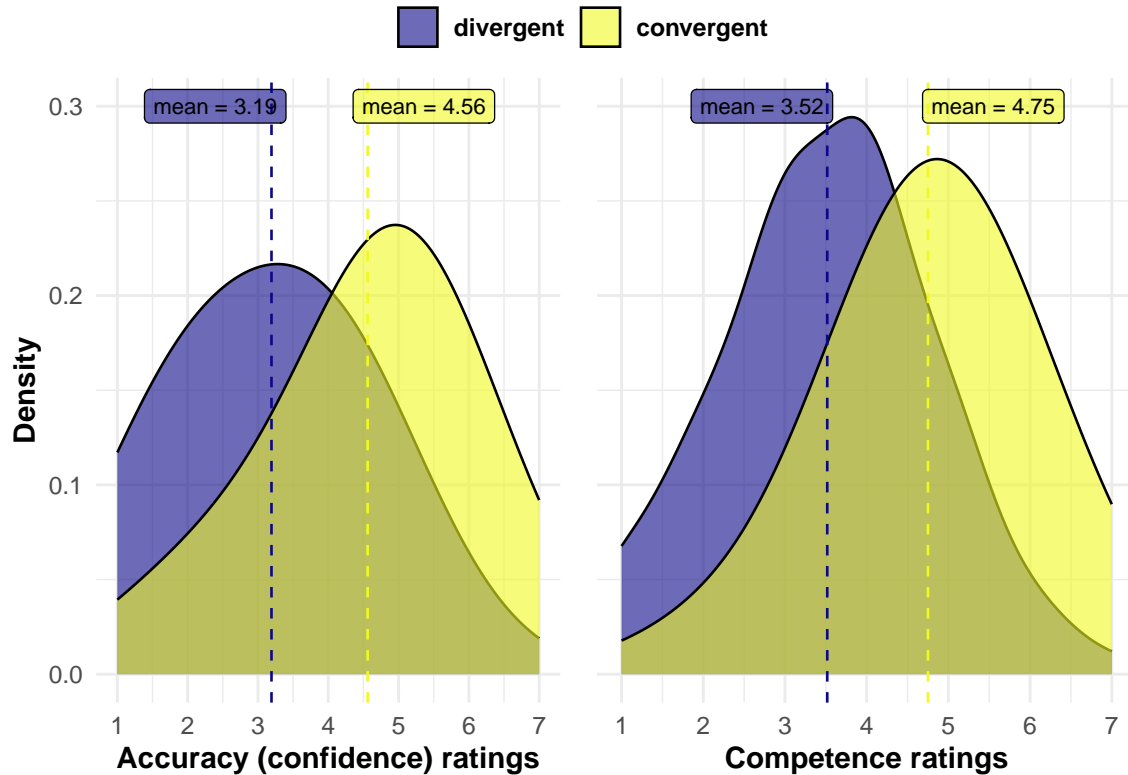

Figure 2. Distributions of accuracy and competence by level of convergence.

Experiment 2

In experiment 1, informants were always independent. In experiment 2, we set out to test whether informational dependency alters inferences from convergence. We used the same setting as in experiment one, in which participants saw the fictional results of an estimation game. However, this time participants only saw convergent sets of estimates. We manipulated informational dependency: players were independent either independent or could "discuss at lenght" before making an estimate. Our hypotheses were:

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).

Participants. We recruited 200 participants from the UK via Prolific. 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 ≥ 0.2 (alpha = .05, two-tailed).

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

```
##      Condition
## 1 Independence
## 2   Discussion
##
## 1 Players are asked to make completely independent decisions - they cannot see each othe
## 2                                                     Players are asked to talk wi
```

Stimuli. We a subset of the stimuli we had generated for experiment one, namely convergent ones with three estimates. By contrast to experiment one, participants saw only two stimuli in total (one set of estimates per condition). Otherwise, we proceeded just as in experiment one: 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.

Dependent variables. We relied on the same set of questions as in experiment one.

Results. To account for dependencies of observations due to our within-participant design, we ran mixed models, with a random intercept for participants, using the `lme4` package and its `lmer()` function in R. Figure 3 visualizes the results and table 2 contains descriptive results. The data does not support our hypotheses. Participants were slightly less confident about their estimates when the converging informants were independent Δ

Table 2

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)

Accuracy = -0.26 [-0.462, -0.048]), compared to when they discussed (Accuracy = 4.03 [3.829, 4.231]). The effect is small, but in the opposite direction of what we had predicted. We do not find an effect regarding competence.
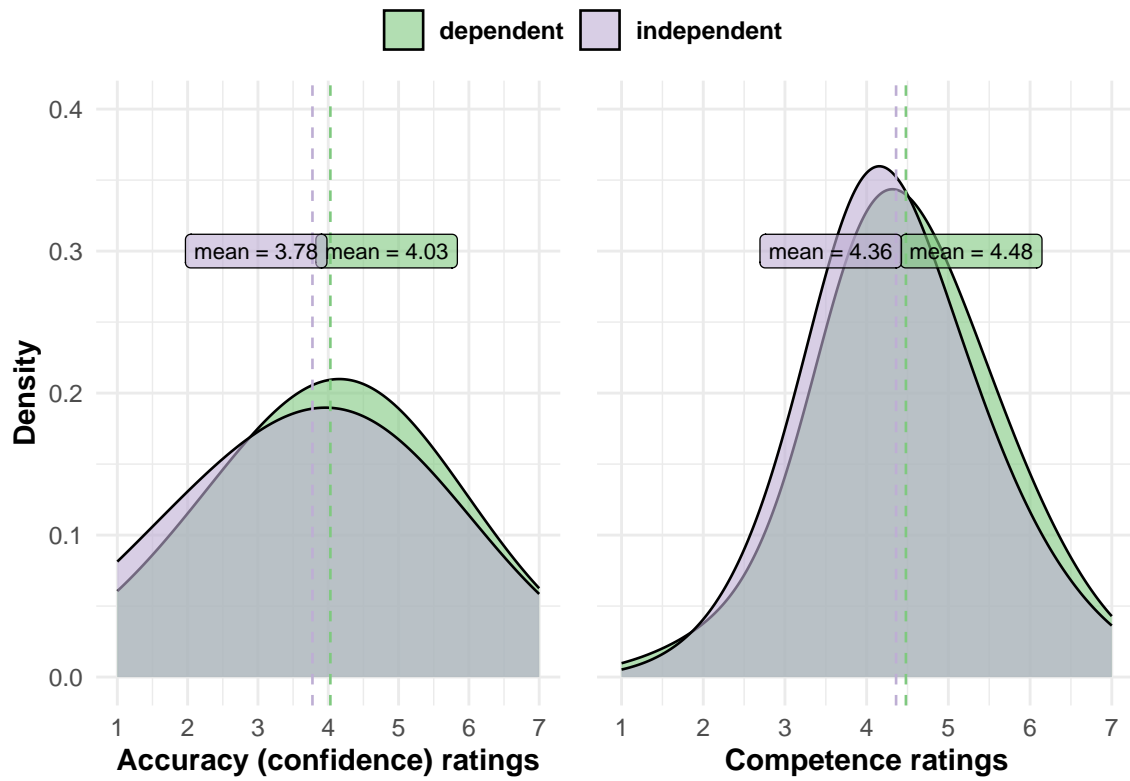

Figure 3. Distributions of accuracy and competence by level of informational dependency.

Experiment 3

The aim of experiment 3 was twofold: First, we wanted to replicate the results of experiment 1 in a slightly more ecological scenario. In experiment 3, participants were told they saw (fictional) predictions of experts for stock values. Since we expected the findings of experiment one to replicate, our first set of hypotheses was:

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.

Second, as in experiment two, we wanted to test the independence assumption – but this time, against a different case of informational dependency, namely when informants shared a conflict of interest. In this condition, experts were described to gain from estimating the value of a stock at a certain price – and that price always happened to be the mean of the the expert's estimates. Our second set of hypotheses was:

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.

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 OSF, and the procedure is described in the pre-registration document. The simulation suggested that 100 participants provide a significant interaction term between 95% and 97% (alpha for significance = 0.05) of the time. Due to uncertainty about the assumptions we made and because we had resources for a larger sample, we recruited 199 participants for this study – again, from the UK and via Prolific. Again, not a single participant failed our attention check.

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 which condition the participant was assigned to (see Table **??**; for more detailed materials see pre-registration document on OSF).

```
##                 Condition
## 1          Independence
## 2 Conflict of interest
##
## 1                                  Experts are independent of each other, and have no conflict o
## 2 All three experts have invested in the specific stock whose value they are predicting,
```

Design.   We manipulated two factors: informational dependency (two levels, independence and conflict of interest, see Table **??**; between participants) and convergence (two levels, convergence and divergence; within participants).

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.

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 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 randomize 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.

Dependent variables. By contrast to the previous experiments, we operationalized accuracy as an actual question on accuracy this time. For each set of estimates participants responded to the following questions: First, "On average, how accurate do you think these three predictions are?" on a 7-point Likert scale ("not accurate at all" to "extremely accurate"); second, "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").

Results. 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, using the `lme4` package and its `lmer()` function in R. Figure 4 visualizes

Table 3

	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)

the results and Table 3 contains descriptive results.

We find evidence all four hypotheses. As for the first set of hypotheses, to match the setting of experiment one, we reduced the sample of experiment three 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 (see left side of Fig. 4). We find a positive effect of convergence on accuracy (Δ Accuracy = 1.88 [1.658, 2.102]; baseline Accuracy divergent: 3.40 [3.213, 3.587]). We find very similar results for competence (Δ Competence = 1.62 [1.411, 1.839]; baseline Competence divergent: 3.61 [3.414, 3.806]).

The second set of hypotheses targeted the interaction of informational dependency and convergence (for a visual representation of these interactions, see Fig. 4). In the independence condition, the effect of convergence on accuracy was more positive (Δ Convergence = 0.99 [0.634, 1.348]) than in the conflict of interest condition (Convergence = 0.89 [0.636, 1.142]). Likewise the effect of convergence on competence is more positive (Δ Convergence = 0.80 [0.474, 1.13]) than in the conflict of interest condition (Convergence = 0.82 [0.591, 1.056]).

Categorical estimates: Experiments 4 to 6

In the first series of experiments (experiments 1-3) we tested inferences from convergence in a numerical choice setting: Participants saw (fictive) players' numeric estimates on a scale from 1000 to 2000. The degree of convergence varied by the distance between estimates.

In the second series of experiments (experiments 4-6), we test inferences from convergence in a categorical choice setting. In the categorical scenario, the fictive players make choices on a set of response options (i.e. categories). Convergence varies by the ratio of people agreeing on an option. Experiment four and five can be considered robustness checks as to whether the results of the first series hold in a categorical choice setting. Experiment six tests a new context factor: the number of choice options.

Experiment 4

In line with our findings in numerical choice settings, we predicted that:

H1: Participants perceive an estimate of an independent informant as more accurate the more it converges with the estimates of other informants.
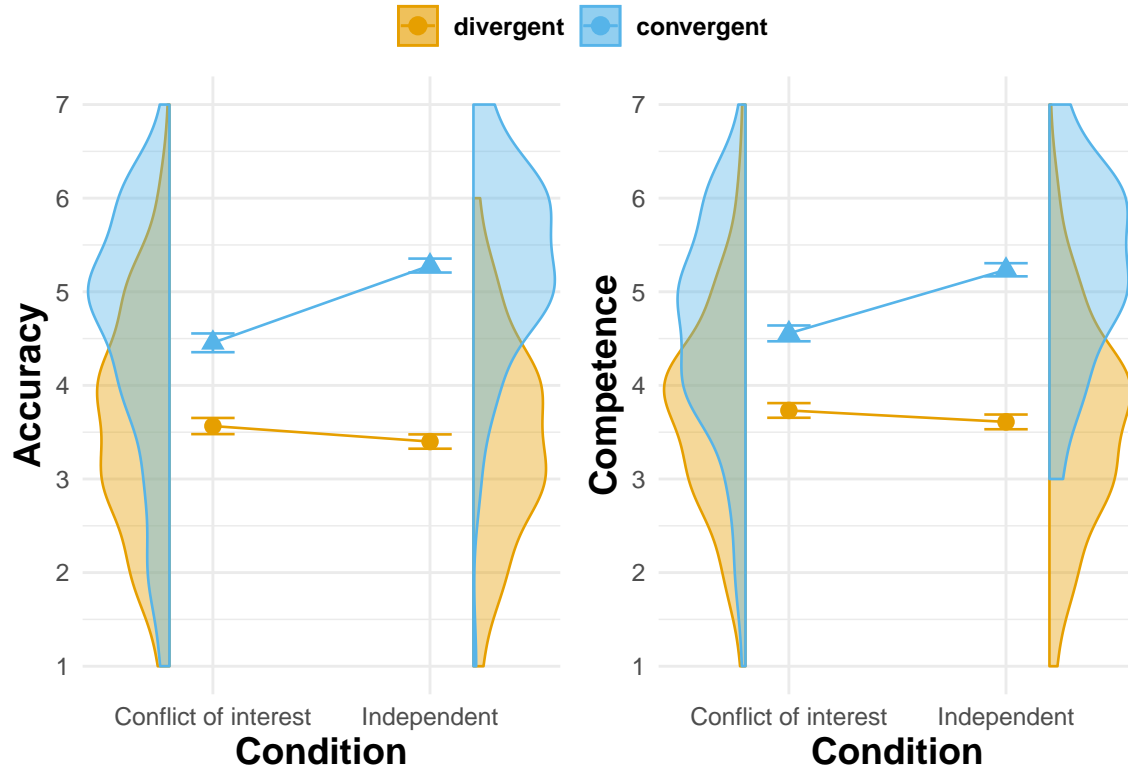
Figure 4. Distributions of accuracy and competence by level of informational dependency.

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

Participants. We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found in the `power_Exp4.Rmd` document. 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, we recruited 200 participants.

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

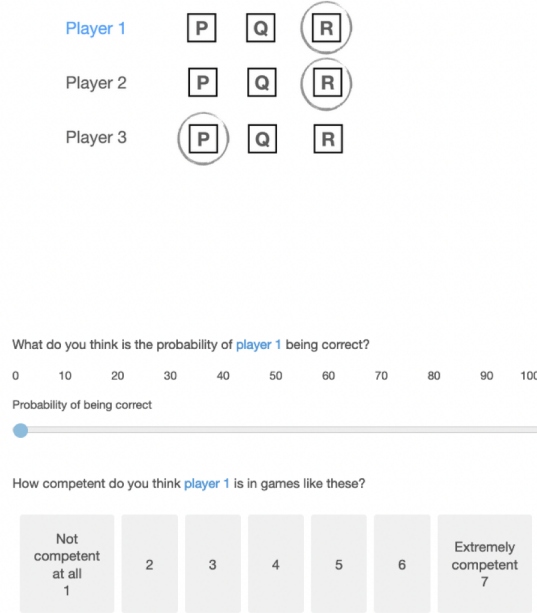presented to the results of eight such games and had to answer questions (see Fig. 5).

Figure 5. An example of a stimulus for the majority condition

Design. We manipulated convergence 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) majority against the focal player's estimate, 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 values in squared parenthesis.

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

Dependent variables. To assess 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. As for 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").

Results. To account from dependencies of observations due to our within-participant design, we ran mixed models, with random intercept and random slope for participants, using the `lme4` package and its `lmer()` function in R. Figure 6 visualizes the results and table 4 contains descriptive results.

Table 4

Convergence	Accuracy	Competence
opposing majority (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)

As in the numerical setting, we found a positive effect of convergence on both accuracy (Accuracy = 16.84 [15.009, 18.668]) and competence (Competence = [0.578, 0.788]).
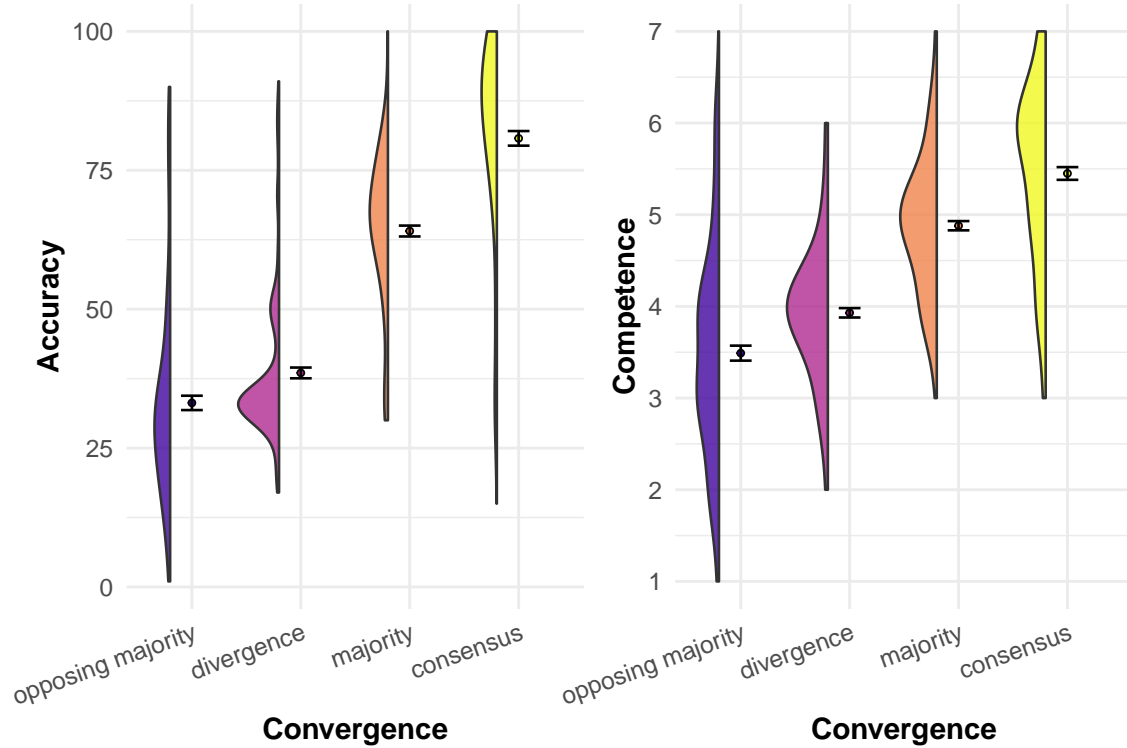

Figure 6. Distributions of accuracy and competence by level of convergence.

Experiment 5

In line with our findings in numerical choice settings, we predicted that:

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).

Participants.   We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found in the `power_Exp5.Rmd` document. 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 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 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.

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." They were then presented to the results of eight such games and had to answer questions (see Fig. 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. 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.

```
##                            Condition
## 1         Independence condition
## 2 Conflict of interest condition
##
## 1                                                                        The three advisors are i
## 2 The three advisors have already invested in one of the three options, the same option
```

Stimuli.   Identical to experiment 4, all participants see all four conditions, with two stimuli (i.e. game results) per condition. Each participant therefore sees eight stimuli in total (4 convergence levels x 2 stimuli).

Table 5

	Accuracy		Competence	
	Conflict of interest	Independent	Conflict of interest	Independent
opposing majority (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)

Dependent variables. To assess accuracy, we asked participants: "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. As for competence, we asked participants: "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 (2)"extremely competent").

Results. 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, using the `lme4` package and its `lmer()` function in R. Figure 7 visualizes the results and table 5 contains descriptive results.

Akin to the results of experiment 3 in the numerical setting, 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 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 4 and replicated the results (see blue colored distributions in Fig. 7). We find a positive effect of convergence on both accuracy (Convergence = 12.34 [10.362, 14.311]) and competence (Convergence = 0.56 [0.459, 0.665]). The second set of hypotheses targeted the interaction of informational dependency and convergence (for a visual representation of these interactions, see Fig. 7). In the independence condition, the effect of convergence on accuracy was more positive (Δ Convergence = 3.01 [0.027, 5.988]) than in the conflict of interest condition (Convergence = 9.33 [7.232, 11.426]). Likewise, the effect of convergence on competence was more positive (Δ Convergence = 0.16 [0.014, 0.316]) than in the conflict of interest condition (Convergence = 0.40 [0.291, 0.503]).

Experiment 6

Experiment four and five tested if the results from the first series of experiments hold in a categorical choice setting. In Experiment six we tested a new context factor: the number of choice options. In experiments four and five, scenarios always involved three choice options. Here, we varied between three and ten options. The design we use to manipulate convergence is otherwise identical to experiment 4 (players playing games).

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
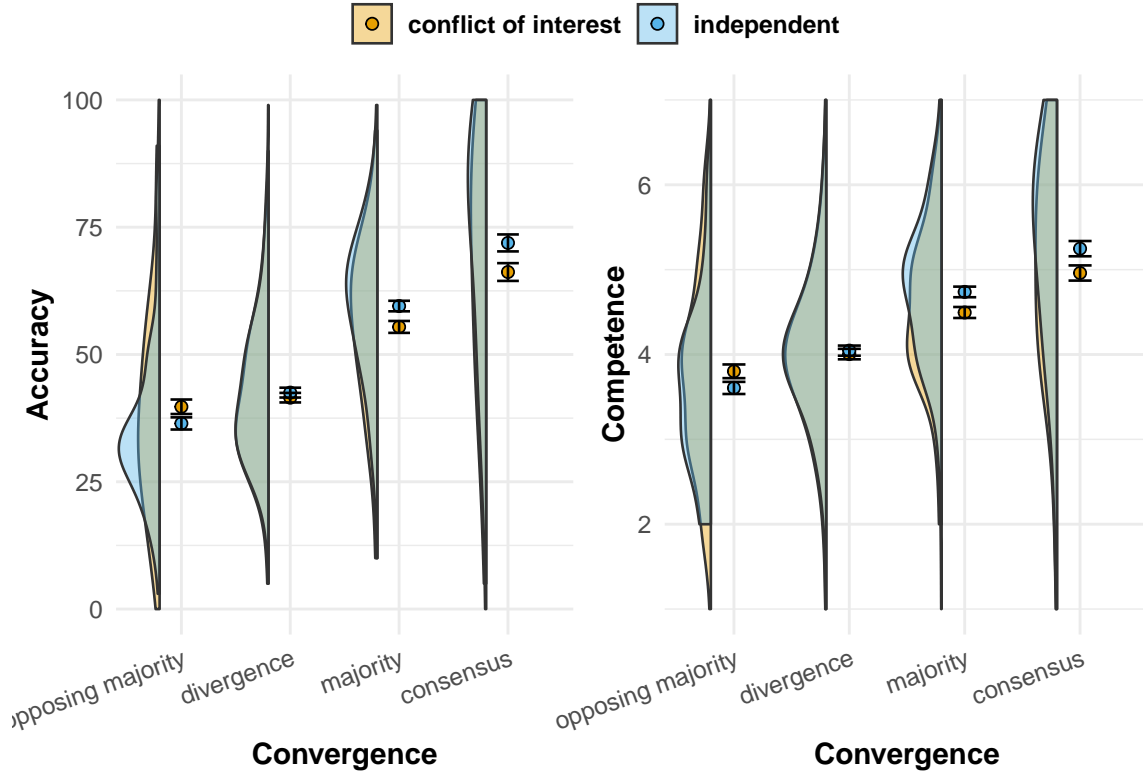
Figure 7. Interaction of convergence and informational dependency.

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.

Participants. We ran a power simulation to inform our choice of sample size. All assumptions and details on the procedure can be found in the `power_Exp6.Rmd` document. 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% (power = 0.872). Due to budget constraints, we considered a sample of 300 participants as good enough, although slightly below our threshold.

Procedure. We used the same experimental setting as in experiment 4.

Design. We manipulated convergence within participants in the same way we did in experiment 4. In addition, between participants, we manipulated the number of choice options. Participants were randomly assigned to either to see stimuli with '3' options, or with '10' options. Participants assigned to the '10' options condition, participants were

assigned to one of two distinct sub-conditions: one in which the range of the answers corresponds to the range of the '3' options condition, and another with increased range (see Appendix XX). We added the increased range condition because we anticipated the possibility that participants might not consider all options as relevant when they only see scenarios in which all answers cluster. We found no differences between the two sub-conditions and collapsed them into a single '10' options condition.

Table 6
Example of a consensus stimulus for the two 'Number of option' conditions

Number of options: 3	Number of options: 10

Stimuli. Identical to experiment 4, all participants saw all four conditions, with two stimuli (i.e. game results) per condition. Each participant therefore saw eight stimuli in total (4 convergence levels x 2 stimuli).

Dependent variables. We used the same items as in experiment 4. To assess 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. As for 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").

Results. 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, using the `lme4` package and its `lmer()` function in R. Figure 7 visualizes the results and table 5 contains descriptive results.

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 (see orange colored distributions in Fig. 8). We find a positive effect of convergence on both accuracy (Convergence = 15.68 [14.112, 17.246]) and competence (Convergence = 0.65 [0.564, 0.736]). We do not find evidence that the number of choice options alters the effect of convergence on either accuracy or competence.

Table 7

	Accuracy		Competence	
	10 options	3 options	10 options	3 options
opposing majority (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)
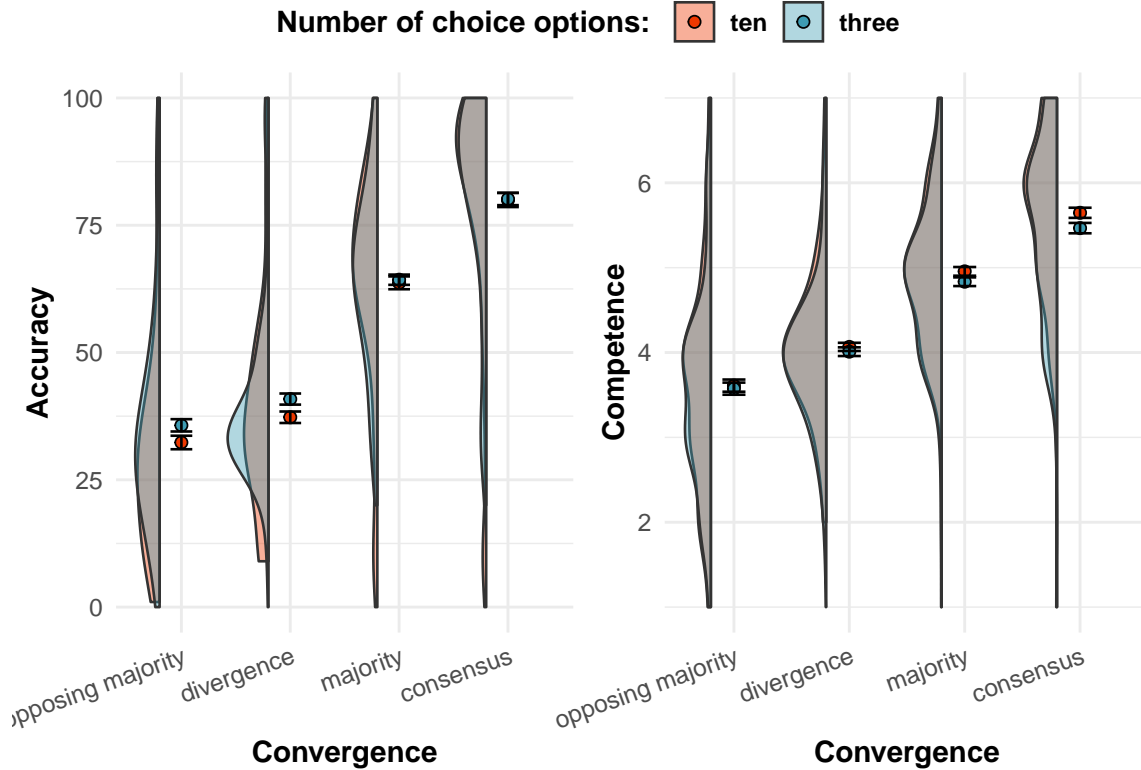
Figure 8. Interaction of convergence and informational dependency.

Discussion

This literature review shows that there is evidence for both adults and children to be susceptible to convergence. It is not always clear, however, whether people infer accuracy from. It also shows (although not really outlined above) that this is only true in boundary conditions - when no bettter information (e.g. strong priors) or information about the competence o informants is available.

When that is the case, peopl tend to favour their own estimates, a phenomenon known as egocentric discounting, or prefer in

Inferences of convergence is one mechanism that work inside a system of epistemic vigiliance.

In many cases, when better knowledge is available, this cognitive mechanism is out-

performed by others.

- egocentric discounting in adults
- perceptual cues + performance hints in children
- older children are more critical of consensus

there are, however, many real-world scenarios in which people have little priors on an estimation task, and little knowledge about the competence of the advisor. In these cases, people have been shown to be especially sensitive to consensus.

Limitations

However, as the original CJT, they still require a minimum of individual-level competence for a majority decision to be accurate (also, non-strategizing and independence). Therefore, without any information about individual voters' competence, the CJT does not justify inferring accuracy from a majority vote.

Data availability. The extracted data used to produce our results are available on the OSF project page (https://osf.io/96zbp/?view_only= d2f3147f652e44e2a0414d7d6d9a6c29).

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/96zbp/?view_ only=d2f3147f652e44e2a0414d7d6d9a6c29).

Competing interest. The authors declare having no competing interests.

References

Austen-Smith, D., & Banks, J. S. (1996). Information Aggregation, Rationality, and the Condorcet Jury Theorem. *The American Political Science Review*, *90*(1), 34–45. Retrieved from http://www.jstor.org/stable/2082796

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, David 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, David 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, David 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

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

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.

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

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

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

Haun, Daniel B. M., Rekers, Y., & Tomasello, M. (2012). Majority-Biased Transmission in Chimpanzees and Human Children, but Not Orangutans. *Current Biology*, *22*(8), 727–731. https://doi.org/10.1016/j.cub.2012.03.006

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

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

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

Leeuwen, E. J. C. van, Cohen, E., Collier-Baker, E., Rapold, C. J., Schäfer, M., Schütte, S., & Haun, D. B. M. (2018). The development of human social learning across seven societies. *Nature Communications*, *9*(1), 2076. https://doi.org/10.1038/s41467-018-04468-2

Leeuwen, E. J. C. van, Kendal, R. L., Tennie, C., & Haun, D. B. M. (2015). Conformity and its look-a-likes. *Animal Behaviour*, *110*, e1–e4. https://doi.org/10.1016/j.anbehav.

2015.07.030

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

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

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, Thomas 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

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

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

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

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

Yaniv, I. (2004a). 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. (2004b). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78.

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

Appendix A

Are we justified in inferring competence and accuracy from convergence? To provide a normative answer, 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. 9).
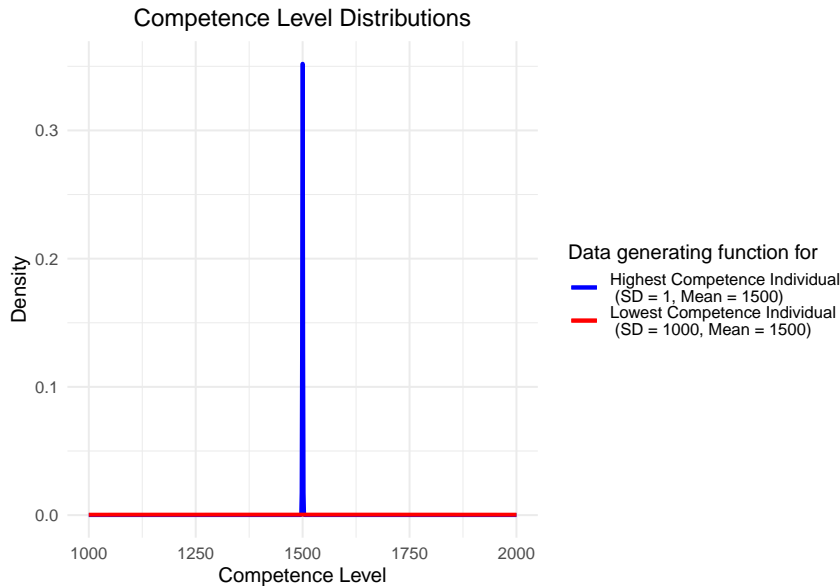
Figure 9. Range of possible data generating functions for individuals.

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. 10).
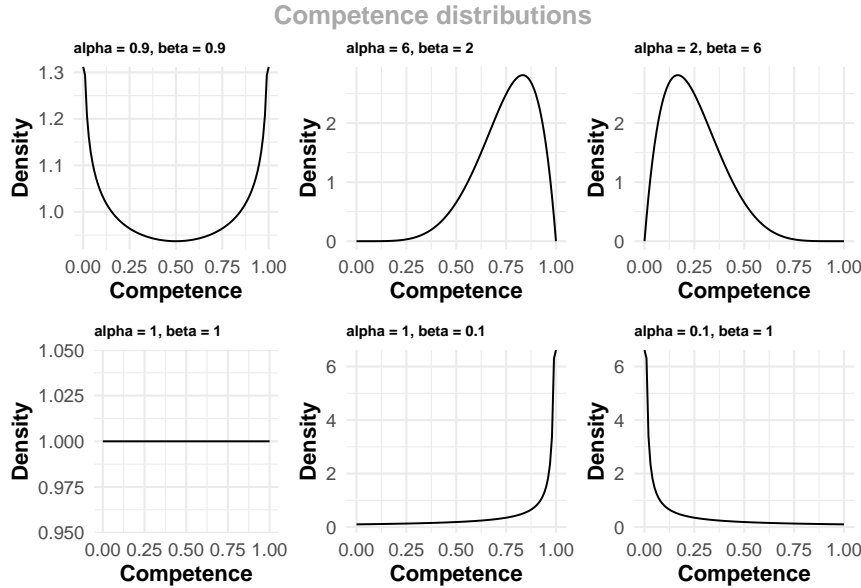
Figure 10. The different population competence distributions we considered in our simulations.

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 100000 (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. 16 to 21, for accuracy, and Fig. 22 to 27, 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. 11). As for accuracy, we find a positive correlation between convergence and competence across all underlying competence distributions. However, this correlations are weaker than for accuracy, and do not increase with sample size (see Fig. 11).

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 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.
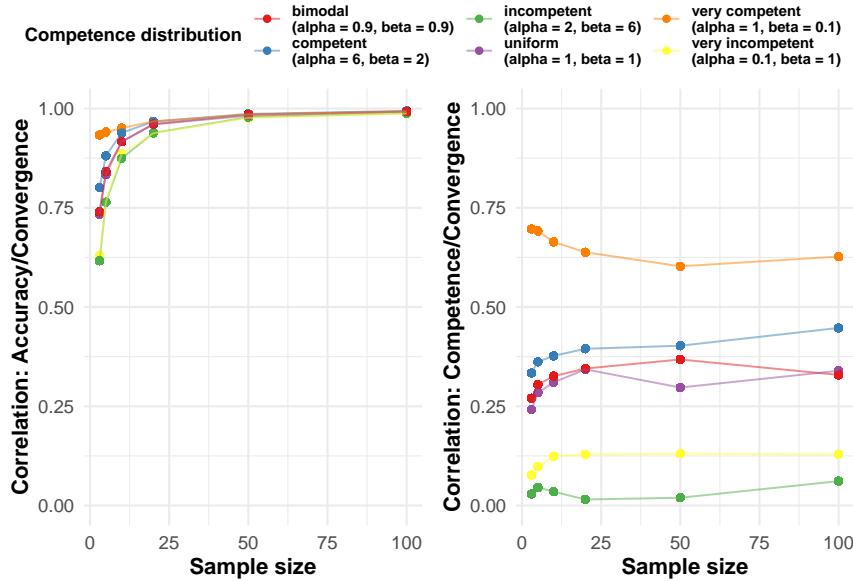
Figure 11. Correlation between accuracy (left) or competence (right) and convergence, as a function of sample size, grouped by the data underlying competence distributions.

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. 10). 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 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 999000 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. 12 shows the average accuracy, and Fig. 13 the average competence value for each share of votes, for different

competence levels and varying the sample size. Fig. 14 and Fig. 15 display the same relationship, but varying the number of choice options instead. The figures dispaly 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.

Number of choice options: 3; Iterations per sample size: 100; Population size per iteration: 999
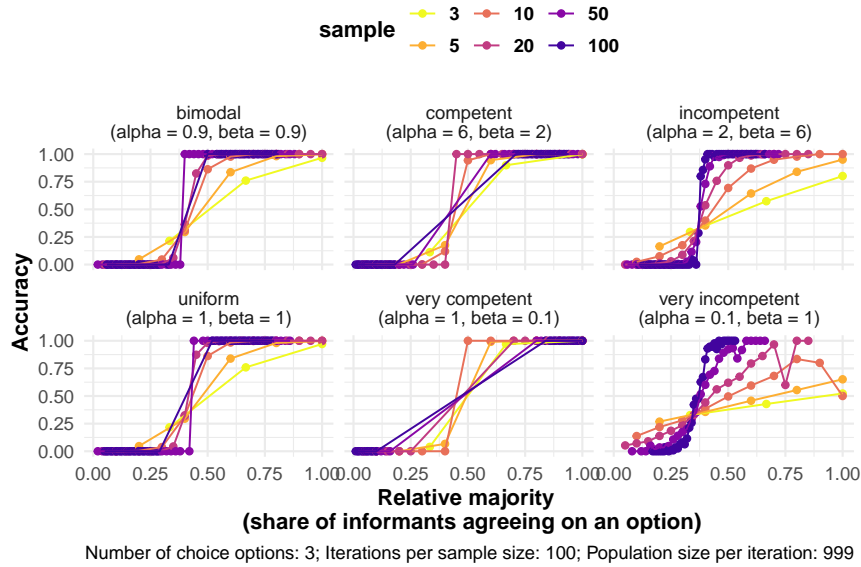
Figure 12. Accuracy as a function of vote share for an option, for different population competence distributions and sample sizes. Points represent averages across all simulations within the respective segment. The number of choice options is three.

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 numeric and categorical choice settings.
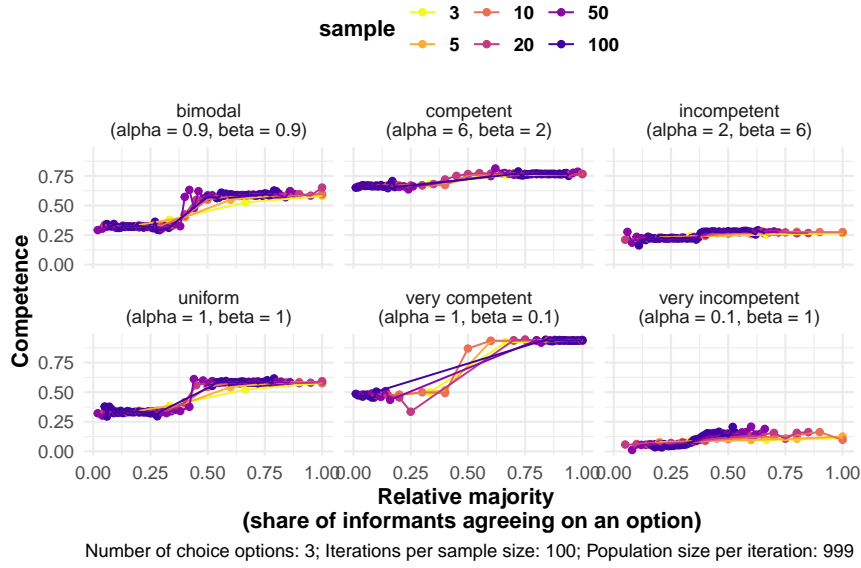
Figure 13. Competence as a function of vote share for an option, for different population competence distributions and sample sizes. Points represent averages across all simulations within the respective segment. The number of choice options is three.
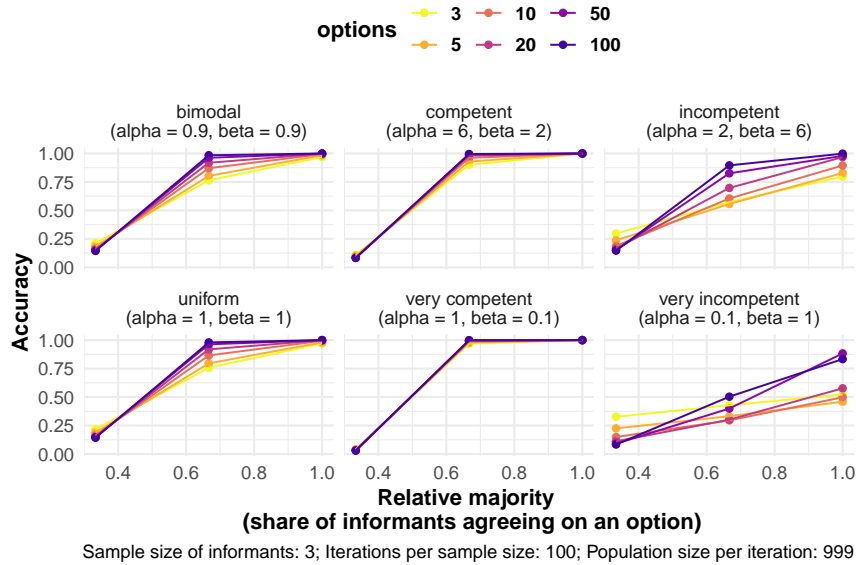
Figure 14. Accuracy as a function of vote share for an option, for different population competence distributions and number of choice options. Points represent averages across all simulations within the respective segment. The sample size is three.
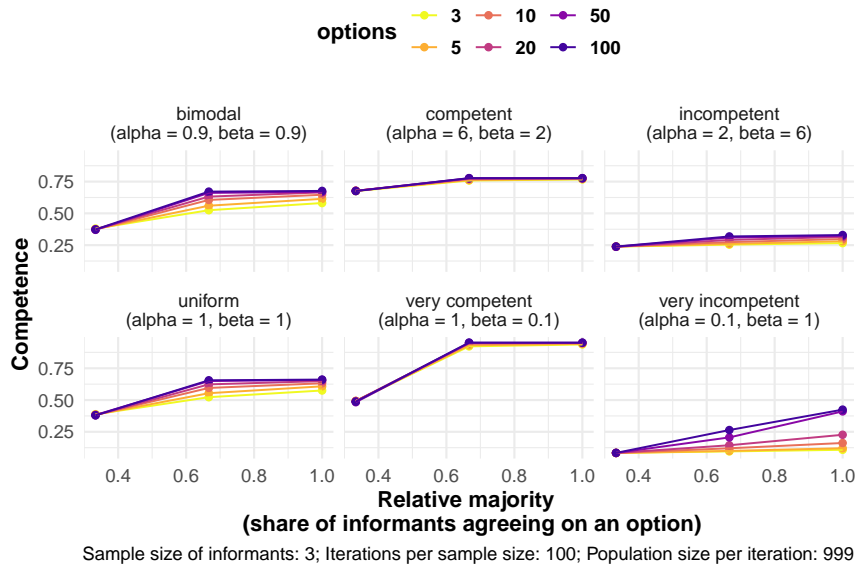
Figure 15. Competence as a function of vote share for an option, for different population competence distributions and number of choice options. Points represent averages across all simulations within the respective segment. The sample size is three.
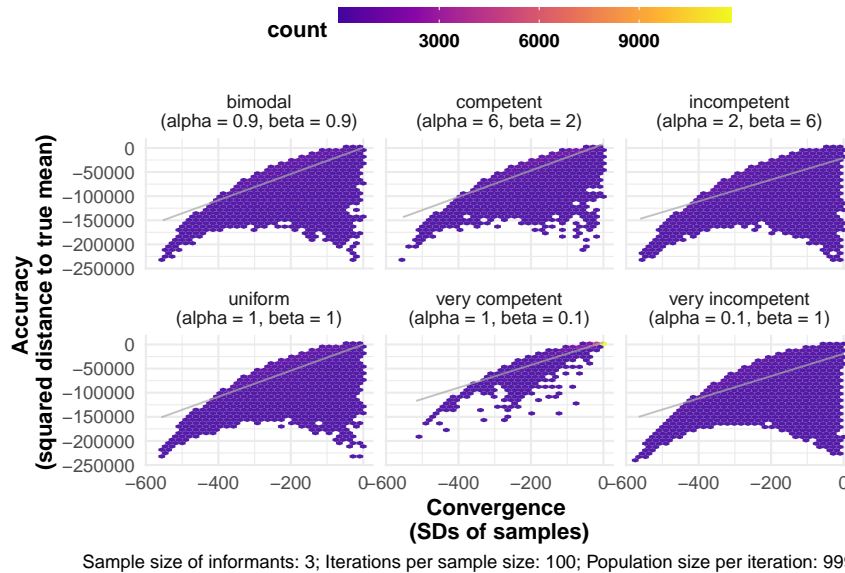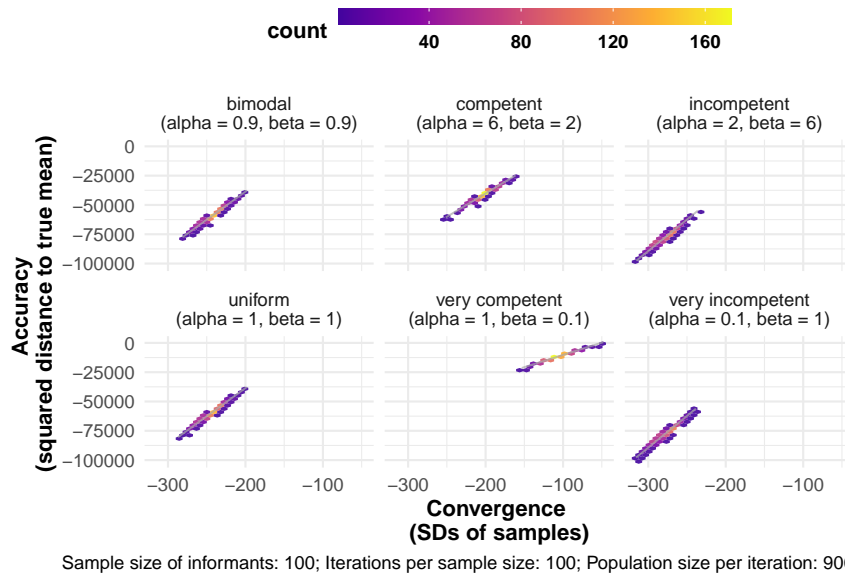

Figure 16. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.
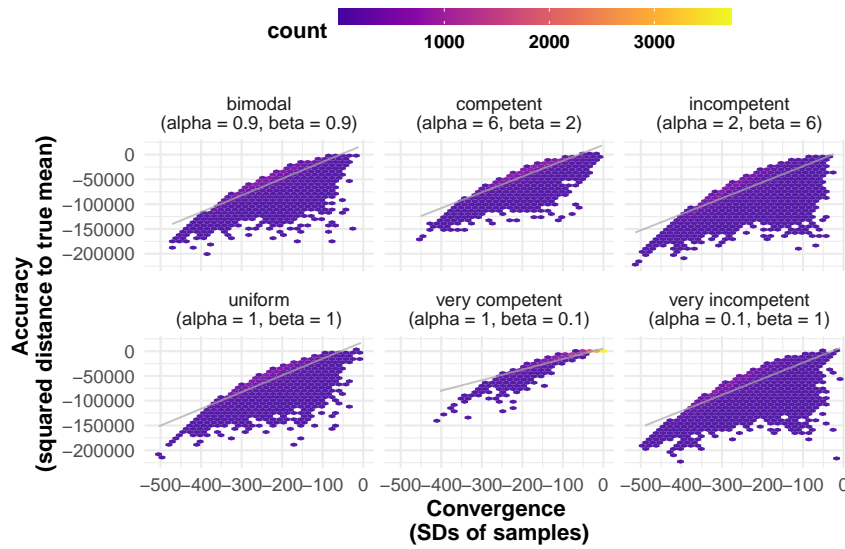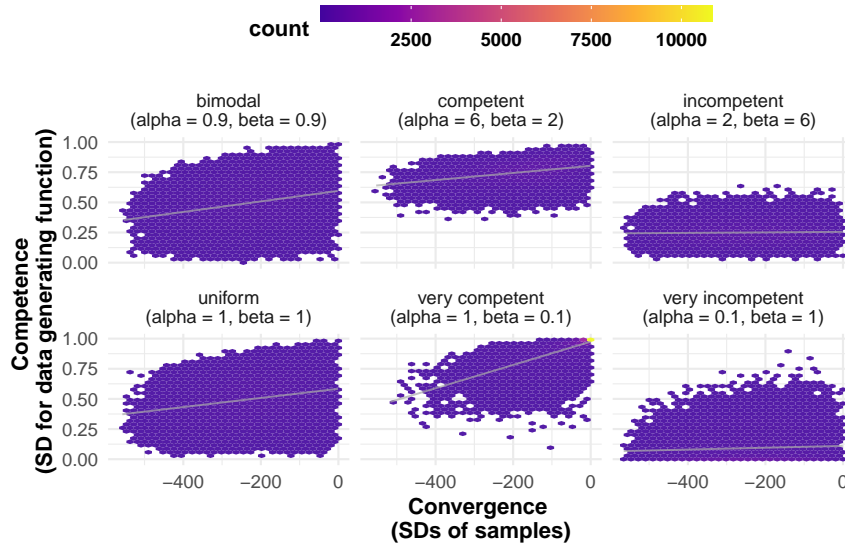
Figure 17. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.
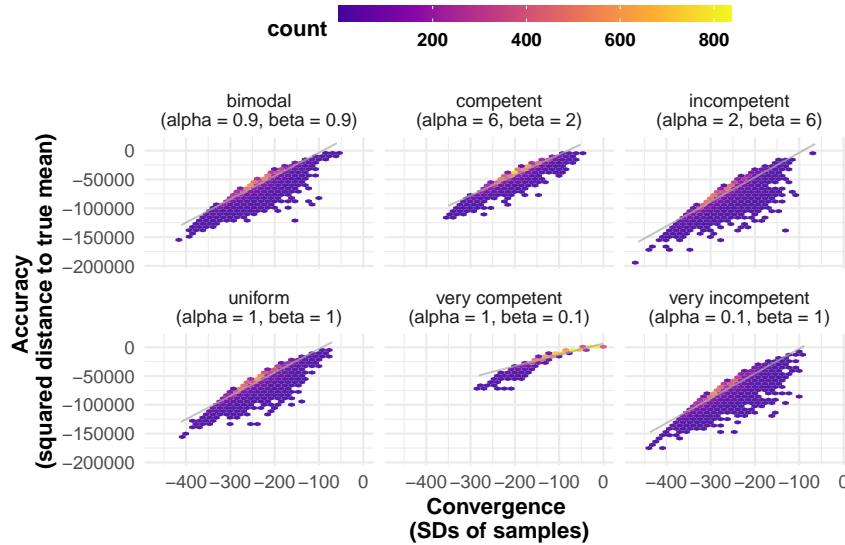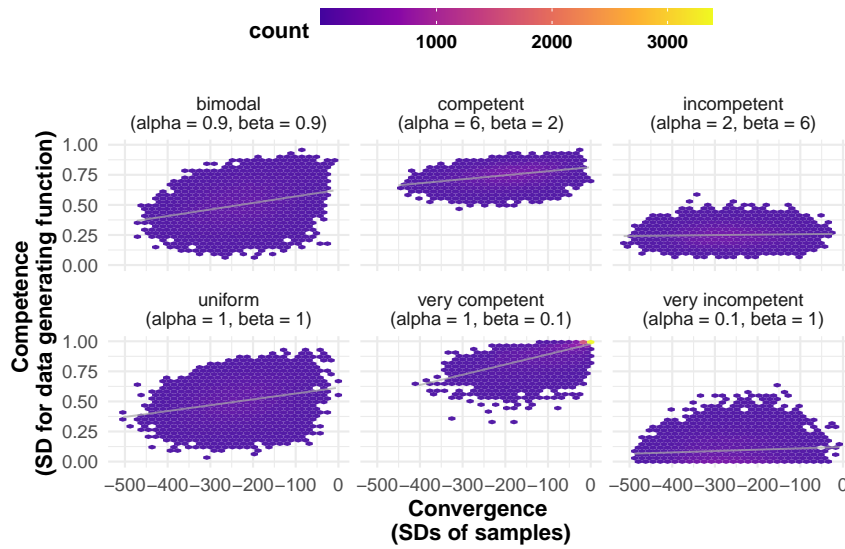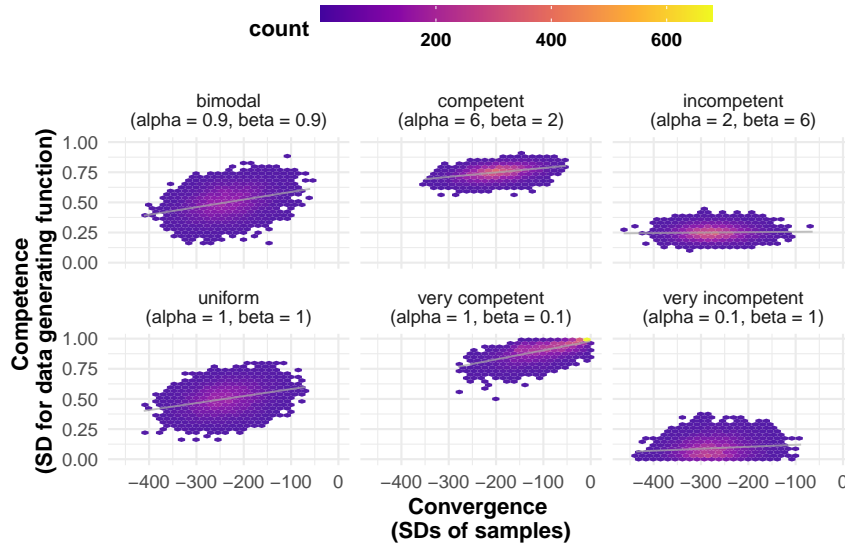

Figure 18. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.
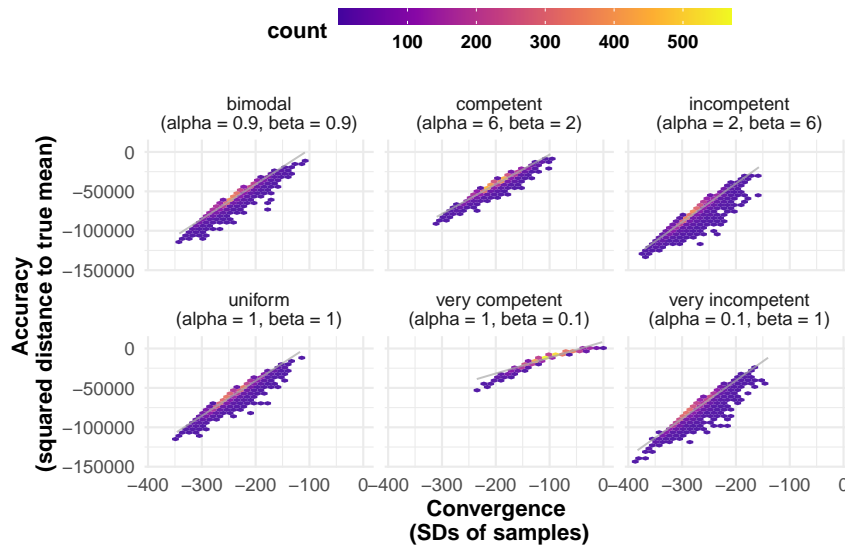
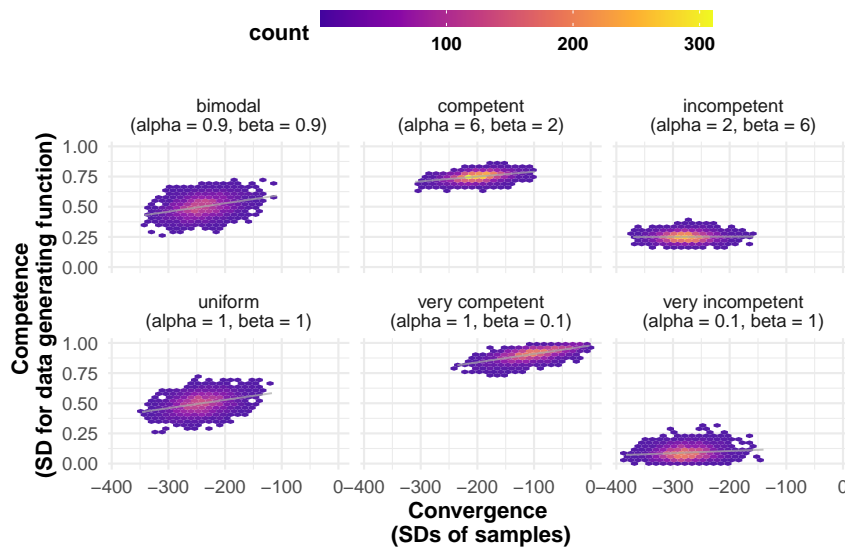*Figure 19*. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.
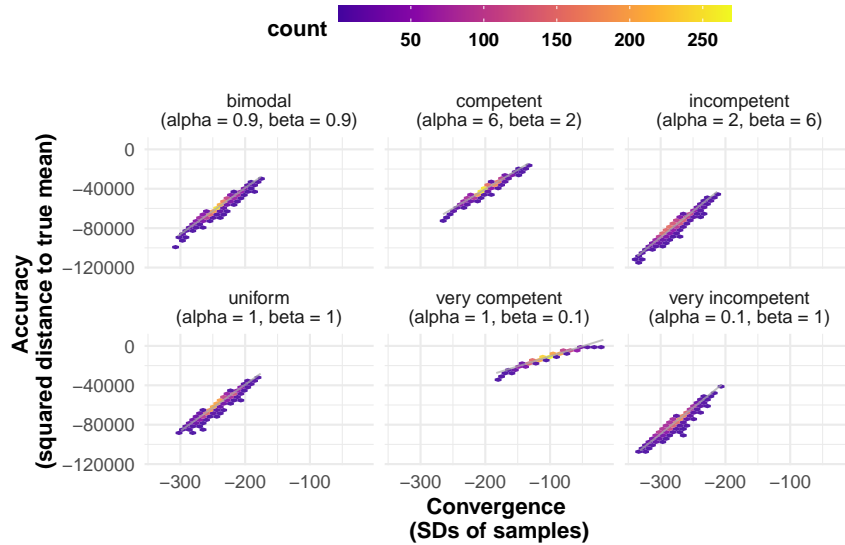
Figure 20. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.

Figure 21. Simulation results showing the relationship between convergence and accuracy for different population competence distributions.

Figure 22. Simulation results showing the relationship between convergence and competence for different population competence distributions.

Figure 23. Simulation results showing the relationship between convergence and competence for different population competence distributions.
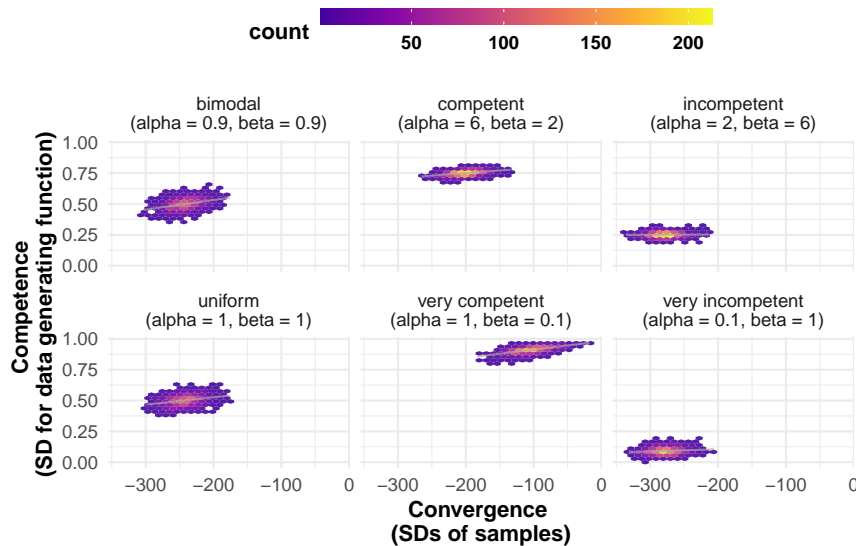

Figure 24. Simulation results showing the relationship between convergence and competence for different population competence distributions.
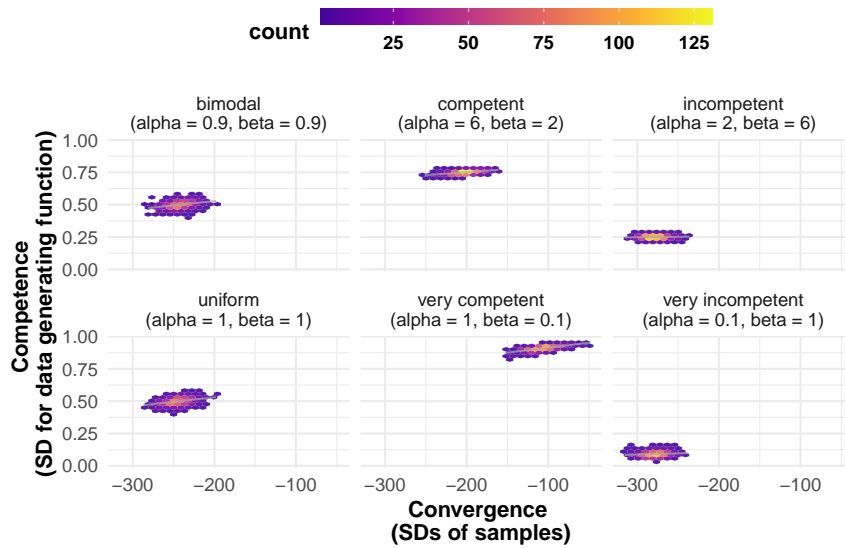
Figure 25. Simulation results showing the relationship between convergence and competence for different population competence distributions.

Figure 26. Simulation results showing the relationship between convergence and competence for different population competence distributions.

Sample size of informants: 100; Iterations per sample size: 100; Population size per iteration: 900

Figure 27. Simulation results showing the relationship between convergence and competence for different population competence distributions.