How social influence can undermine the wisdom of crowd effect

Jan Lorenz[a,1,2], Heiko Rauhut[b,1,2], Frank Schweitzer[a], and Dirk Helbing[b,c,d]

[a]Chair of Systems Design, ETH Zürich, CH-8032 Zurich, Switzerland; [b]Chair of Sociology, in particular of Modeling and Simulation, ETH Zürich, CH-8092 Zurich, Switzerland; [c]Santa Fe Institute, Santa Fe, NM 87501; and [d]Collegium Budapest–Institute for Advanced Study, 1014 Budapest, Hungary

Edited by Burton H. Singer, University of Florida, Gainesville, FL, and approved April 13, 2011 (received for review June 23, 2010)

Social groups can be remarkably smart and knowledgeable when their averaged judgements are compared with the judgements of individuals. Already Galton [Galton F (1907) *Nature* 75:7] found evidence that the median estimate of a group can be more accurate than estimates of experts. This wisdom of crowd effect was recently supported by examples from stock markets, political elections, and quiz shows [Surowiecki J (2004) *The Wisdom of Crowds*]. In contrast, we demonstrate by experimental evidence (*N* = 144) that even mild social influence can undermine the wisdom of crowd effect in simple estimation tasks. In the experiment, subjects could reconsider their response to factual questions after having received average or full information of the responses of other subjects. We compare subjects' convergence of estimates and improvements in accuracy over five consecutive estimation periods with a control condition, in which no information about others' responses was provided. Although groups are initially "wise," knowledge about estimates of others narrows the diversity of opinions to such an extent that it undermines the wisdom of crowd effect in three different ways. The "social influence effect" diminishes the diversity of the crowd without improvements of its collective error. The "range reduction effect" moves the position of the truth to peripheral regions of the range of estimates so that the crowd becomes less reliable in providing expertise for external observers. The "confidence effect" boosts individuals' confidence after convergence of their estimates despite lack of improved accuracy. Examples of the revealed mechanism range from misled elites to the recent global financial crisis.

collective judgment | estimate aggregation | experimental social science | swarm intelligence | overconfidence

Under the right circumstances, the average of many individuals' estimates can be surprisingly close to the truth, although their separate values lie remarkably far from it. There is evidence from guessing tasks (1) and problem-solving experiments (2–4) that the aggregate of many people's estimates tends to be closer to the true value than all of the separate individual or even expert guesses. This phenomenon is referred to as the "wisdom of crowd effect" (5). Even individuals can apply this mechanism and improve their decisions by averaging multiple perspectives from their own reasoning (6–8).

In the following, we will call an aggregate measure of a collection of individual estimates "wise" if it comes close to the true value, even though the individual estimates are largely dispersed. In this case, single estimates are likely to lie far away from the truth, whereas its aggregate lies close to it. The wisdom of crowds effect works if estimation errors of individuals are large but unbiased such that they cancel each other out. Thus, the heterogeneity of numerous decision-makers generates a more accurate aggregate estimate than the estimates of single lay or expert decision-makers. This can be quantified by the "diversity prediction theorem" (9), which states that the collective error is equal to the average individual error minus the group's diversity (*Materials and Methods*).

The wisdom of crowd effect is a statistical phenomenon and not a social psychological effect, because it is based on a mathematical aggregation of individual estimates. Nevertheless, social influence plays a role in individual decision-making and affects individual estimating. Therefore, social influence can also have an impact on the statistical aggregate and the resulting collective wisdom of the respective crowd. As social influence among human group members may trigger individuals to revise their estimates (10), it can have a substantial impact on the statistical wisdom of crowd effect in societies. When individuals become aware of the estimates of others, they may revise their own estimates for various reasons: People may suspect that others have better information (11, 12), they may partially follow the wisdom of the crowd (13), there may be peer pressure toward conformity (14–17), or the group may engage in a process of deliberation about the facts. An example of deliberation about facts would be the task of the Intergovernmental Panel on Climate Change.

Although there is evidence from social psychology that humans have an inclination to adjust their opinions to those of others so that they gradually converge toward consensus (4, 18), many existing studies have two drawbacks. First, consensus formation has often been investigated for questions for which there are no well defined correct answers. Typical examples are attitudes toward abortion, nuclear power, war on terror, or election polls. Another case are "cultural" markets of musical tastes, in which it has been demonstrated that almost any song of average quality may become a hit if social influence is introduced by publishing the number of downloads (19). In this case, the popularity of a song and its perceived quality emerge through the process of interactive downloading and rating. The herding effects created in this way prevent an objective measurement of quality. Therefore, such settings do not reveal whether social influence works in favor or to the disadvantage of the wisdom of the group.

The second drawback of existing studies is that correct answers are often not rewarded with monetary incentives, which makes correct estimations less important and conformity costless. In contrast, we study the interconnection between social influence and the wisdom in groups by using factual questions and monetary incentives for good individual guesses. First, this allows disentanglement of social influence from the wisdom of the group. Second, the incentives trigger the ability to use information of others only for improving own estimates and not for aligning with others for the sake of conformity. This allows investigation of how social influence affects the group's wisdom.

Author contributions: J.L., H.R., F.S., and D.H. designed research; J.L. and H.R. performed research; J.L. and H.R. analyzed data; and J.L. and H.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]J.L. and H.R. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: post@janlo.de or rauhut@gess.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1008636108/-/DCSupplemental.

In this article, we will demonstrate that social influence has three effects that can undermine the wisdom of crowds. Two of the effects are changes of statistical aggregates and one is psychological. The "social influence effect" describes the fact that social influence diminishes the diversity of the group without improvements in accuracy. The "range reduction effect" moves the position of the truth to peripheral regions. This corrupts the wisdom of the crowd from an observer's perspective in the sense that the group becomes less reliable in guiding decision-makers. The "confidence effect" boosts individuals' confidence. [A related effect is known as overconfidence (20–22).] This boost in confidence subverts the wisdom of crowd effect psychologically, because individuals' perceptions contradict the aggregate outcomes of a lack of improvements in accuracy and a decreased reliability of the group's range of estimates.

Recent theoretical studies have analyzed the wisdom of crowd effect in the context of information diffusion in networks of simulated agents in which too little and too much dissemination of information inhibits actors to find optimal solutions, because there is a need to maintain diversity on the one hand and information flow on the other (23). However, from an empirical point of view, it is not clear whether the reduction of diversity is strong enough to severely undermine the wisdom of crowd effect in reality. In the following, we demonstrate by means of laboratory experiments that the wisdom of crowd effect is undermined in all three of the aforementioned ways, even when social influence is relatively mild. Furthermore, the severity of statistical undermining is quantified by a new indicator, which measures the centrality of the truth within a given set of estimates.

Experimental Design

To identify how social influence affects the wisdom of crowds, we conducted a laboratory experiment with real monetary stakes. A total of 144 participants were recruited from a sample of more than 8000 students at Eidgenössische Technische Hochschule Zürich in Zurich, Switzerland. Twelve experimental sessions were conducted, each consisting of 12 subjects.

The participants had to solve six different estimation tasks testing their real-world knowledge regarding geographical facts and crime statistics. We selected questions for which subjects were unlikely to know the exact answer but also avoided those for which they did not have a clue at all. Each question was simultaneously presented to all subjects in a computer laboratory by using the z-tree software (24). Each subject sat in an isolated cubicle in front of a computer with no visual, verbal, or chat contact with each other and was asked to enter the estimates privately without communicating with other subjects. All subjects were told all details of the experimental procedures and payments by printed instructions, which described the payment rules, the anonymity warranty, the obligation to adhere to the no-communication policy, and the obligation to make no use of any auxiliary devices such as the Internet or mobile phones. A test in the beginning ensured that subjects understood the payment rules.

In the first estimation period, all subjects had to respond to the first knowledge question on their own. After all 12 group members made an estimate, everybody was asked to give another estimate. In total, we elicited five consecutive responses for each knowledge question. Additionally, we elicited after the first and final estimate for each question subjects' confidence in their estimate on a six-point Likert scale (1, very uncertain; 6, very certain). The confidence values were not communicated to others. After the fifth (and final) estimate of each question, an evaluation was provided that included the true answer, the five estimates of the respective subject, the payments for each of the five estimates, and the total payment for all estimates for one question.

We tested three different information conditions regarding what each subject learned about the estimates of the other subjects. Subjects could base their second, third, fourth, and fifth estimate on either aggregated or nonaggregated information regarding other people's estimates. Two of our conditions were different operationalizations of social influence, and the third condition served as a control condition without social influence. The reason to use two different kinds of social influence was to demonstrate the robustness of our effects with regard to the specific kind of social influence.

In the "aggregated information" condition, subjects could reconsider their estimate after having received the average (arithmetic mean) of all 12 estimates of the former round. Subjects were also reminded about their last estimate from the previous round. In the "full information" treatment, subjects received a figure of the trajectories of all subjects' estimates from all previous rounds. In this figure, the estimates of each of the 12 subjects over all previous rounds were represented by one line, adding up to 12 separate lines (one for each subject). In addition, the numerical values of all subjects' estimates from the last round were presented and subjects were reminded about their own last estimate.* The "no information" treatment served as a control and revealed no information about the other subjects' estimates. In this condition, subjects had to answer the same question five times on their own. They were reminded about their latest estimate.

In each session, two questions were posed in the control, two in the aggregated, and two in the full information treatment. The order of questions and the order of treatments was randomized across experimental sessions. This means that the same six knowledge questions were posed in each experimental session, but in a different order and in different information treatments. The allocation of subjects to a random sequence of questions and information treatments ruled out order effects with regard to questions and treatments.

Subjects received monetary payments for each good estimate. Possible rewards were 4, 2, or 1 points if their estimates fell into the 10%, 20%, or 40% intervals around the truth; otherwise, they received no points. The rewards applied to all rounds to make sure that individuals took all of their decisions seriously. The correct answer and the rewards for all five estimates were only disclosed to the subjects after the fifth estimate. This reward structure introduced incentives only to find the truth and avoided incentives to conform with others. Furthermore, there were no incentives for strategic considerations. For example, there was no benefit of being better than others or of misleading others, because this did not affect individuals' payments. There was also no possibility to help others by deviating from the strategy to find the best estimate. Therefore, our experimental design put subjects into a situation in which they would try to get as close to the truth as possible by using their own knowledge and the estimates of others.

Materials and Methods provides more in-depth information regarding the knowledge questions, payment rules, and data structure. *SI Appendix* contains all the details of the experimental procedures that were presented to the subjects. Dataset S1 contains the raw data.

Results

Aggregation of the Wisdom of Crowds. The empirical analysis of the wisdom of crowds requires an appropriate aggregation measure. [Already in reply to Galton's article (1), there was a discussion of how to find the best aggregation measure for the wisdom of crowds (25, 26). It is common to use the unweighted

*In real-life situations with social influence, there may be additional effects, from which our experiment has abstracted: this includes competition, group pressure, and authority effects. For example, a criminologist could say: "I know the number of victims." In contrast to such possibilities, our comparably mild and parsimonious kind of information feedback has the advantage that it enables a particularly controlled experimental setting in which there is little ambiguity about which kind of information feedback and social influence played a role.

Table 1. **The wisdom of crowd effect exists with respect to the geometric mean but not with respect to the arithmetic mean**

Question	True value	Wisdom-of-crowd aggregation		Median
		Arithmetic mean	Geometric mean	
1. Population density of Switzerland	184	2,644 (+1,337.2%)	132 (−28.1%)	130 (−29.3%)
2. Border length, Switzerland/Italy	734	1,959 (+166.9%)	338 (−54%)	300 (−59.1%)
3. New immigrants to Zurich	10,067	26,773 (+165.9%)	8,178 (−18.8%)	10,000 (−0.7%)
4. Murders, 2006, Switzerland	198	838 (+323.2%)	174 (−11.9%)	170 (−14.1%)
5. Rapes, 2006, Switzerland	639	1,017 (+59.1%)	285 (−55.4%)	250 (−60.9%)
6. Assaults, 2006, Switzerland	9,272	135,051 (+1,356.5%)	6,039 (−34.9%)	4,000 (−56.9%)

The aggregate measures arithmetic mean, geometric mean, and median are computed on the set of all first estimates regardless of the information condition. Values in parentheses are deviations from the true value as percentages.

arithmetic mean, but there are many reasonable alternatives, giving ample room for adjustments or "tuning" (27–30).] In our case, the arithmetic mean performs poorly, as we have validated by comparing its distance to the truth with the individual distances to the truth. In only 21.3% of the cases is the arithmetic mean closer to the truth than the individual first estimates. This is because the estimates of our type of questions are not normally distributed but right-skewed. In other words, the majority of estimates are low and a minority of estimates are scattered in a fat right tail, as it is the case for log-normal distributions.

As a large number of our subjects had problems choosing the right order of magnitude of their responses, they faced a problem of logarithmic nature (31). When using logarithms of estimates, the arithmetic mean is closer to the logarithm of the truth than the individuals' estimates in 77.1% of the cases. This confirms that the geometric mean (i.e., exponential of the mean of the logarithmized data) is an accurate measure of the wisdom of crowds for our data (Table 1). In particular, log-normal distributions are justified for variables with high variance with a range of positive values only (32), which is the case for our data.[†] We further divided each estimate by the respective true value before taking the logarithm to make the distributions of estimates comparable across different questions. This yielded approximately normal distributions and true values corresponding to zero.

Social Influence Effect. The first kind of undermining of the wisdom of crowds is a statistical effect, which we call social influence effect. This effect denotes the fact that social influence diminishes diversity in groups without improving its accuracy. This means that, on average, groups cannot make use of information exchange, but engage in a convergence process that does not yield improvements of the collective.[‡]

Fig. 1 gives evidence for the social influence effect. Here, group diversities and collective errors for each question and each time step are computed on the transformed data set. Fig. 1*A* shows for each information condition exemplary responses to one question over the five time steps in one group. By comparing the no information condition with the aggregated and full information conditions, the typical effects of social influence can clearly be seen. It is evident that social influence promotes a convergence of estimates. Fig. 1*B* shows, for the same exemplary sessions, core ranges of estimates and two types of aggregate measures: the arithmetic mean and the geometric mean. Fig. 1*C* provides the respective numbers for the exemplary sessions. We provide a test for the complete data in Fig. 1 *D* and *E*, demonstrating that social influence strongly reduces the group's diversity[§] without significantly reducing its collective errors.[¶]

The robustness of the social influence effect is supported by further statistical significance tests (*SI Appendix*). A Kolmogorov–Smirnov test confirms that the distribution of estimates changes significantly if social influence is allowed for. This applies particularly to the variance of the distribution, as an F test shows. In addition, Kolmogorov–Smirnov and t tests for the group data demonstrate that the group diversity is significantly reduced under social influence, whereas the collective error changes only slightly. In the control condition without social influence, these effects are almost null.

Range Reduction Effect. Let us take the perspective of a person who, or government that, needs advice and requests expertise from different specialists. If all predictions are narrowly distributed around a wrong value, a decision-maker would gain confidence in advice that is actually misleading. In fact, the close clustering around a wrong value makes the group less "wise" in the sense that the group delivers a wrong hint regarding the location of the truth. This is the case because the truth would not be located centrally but at outer regions of the range of estimates. We quantify this by a wisdom-of-crowd indicator, which

[†]Note that the framework of the diversity prediction theorem (9) can also be applied to logarithmically transformed data. For the case of logarithmically transformed data, the collective error of the logarithms is the logarithm of the geometric mean and one SD is the logarithm of the geometric SD. Considering the logarithmic nature of our data, one may argue that the geometric mean would have been a better design choice than the arithmetic mean for the information feedback in the aggregated information condition. However, this measure is hard to understand for most subjects because it necessitates confidence with logarithmic transformations. As the simple average (i.e., arithmetic mean) is known from daily life, this information is more meaningful for subjects. Hence, we decided for the arithmetic mean.

[‡]The empirical measurement of the social influence effect requires questions with moderate difficulty. In particular, subjects should not have a precise factual knowledge of an issue, because this would prevent adaption and social influence. We can empirically confirm that this was not the case for our questions and subjects: in only 1.5% of all cases, subjects responded at all five times in the most inner payment range of one particular question. This means in absolute values that 13 of 864 consecutive response runs were responded in the full payment range (144 subjects responded to six questions in a run of five consecutive responses). Three of these 13 "high-success runs" were performed by the same person and two from another person. All other high-success runs were performed by different persons.

[§]It deserves to be mentioned that the initial diversity seems to be higher in the no information condition. It could be that subjects anticipate to feel uneasy if their published estimates are too distant from those of others. This could foster that their initial estimates tend to be more "conservative" in the conditions with information feedback. Interestingly, this discrepancy in initial variance is mainly caused by the questions about crime statistics and not about geographical facts.

[¶]Note that the collective error slightly declines under social influence, especially in the aggregated information condition, which is partially supported by the significance tests (*SI Appendix*). This is a result of two empirical facts. First, the distributions of estimates are right-skewed. As a consequence, the arithmetic mean is usually much larger than most estimates and also much larger than the true value. Second, it is an empirical fact for our choice of questions that the geometric mean (which is our aggregation measure to compute the collective error) is always slightly lower than the true value (Table 1). The mechanism of presenting the arithmetic mean in the aggregated condition thus triggers an upward drift toward the true value. This issue is interesting but deserves future studies, as this effect may be different for different sets of questions.

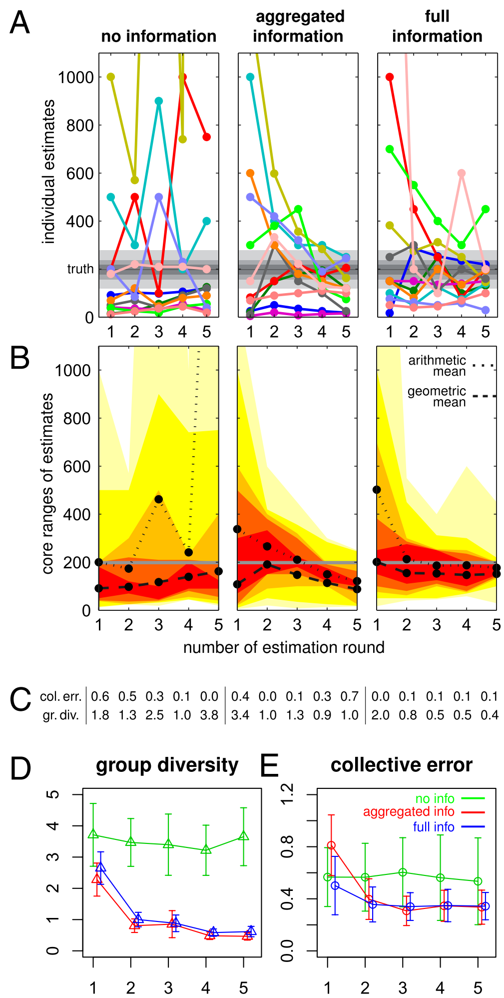
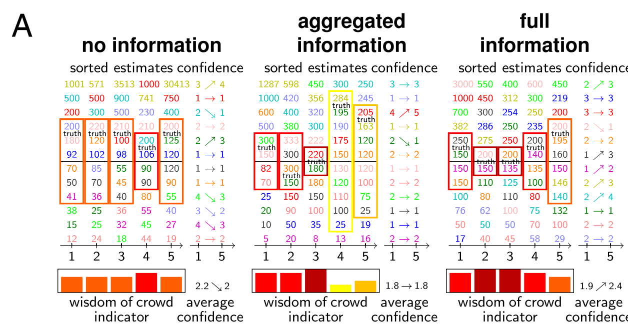


Fig. 1. Social influence effect: Social influence diminishes group diversity without diminishing the collective error. (*A*) Typical examples of experimental sessions for all three information conditions, displaying the five individual responses to the question, "How many murders were registered in Switzerland in 2006?" In the no information condition, there is no convergence of estimates, whereas estimates converge in the aggregated and full information conditions. Each subject's decision over the five consecutive responses is represented by a different color. Black lines correspond to the true value; surrounding gray zones represent the areas in which the subjects received monetary payments (payments decrease from dark to light gray). (*B*) Representation of the same data in aggregated form. The arithmetic mean is represented by a dotted line and the geometric mean by a dashed one. The ranges of the 4, 6, 8, 10, and 12 inner estimates are represented by red, dark orange, light orange, dark yellow, and light yellow colors. (*C*) Values of the collective error (squared deviation of group average from the truth) and group diversity (average squared deviation from the group average) for all 12 estimates at each time step for the same exemplary session as in *A* and *B* (computed on logarithms of the estimates and normalized by true values). (*D*) Average group diversity for all information conditions over time (each data point represents 24 groups, aggregated over all questions; computed on logarithms of estimates and normalized by the respective true value; error-bars represent 10% confidence intervals). (*E*) Average collective error (data and presentation analogous to *D*).

generalizes the concept of "bracketing the truth" (33) to more than two persons. Our indicator considers a group to be maximally wise if the truth lies between the two most central values of all estimates (in our case, between the sixth and the seventh largest of 12 estimates). If the four most central values are needed to enclose the true value, the level of wisdom is considered to be lower, and if the six most central values are needed, it is even lower, and so forth. If it lies outside the range of estimates of all individuals, there is no wisdom of crowd effect at all (a precise definition is provided in *Methods and Materials*).

Fig. 2*A* shows bar plots of the wisdom-of-crowd indicator over time for the three treatments for the same exemplary question as in Fig. 1. Furthermore, the corresponding core range of sorted estimates enclosing the true value is reported. The figure demonstrates that the wisdom-of-crowd indicator tends to decline over time under conditions of social influence. This effect is substantial and statistically significant for all questions, which is confirmed by the regression model in Fig. 2*B*. It is revealed that the wisdom of crowd indicator is about one unit lower in conditions with information exchange compared with the control condition. Note that the reduction is stronger under the aggre-



B

	(1) groups' wisdom-of-crowd indicator	(2) individuals' increase in confidence
intercept	3.92***	-0.049
	(15.1)	(-0.97)
aggregated information	-1.39**	0.20**
	(-3.29)	(3.08)
full information	-0.98*	0.28***
	(-2.21)	(4.14)
N	288	864

t statistics in parentheses (robust std. errors), $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Fig. 2. Although the wisdom of the crowd diminishes over time, individuals gain confidence in their own estimates. (*A*) Sorted estimates over successive rounds for the same exemplary sessions as displayed in Fig. 1. Boxes represent the most inner estimates that still include the true value. The range of the box depicts the wisdom-of-crowd indicator. The maximum value of 6 represents the highest value for the wisdom of the crowd meaning, that 6 values are below and 6 above the true value. A wisdom-of-crowd indicator of 0 denotes that the truth is outside of the range of estimates. The colors of the boxes and bars for the wisdom-of-crowd indicator are analogous to *B* (for values 0–5, aubergine is additionally introduced for the value of 6). Colored numbers represent individuals (analogous to colors in Fig. 1*A*). The right column shows the certainty of the estimate as reported by the respective individual after estimate 1 and estimate 5 (1, very uncertain; 6, very certain). These values were not propagated to other subjects. (*B*) Confirmation of exemplary trends by regression models taking all questions into account. Linear regressions on (model 1) the groups' wisdom-of-crowd indicator and (model 2) individuals' change in the certainty of their estimates. The predictors in both models are the experimental treatments, implemented as dummy variables and coded with 1 as the experimental condition (aggregated or full information) and 0 otherwise. The control condition is the reference category, represented by the intercept of the respective regression model. In model 1, the wisdom-of-crowd indicator is calculated for the pooled second, third, fourth, and fifth time step. The first time step is excluded because the initial period had no information feedback and can therefore not yield treatment differences. In model 2, individuals' increase in confidence is the outcome variable, which is the difference between the initial and final individuals' certainty in their estimates. Robust SEs are calculated, taking the clustering within subjects into account.

gated information condition compared with the "full information condition. [This matches previous findings on the effects of social influence in terms of gossip on the behavior, which seems to be stronger if gossip comes from fewer sources (34).] An increase of one unit means that one has to consider one additional person in the upper range and one additional person in the lower range of sorted estimates so that the truth is included in the selected range. This effect demonstrates that the truth becomes less central if social influence is allowed for. Another interpretation of this effect is that the group becomes less reliable in estimating the truth if it has been exposed to social influence.

Confidence Effect. The third kind of undermining of the wisdom of crowds concerns the psychological consequences of the two aforementioned statistical effects. The confidence effect reflects that opinion convergence boosts individuals' confidence in their estimates despite a lack of collective improvements in accuracy. Fig. 2*A* shows the individuals' self-reported change in confidence regarding their initial and final estimates for the same exemplary sessions as studied before. It can be seen that individuals in these exemplary sessions become more confident in the full information condition and less confident in the control condition. We analyze the general effects over all sessions and questions with regression models in Fig. 2*B*. This regression analysis demonstrates that individuals' confidence is substantially and significantly boosted in the aggregated and full information conditions in comparison with the control condition without social influence. We can interpret this psychological effect in comparison with the statistical effects: the confidence measure can be regarded as "subjective"—and self-reported reliability of estimates and the wisdom of crowd indicator as "objective"—statistical measure of reliability. The comparison of both illustrates that social influence undermines the wisdom of crowds by boosting the subjective and decreasing the objective reliability of the crowd.

Discussion

Based on the wisdom of crowd effect, groups can be remarkably accurate in estimating vaguely known facts. From the perspective of decision-makers, it would be valuable to request multiple independent opinions and aggregate these as the basis of their judgements. Real-life examples are predictions of economic growth rates, market potentials, the increase of the world temperature, tax estimations, the assessment of the impact of new technologies, or estimating the amount of finite natural resources.

However, it is hardly feasible to receive independent opinions in society, because people are embedded in social networks and typically influence each other to a certain extent. It is remarkable how little social influence is required to produce herding behavior and negative side effects for the mechanism underlying the wisdom of crowds. In our experiment, we provided just the bare information of the estimates of others (in a similar way as the previous stock price is known to traders trying to make money with their estimates of the fundamental value of a stock). We did not allow for group leader effects, persuasion, or any other kind of social psychological influence. We just provided noncompetitive monetary incentives for the estimation of correct values. These incentives were designed such that the information of others could just be used to update the own knowledge. There was no premium to coordinate with others' opinions.

Our experimental results show that social influence triggers the convergence of individual estimates and substantially reduces the diversity of the group without improving its accuracy. The remaining diversity is often so small that the correct value shifts from the center to outer regions of the range of estimates. Thus, when taking committee decisions or following the advise of an expert group that was exposed to social influence, their opinions may result in a set of predictions that does not even enclose the correct value anymore. From the perspective of decision-makers, such advice may be thoroughly misleading, because closely related, seemingly independent advice may pretend certainty despite substantial deviations from the correct solution.

Psychologically, however, the convergence of estimates significantly boosts individuals' confidence. This confidence gain happens despite a lack of improvements, giving evidence for a psychological trap whereby individuals are led into the false belief of collective accuracy as a result of their convergence. Nevertheless, the statistical effects of undermining are less severe for easier questions and if individuals are more confident in their answers (*SI Appendix*). This gives weight to the conclusion that the negative effects of social influence occur especially in a certain range of question difficulty and individuals' confidence, a conjecture that should be explored in follow-up studies.

Our results underpin the value of collecting individuals' estimates in the absence of social influence. However, in democratic societies, it is difficult to accomplish such a collection of independent estimates, because the loss of diversity in estimates appears to be a necessary byproduct of transparent decision-making processes. For example, opinion polls and the mass media largely promote information feedback and therefore trigger convergence of how we judge the facts. The wisdom of crowd effect is valuable for society, but using it multiple times creates collective overconfidence in possibly false beliefs.

Presumably, herding is even more pronounced for opinions or attitudes for which no predefined correct answers exist. For example, prospective research may investigate herding and consensus formation on predictions of climate change or election outcomes. However, long-term predictions may have short-term consequences on the system itself: pessimistic predictions for climate change may entail international political consequences, or election polls may change the popularity of parties that have been exposed as those with the least support. These feedback loops hinder the disentanglement of herding behavior from the wisdom of crowds.

Materials and Methods

Knowledge Questions, Payment Rules, and Data Structure. The following questions were used:

1. What is the population density in Switzerland in inhabitants per square kilometer?
2. What is the length of the border between Switzerland and Italy in kilometers?
3. How many more inhabitants did Zurich gain in 2006?
4. How many murders were officially registered in Switzerland in 2006?
5. How many rapes were officially registered in Switzerland in 2006?
6. How many assaults were officially registered in Switzerland in 2006?

All questions imply nonnegative or positive numbers as answers. Note that question 3 may have also allowed a negative gain of inhabitants, but the question was phrased such that it implied a gain and not a loss. Furthermore, entering negative numbers was not supported by our program.

Subjects received monetary payments in Swiss Francs (CHF) for each good estimate, taking the distance between the estimate and the true value into account. Three different intervals for monetary payments were used: 0% to 10% deviation (1.40 CHF), 11% to 20% deviation (0.70 CHF), and 21% to 40% deviation (0.35 CHF). Estimates that were more than 40% away from the true value were not financially rewarded. Rewards were communicated in experimental points and paid in CHF without requiring a signature after the experiment.

Our data (Dataset S1) comprise 12 groups in which 12 subjects responded five times to six knowledge questions in separate cubicles. Two questions were posed in the control, two in the aggregated, and two in the full information treatment. Thus, for each treatment, we had 24 groups: four groups for each of the six questions. The order of the questions and treatments was randomized among the experimental sessions.

Ten values were removed from the statistical analysis of the data set. Five of them were dramatic outliers from the same person in the same run. They were 1,000 times larger than the second largest estimate; thus, the subject seemed to have confused meters and kilometers. Four estimates were detected as "fun moves." The "fun" under the aggregate information condition was to make an incredibly large estimate to test how much the mean of the group would increase. Analogously, under the full information condition, the fun was to make incredibly high guesses to produce steep lines. The latter fun moves did not affect the rest of the information because there was always also a list of estimates from the previous round. One zero estimate was removed only when logarithmic data were used for computational reasons.

Measures. For a set of estimates $x_1, \ldots, x_n$ and the true value, we took the following measures to quantify the impact of social influence on the wisdom of crowds effect: The mean is denoted by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The collective error is the squared deviation of the average from the truth $(\text{truth} - \bar{x})^2$, the group diversity is the variance of estimates (average of squared deviations from the average) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The diversity prediction theorem (see ref. 9, p. 211, and ref. 35) states that collective error plus group diversity equals the average individual error, which is the average of the squared deviations from the truth $\left(\frac{1}{n}\sum_{i=1}^{n}(\text{truth} - x_i)^2\right)$. The proof is elementary. The collective error is also called "population bias" (6). A low collective error combined with a high group diversity implies that the wisdom of crowd effect works well, because asking many instead of one drastically improves accuracy. Both measures are used in Fig. 1 *D* and *E*. Notice that $x_i = \log(\tilde{x}_i/\text{truth})$ are logarithms of the raw data point $\tilde{x}_i$ normalized by the corresponding true values. The arithmetic mean of these logarithms corresponds to the geometric mean. Logarithms transform the results from arithmetic mean to geometric means, where a wisdom of crowd effect exists. Normalization of raw estimates by the truth makes different questions comparable.

For the definition of the wisdom-of-crowd indicator (Fig. 2) let $\hat{x}_1, \ldots, \hat{x}_n$ be the sorted estimates. Then, the wisdom of crowd indicator is $\max\{i | \hat{x}_i \leq \text{truth} \leq \hat{x}_{n-i+1}\}$. The wisdom-of-crowd indicator achieves its maximum at [*n* / 2] when the truth lies between the most central estimates (or at the most central estimate). Its minimum of zero is achieved when the truth lies below the minimal or above the maximal estimate. Notice, that a high wisdom-of-crowd indicator implies that the truth is close to the median. Thus, it implicitly defines the median as the appropriate measure of aggregation. In our empirical case this is not in conflict with the choice of the geometric mean as can be seen by the similarity of the geometric mean and the median in Table 1. A theoretical reason is that the geometric mean and the median coincide for a log-normal distribution.

ACKNOWLEDGMENTS. This paper benefited from comments made by Stefan Herzog, Michael Mäs, Ryan O. Murphy, two anonymous reviewers, and research assistance by Hanna Thorn and Silvana Jud. This study was supported by funding from Eidgenössische Technische Hochschule Zürich. Preliminary versions of the experiment were designed and conducted by J.L. in 2006 (36) and 2005.

1. Galton F (1907) Vox populi. *Nature* 75:7.
2. Lorge I, Fox D, Davitz J, Brenner M (1958) A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychol Bull* 55: 337–372.
3. Hommes C, Sonnenmans J, Tuinstra J, van de Velden H (2005) Coordination of expectations in asset pricing experiments. *Rev Financial Stud* 18:955–980.
4. Yaniv I, Milyavsky M (2007) Using advice from multiple sources to revise and improve judgments. *Organization Behav Hum Decis Process* 103:104–120.
5. Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Doubleday Books, New York).
6. Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psychol Sci* 19:645–647.
7. Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychol Sci* 20:231–237.
8. Rauhut H, Lorenz J (2010) The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *J Math Psychol* 55:191–197.
9. Page SE (2007) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton Univ Press, Princeton, NJ).
10. Mason WA, Conrey FR, Smith ER (2007) Situating social influence processes: dynamic, multidirectional flows of influence within social networks. *Pers Soc Psychol Rev* 11: 279–300.
11. Banerjee AV (1992) A simple model of herd behavior. *Q J Econ* 107:797–817.
12. Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *J Polit Econ* 100:992–1026.
13. Mannes AE (2009) Are we wise about the wisdom of crowds? the use of group judgments in belief revision. *Manage Sci* 55:1267–1279.
14. Allport GW (1924) The study of the undivided personality. *J Abnorm Psychol Soc Psychol* 19:132–141.
15. Asch SE (1955) Opinions and social pressure. *Sci Am* 193:31–35.
16. O'Gorman HJ (1986) The discovery of pluralistic ignorance: An ironic lesson. *J Hist Behav Sci* 22:333–347.
17. Miller DT, McFarland C (1991) When social comparison goes awry: The case of pluralistic ignorance. *Social Comparison: Contemporary Theory and Research*, eds Suls J, Wills T (Lawrence Erlbaum, Hillsdale, NJ), pp 287–313.
18. Goldstone RL, Gureckis TM (2009) Collective behavior. *Topics Cogn Sci* 1:412–438.
19. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–856.
20. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131.
21. Plous S (1993) *The Psychology of Judgment and Decision Making* (McGraw-Hill, New York).
22. Dawes RM, Mulford M (1996) The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organization Behav Hum Decis Process* 65:201–211.
23. Lazer D, Friedman A (2007) The network structure of exploration and exploitation. *Admin Sci Q* 52:667–694.
24. Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* 10:171–178.
25. Hooker RH (1907) Mean or median. *Nature* 75:487.
26. Galton F (1907) The ballot-box. *Nature* 75:509–510.
27. Genest C, Zidek J (1986) V Combining probability distributions: A critique and an annotated bibliography. *Stat Sci* 1:114–135.
28. Dawid AP, et al. (1995) Coherent combination of experts' opinions. *TEST* 4:263–313.
29. Chen KY, Fine LR, Huberman BA (2004) Eliminating public knowledge biases in information-aggregation mechanisms. *Manage Sci* 50:983–994.
30. Golub B, Jackson MO (2010) Naïve learning in social networks: Convergence, influence, and the wisdom of crowds. *Am Econ J Microecon* 2:112–149.
31. Dehaene S, Izard V, Spelke E, Pica P (2008) Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science* 320:1217–1220.
32. Limpert E, Stahel WA, Abby M (2001) Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51:341–352.
33. Larrick RP, Soll JB (2006) Intuitions about combining opinions: misappreciation of the averaging principle. *Manage Sci* 52:111–127.
34. Sommerfeld RD, Krambeck HJ, Milinski M (2008) Multiple gossip statements and their effect on reputation and trustworthiness. *Proc Biol Sci* 275:2529–2536.
35. Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning. *Adv Neural Inform Process Syst* 7:231–238.
36. Lorenz J (2007) Repeated averaging and bounded confidence – modeling, analysis and simulation of continuous opinion dynamics. PhD thesis (Universität Bremen, Bremen, Germany).