Decision
2014, Vol. 1, No. 2, 79–101

2325-9965/14/$12.00 DOI: 10.1037/dec0000004

When Is a Crowd Wise?

Clintin P. Davis-Stober
University of Missouri

David V. Budescu
Fordham University

Jason Dana
Yale University

Stephen B. Broomell
Carnegie Mellon University

Numerous studies and anecdotes demonstrate the “wisdom of the crowd,” the surprising accuracy of a group’s aggregated judgments. Less is known, however, about the generality of crowd wisdom. For example, are crowds wise even if their members have systematic judgmental biases, or can influence each other before members render their judgments? If so, are there situations in which we can expect a crowd to be less accurate than skilled individuals? We provide a precise but general definition of crowd wisdom: *A crowd is wise if a linear aggregate, for example a mean, of its members’ judgments is closer to the target value than a randomly, but not necessarily uniformly, sampled member of the crowd*. Building on this definition, we develop a theoretical framework for examining, a priori, when and to what degree a crowd will be wise. We systematically investigate the boundary conditions for crowd wisdom within this framework and determine conditions under which the accuracy advantage for crowds is maximized. Our results demonstrate that crowd wisdom is highly robust: Even if judgments are biased and correlated, one would need to nearly deterministically select only a highly skilled judge before an individual’s judgment could be expected to be more accurate than a simple averaging of the crowd. Our results also provide an accuracy rationale behind the need for diversity of judgments among group members. Contrary to folk explanations of crowd wisdom which hold that judgments should ideally be independent so that errors cancel out, we find that crowd wisdom is maximized when judgments systematically differ as much as possible. We reanalyze data from 2 published studies that confirm our theoretical results.

Keywords: wisdom of the crowd, judgment, forecasting, aggregation, prediction

Supplemental materials: http://dx.doi.org/10.1037/dec0000004.supp

Galton (1907) provided perhaps the first documentation of the “Wisdom of the Crowd” (Surowiecki, 2004) or “swarm intelligence” (Krause, Ruxton, & Krause, 2009) effect when he analyzed 787 individuals’ guesses of the weight of a slaughtered and dressed ox. The individuals were entered in a contest at a livestock show for which they were charged a small fee, thus motivating them to guess well. Because they competed for prizes, their discussions of the guesses were likely limited, resulting in a series of individual judgments uninfluenced by the guesses of others. Some competitors were highly skilled in this judgment, such as butchers and farmers, while others were novices. The guesses, which Galton called “unbiased by passion and oratory,” were nearly symmetric about the correct answer of 1,198 pounds, and the wisdom of the crowd, as captured by the median guess of 1,207 pounds, was accurate within 0.8%.

This article was published Online First February 3, 2014.

Clintin P. Davis-Stober, Department of Psychological Sciences, University of Missouri; David V. Budescu, Department of Psychology, Fordham University; Jason Dana, School of Management, Yale University; Stephen B. Broomell, Department of Social and Decision Sciences, Carnegie Mellon University.

This work was supported, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20059.

Correspondence concerning this article should be addressed to Clintin P. Davis-Stober, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211. E-mail: stoberc@missouri.edu

It has since been well established that aggregating judgments or predictions across individuals can be surprisingly accurate in a variety of domains, including prediction markets, political polls, game shows, and forecasting (see Surowiecki, 2004). Under Galton's conditions of individuals having largely unbiased and independent judgments, the aggregated judgment of a group of individuals is uncontroversially better, on average, than the individual judgments themselves (e.g., Armstrong, 2001; Clemen, 1989; Galton, 1907; Surowiecki, 2004; Winkler, 1971). The boundary conditions of crowd wisdom, however, are not as well understood. For example, when group members are allowed access to other members' predictions, as opposed to making them independently, their predictions become more positively correlated, and the crowd's performance can diminish (Lorenz, Rauhut, Schweitzer, & Helbing, 2011). In the context of handicapping sports results, individuals have been found to make systematically biased predictions, so that their aggregated judgments may not be wise (Simmons, Nelson, Galak, & Frederick, 2011). How robust is crowd wisdom to factors such as nonindependence and bias of crowd members' judgments? If the conditions for crowd wisdom are less than ideal, is it better to aggregate judgments or, for instance, rely on a skilled individual judge? Would it be better to add a highly skilled crowd member or a less skilled one who makes systematically different predictions than other members, increasing diversity?

We provide a simple, precise definition of the wisdom-of-the-crowd effect and a systematic way to examine its boundary conditions. We define a crowd as wise if a linear aggregate of its members' judgments of a criterion value has less expected squared error than the judgments of an individual sampled randomly, but not necessarily uniformly, from the crowd. Previous definitions of the wisdom of the crowd effect have largely focused on comparing the crowd's accuracy with that of the average individual member (Larrick, Mannes, & Soll, 2012). Our definition generalizes prior approaches in a couple of ways. First, we consider crowds created by any linear aggregate, not just simple averaging. Second, our definition allows the comparison of the crowd to an individual selected according to a distribution that could reflect past individual performance; for example, their skill, or other attributes. On the basis of our definition, we develop a framework for analyzing crowd wisdom that includes various aggregation and sampling rules. These rules include both weighting the aggregate and sampling the individual according to skill, where skill is operationalized as predictive validity; that is, the correlation between a judge's prediction and the criterion. Although the amount of the crowd's wisdom—the expected difference between individual error and crowd error—is nonlinear in the amount of bias and nonindependence of the judgments, our results yield simple and general rules specifying when a simple average will be wise. While a simple average of the crowd is not always wise if individuals are not sampled uniformly at random, we show that there always exists some a priori aggregation rule that makes the crowd wise.

Our results suggest that crowd wisdom is robust to different choices of aggregation and sampling rules. That is, how one aggregates the judgments or chooses an individual judge rarely affects the qualitative conclusion that even a crowd that is a simple average of judges is wiser than the individual. By identifying conditions for crowd wisdom, our results also provide guidance for constructing an optimally wise group—a group whose accuracy most exceeds that of its individual members—with two surprising conclusions emerging. First, a crowd becomes wisest when it is maximally informative, which entails that its members' judgments are as *negatively* correlated with each other as possible, as opposed to being independent. Thus, the best judge to add to a crowd is one that is maximally different from others. One intuitive analogy of this result is to think of the group as a financial portfolio: Sometimes it is better to diversify performance by "hedging" and including an asset that performs well when other assets perform poorly. This result provides mathematical support for the idea that crowds with more diversity are wiser (Hong & Page, 2004). Furthermore, our theoretical framework provides a mechanism for determining when it would be better for the overall group prediction to add a group member who, perhaps, is less skilled than the alternative members, but provides diverse predictions. In other words, our framework provides a quantification of the accuracy–diversity trade-off.

A second surprising conclusion is that while the absolute accuracy of the crowd depends on the direction and magnitude of members' bias, it is almost always preferable to use a weighted aggregate of judgments rather than select the single best group member, even if the crowd members are biased. Unless the best group member can be selected deterministically, as in certain intellective tasks (Laughlin, 1996), the decrease in variance of predictions caused by aggregating judgments will offset the bias, a manifestation of the well-known bias/variance trade-off (Gigone & Hastie, 1997).

We define accuracy as the average squared error of prediction (whether a group or individual). This is a common "gold standard" accuracy metric within the field of statistics (Lehmann & Casella, 1998). This accuracy metric allows us to derive distribution-free results on crowd wisdom. In other words, we make no assumptions regarding the underlying distributional form of the individual or group's predictions, such as normality, nor do we impose any constraints on the distribution's shape such as symmetry or unimodality. Alternative accuracy definitions (e.g., average absolute error) can change the conclusions of our model, though our approach is one that could, in theory, be extended to any accuracy metric.

We present an application of our framework to experimental studies by reanalyzing the data collected, analyzed and published by Vul and Pashler (2008) and Simmons et al. (2011). Our analysis finds a "wisdom of the crowd" effect when applied to the group of individuals from Vul and Pashler (2008), extending the original analysis which examined the accuracy of pooled repeated judgments *within* individuals. Our reanalysis of the Simmons et al. (2011) data supports the overall treatment effect of increasing individual bias by manipulating the sports betting information available to them. In contrast to the original findings reported by Simmons et al. (2011), our reanalysis, guided by our new formulation, finds an overall improvement of the crowd's predictions relative to individuals across all treatments in the study. In other words, while the members are individually biased and the crowd not particularly accurate, the crowd is still wise relative to the individual.

In the next section, we present the general definition of crowd wisdom and our basic sampling assumptions. We then derive a family of inequalities for evaluating the wisdom of the crowd effect. We then analyze several special cases, including comparing an equally weighted linear aggregate of the judges to probabilistically selecting an individual judge according to his or her skill. We then apply our framework to a reanalysis of two data sets. We conclude with a discussion and present future directions for this work.

The General Model

The Crowd Prediction

Consider a set of N-many decision makers (DMs), where each DM makes a judgment about the unknown value of a criterion. We model the criterion being predicted (or estimated) by the group members as a random variable with finite mean and variance. In this way, we conceptualize our framework as applying to random criteria, as in prediction, as well as to the special cases of estimating a single fixed quantity (which we accommodate by setting the variance of the criterion to 0). We take this criterion value to be a random variable, Y, with mean μ_y and variance σ_y^2.

Similarly, we assume that each DM's judgment is a random variable. This assumption represents the variability of a DM who gives variable responses to the same task. With this assumption, we can model how a DM's predictions correlate with the criterion as well as other DMs in the crowd. Let the prediction distribution of the i^{th} DM be the random variable X_i with mean μ_{xi} and variance σ_{xi}^2.

A *crowd prediction*, denoted C, is defined as the random variable formed by linearly combining the DMs according to predetermined weights w_i, $C = \sum_{i=1}^{N} w_i X_i$, with the restriction that all w_i are non-negative and, to ensure uniqueness, $\sum_{i=1}^{N} w_i = 1$. The weights, w_i, are not random variables, but rather fixed choices of how to combine crowd member judgments.

At this point, note that we place no a priori restrictions on the μ_{xi} and σ_{xi}^2 values. This allows for the possibility that DMs are *biased*, meaning that their average judgment would not equal the average criterion value, $E[X_i] = \mu_{xi} \neq \mu_y$. Also note that we allow DMs to have different prediction variances, σ_{xi}^2, and arbitrary covariances with other DMs where $\sigma_{xi,xj}$ denotes the covariance of X_i and X_j. In other words, the judgments of the crowd members may be correlated with each other. In this way, we can model the effects of crowd members influencing each others' judgments. This approach builds upon the seminal works of Hogarth (1978) and Winkler (1981), and our analyses generalize those of Einhorn, Hogarth, and Klempner (1977).

Finally, note that we place no a priori restrictions on the possible ranges of the covariance between X_j and Y, denoted $\boldsymbol{\sigma}_{xi,y}$, other than the usual positive semidefinite restrictions on covariance matrices. In other words, some crowd members may have more *skill* than others in that their judgments are better related to the criterion.

To fix these ideas, our framework could be used to evaluate the following types of tasks:

1. A group of N-many financial analysts that predict the weekly changes (or absolute changes) in the value of a market index (such as the DOWJI), or the exchange rate of the U.S. dollar and the Euro.

2. A group of N-many sports prognosticators who predict the number of points scored every week in all NFL games, or the number of goals scored in the Bundesliga.

3. A group of N-many weather forecasters who predict the total amount of monthly rain, or the average monthly temperature in a given location.

4. A group of N-many economists predicting the probability that the unemployment rate next month will be below 8%.

In all of these cases, we have a random target variable (criterion) and repeated random predictions from multiple judges. The individual forecasts and observed realizations of the corresponding random variables allow straightforward estimation of all the parameters (means, variances and covariances) that play a role in our model.

We clarify that our definition and analytic results are defined over a single, abstract prediction task. Often, one is interested in the wisdom of the same crowd across multiple, distinct prediction tasks. In Section 5, we demonstrate how our theory can be extended to such cases by adding some additional assumptions on crowd behavior for our reanalysis of the Vul and Pashler (2008) data set. This allows for an application of the theory to a wide range of empirical data sets.

Our model and results are limited to so-called "statisticized" groups, where the crowd is merely a mechanical linear aggregation of individual judgments, as opposed to, for example, freely interacting deliberative groups like juries or structured group interactions (e.g., Delphi method; Linstone & Turoff, 1975). While this focus is perhaps somewhat limited, it is consistent with much of the literature on crowd wisdom (although see Merkle & Steyvers, 2011, for a Bayesian aggregation model using nonlinear weights). Our definition of a "crowd" prediction as a linear combination of group member predictions contains the simple group average as a special case. Our approach can be seen as a generalization of several previous approaches, such as comparing the group average with an individual selected uniformly random (Einhorn et al., 1977; Wallsten & Diederich, 2001). We extend these approaches by considering other special cases, such as the one where the probability of selecting an individual is proportional to that individual's expected performance (measured by the correlation with the criterion variable).

Prediction of an Individual Selected Randomly

We consider whether the crowd's judgment is expected to be better than an individual crowd member's. Let P be the random variable formed by selecting a single member of the crowd probabilistically, and let p_i denote the probability of selecting the i^{th} crowd member, with $p_i \geq 0, \forall i \in \{1, 2, \ldots, N\}$ and $\sum_{i=1}^{N} p_i = 1$. As a special case, if all p_i values are equal, that is, $p_i = \dfrac{1}{N}, \forall i \in \{1, 2, \ldots, N\}$, then P reduces to selecting any individual DM with equal probability. At the other extreme, if $p_k = 1$ with $p_i = 0, \forall i \in \{1, 2, \ldots, N\}, i \neq k$, then the k^{th} DM is selected with probability one, for example, the highest performing group member is known. In a later example we consider the case where p_i is proportional to the i^{th} DM's correlation with Y.

A Wisdom of the Crowd Criterion

We consider the expected squared loss between each prediction distribution and the criterion distribution Y throughout. We compare the values $E[(C - Y)^2]$ and $E[(P - Y)^2]$ to one another, where "$E[\cdot]$" is the expectation operator. In other words, the prediction model that comes closest, on average, to Y is considered to be more accurate. This accuracy criterion, expected squared-error, is only appropriate for tasks in which "close-ness" of a prediction or judgment can be evaluated on a continuous scale (see Lee, Steyvers, de Young, and Miller, 2012; Yi, Steyvers, Lee, & Dry, 2012, for recent approaches to modeling crowd wisdom for combinatorial and ranking tasks, which would not meet our modeling assumptions).

We define a *wisdom of the crowd effect* to hold if, and only if,

$$E[(C - Y)^2] \leq E[(P - Y)^2], \tag{1}$$

for some crowd aggregate weights, w_i, $i \in \{1, 2, \ldots, N\}$, and selection distribution probability weights p_i, $i \in \{1, 2, \ldots, N\}$. Note that the right-hand side of Inequality (1) is the expected accuracy of selecting an individual according to an arbitrary, prespecified probability distribution, in contrast to previous formulations such as evaluating the arithmetic mean accuracy of individual predictions (Larrick et al., 2012).

Let μ_X be the $N \times 1$ vector of the DMs' mean predictions. Let Σ_{xx} be the covariance matrix of the X_i, $i \in \{1, 2, \ldots, N\}$, random variables. Let $\boldsymbol{\sigma}_{xy}$ denote the $N \times 1$ vector of covariances of Y with each X_i, $i \in \{1, 2, \ldots, N\}$. It is straightforward to show that $E[(C - Y)^2]$ is equal to the following:

$$E[(C - Y)^2] = \left(\mu_X'\boldsymbol{w} - \mu_y\right)^2 + \boldsymbol{w}'\Sigma_{XX}\boldsymbol{w} - 2\boldsymbol{w}'\sigma_{\boldsymbol{xy}} + \sigma_y^2,$$

where $\boldsymbol{w}$ is the $N \times 1$ vector of weights, w_i, $i \in \{1, 2, \ldots, N\}$, defining C.

Next, we consider the random variable $(P - Y)^2$. An application of the iterated expectation theorem (e.g., Bickel & Doksum, 2001) yields:

$$E[(P - Y)^2] = \sum_{i=1}^{N} p_i\left[(\mu_{xi} - \mu_y)^2 + \sigma_{xi}^2 - 2\sigma_{y,xi} + \sigma_y^2\right].$$

Proposition 1. (Wisdom of the Crowd Effect). The aggregate crowd prediction distribution, C, defined by $\boldsymbol{w}$ has lower expected loss than an individual judgment selected according to the probability measure, p_i, $i \in \{1, 2, \ldots, N\}$, if, and only if, the following inequality holds:

$$(\mu_X'\boldsymbol{w} - \mu_y)^2 + \boldsymbol{w}'\Sigma_{XX}\boldsymbol{w} - 2\boldsymbol{w}'\sigma_{\boldsymbol{xy}} + \sigma_y^2 \leq \sum_{i=1}^{N} p_i\left[(\mu_{xi} - \mu_y)^2 + \sigma_{xi}^2 - 2\sigma_{y,xi} + \sigma_y^2\right]. \tag{2}$$

It is possible to rearrange Inequality (2) in a way that separates clearly the various factors that drive the effect. By rearranging terms, we can simplify this expression as follows,

$$\sum_{\substack{i,j=1\\ i\neq j}}^{N} w_i w_j(\sigma_{xi,xj} + \mu_{xi}\mu_{xj}) \leq 2\sum_{i=1}^{N}(w_i - p_i)(\mu_y\mu_{xi} + \sigma_{xi,y}) - \sum_{i=1}^{N}\left(w_i^2 - p_i\right)\left(\mu_{xi}^2 + \sigma_{xi}^2\right). \tag{3}$$

If we additionally assume that $\mu_y = 0$ we obtain:

$$\sum_{\substack{i,j=1\\ i\neq j}}^{N} w_i w_j(\sigma_{xi,xj} + \mu_{xi}\mu_{xj}) \leq 2\sum_{i=1}^{N}(w_i - p_i)\sigma_{xi,y} - \sum_{i=1}^{N}\left(w_i^2 - p_i\right)MSE_{xi}.$$

The right-hand side of this inequality focuses on the N individuals in the crowd. In particular, this expression highlights the effect of the difference between the weights assigned to individuals in the crowd (w_i) and the probabilities of selecting these individuals from the crowd (p_i) on the individual mean squared errors of the individual judges and the individual judges' covariances with the criterion. This expression is maximized when individual judges with high covariances with the criterion ($\sigma_{y,xi}$) and low individual mean squared errors are overweighted in the crowd, relative to their probability of selection. On the other hand, the left-hand side of the inequality is independent of the criterion, Y, and reflects only the interrelation between the various judges in the crowd and their relative weights. It is minimized when there is an inverse relationship between $E(x_i x_j) = \boldsymbol{\sigma}_{xi,xj} + \mu_{xi}\mu_{xj})$ and $w_i w_j$, that is, when pairs of judges with high (low) $E(x_i x_j)$ are assigned relatively low (high) weights.

The above proposition provides an explicit, testable condition to determine whether a given crowd is wise. In the following special cases, we demonstrate how this result can be used to evaluate the relative trade-offs of group member interdependence versus bias. First, we prove a basic result within our framework.

Result 1. Consider the case when $w_i = p_i, \forall i \in \{1, 2, \ldots, N\}$, that is, the aggregation weights providing the crowd prediction are identical to the selection weights used to determine the individual DM prediction distribution. Then a wisdom of the crowd effect always holds.

Proof. See Appendix

Result 1 extends the finding that the crowd member average is more accurate than the average crowd member to any situation in which the aggregation weights are identical to the probability weights used to select the individual. As in previous, related arguments (Dawes, 1970; Hogarth, 1978), Result 1 is a straightforward application of Jensen's inequality.

While Result 1 guarantees the existence of an aggregation method that makes the crowd wise, it is interesting to consider particular aggregation rules, such as an unweighted aggregate, against other methods of selecting individuals. For example, a practical problem that one might face is to choose between the judgment of an available expert and a crowd. Without exact knowledge of the relevant crowd parameters, it is difficult to decide on aggregation rules other than a simple average. We can still run through scenarios involving a simple crowd average, however, to see how likely it would be to find an expert that is more accurate. Furthermore, we may wonder about the extent of crowd wisdom and factors that lead to maximizing the crowd's predictive advantage over the individual. We next provide some comparative analyses using Inequality (1) to demonstrate factors that maximize crowd wisdom. In the section following, we examine special cases in which the aggregation weights do not match the selection weights.

Unweighted Average Versus Selecting an Individual at Random With Equal Probability

Consider the simple case where the crowd, C, is defined by the unweighted (simple) group average, $w_i = \frac{1}{N}, i \in \{1, 2, \ldots, N\}$, and the competitor model P is defined by the uniform distribution $p_i = \frac{1}{N}, i \in \{1, 2, \ldots, N\}$; that is, the competitor individual is selected uniformly at random. This models the case where one has no prior information to suggest or reason to believe that any member of the group is any better at the prediction task than any other. Inequality (2) can be rearranged and written:

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{xi,xj} - \frac{1}{N}\sum_{i=1}^{N}\sigma_{xi}^2 \le \frac{1}{N}\sum_{i=1}^{N}(\mu_{xi} - \mu_y)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\mu_{xi} - \mu_y\right)^2. \quad (4)$$

Recall that this inequality is simply an algebraic rearrangement of the inequality, $E[C - Y)^2] \le E[(P - Y)^2]$, and thus, the magnitude to which (4) deviates from equality is precisely the expected difference between the random variables $(C - Y)^2$ and $(P - Y)^2$. The greater the deviation from equality in (4), the more pronounced the wisdom of the crowd effect.

What does the composition of the "crowd" look like when the inequality $0 \le E[(P - Y)^2] - E[C - Y)^2]$ is maximized? When it is minimized? First, consider the left-hand side of inequality (4). Because the covariance matrix of the judges is positive semidefinite, this side of the inequality is necessarily nonpositive. To simplify matters, assume that all predictions are standardized such that the covariance between judges X_1 and X_2 can be interpreted as correlations, r_{xij}. The left-hand side of inequality (4) is maximized when all the judges are perfectly correlated with each other, that is, $\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} r_{xij} - \frac{1}{N}\sum_{i=1}^{N} r_{xi}^2\right) = \frac{N^2}{N^2} - \frac{N}{N} = 0$. As the judges become less correlated with one another, this value becomes smaller and the wisdom of the crowd effect becomes more pronounced. This result is intuitive because if all the judges provide the same (or almost the same) predictions, there would be little gained by aggregation. Note that when crowd members' judgments are highly correlated, adding in a new member whose judgments are nonredundant improves crowd wisdom, particularly for a small crowd. As is evident from Inequality (4), this improvement can occur even if the new, nonredundant member has substantially lower skill than the existing members. The intuition behind this result is clearer when one considers other linear aggregation problems: If one has highly redundant predictor variables in a multiple regression, adding a predictor that gives new information will help the model even if the new predictor is poorly correlated with the outcome variable. In the context of multiple regression this effect is similar to "suppression" in that a new member can improve the crowd's estimate via his or her relationship with other judges and not the criterion per se (see Tzelgov & Henik, 1991).

It is worth noting that while the left-hand side is smaller for perfectly uncorrelated judges, it is not minimized in this case. This term is minimized when all judges are equally and maximally *negatively* correlated with one another. This result is distinct from the folk explanation of crowd wisdom which holds that independence among judges is necessary so that errors cancel out (variance reduction does factor into the right-hand side of this inequality, as shown below). This result is in line, however, with other mathematical models demonstrating how group diversity can improve overall group accuracy (Hong & Page, 2004). To clarify, this maximal negative correlation is subject to the usual positive semidefinite constraints on the interjudge correlation matrix, which, for large crowds, will be necessarily very small. As the number of judges goes to infinity, the maximal negative correlation of all judges approaches zero.

While the left-hand side of Inequality (4) describes the effects of intercorrelation between judges, the right-hand side describes the effects of judge bias, that is, the expected squared deviation between a judge's prediction and the criterion. This side of the inequality must necessarily be non-negative. This term, $\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi} - \mu_y)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\mu_{xi} - \mu_y\right)^2$, is minimized when the true means of all judge predictions are equal to the mean of the criterion, that is, when all judges are unbiased. In this case, there is little benefit to aggregating the judges from the standpoint of minimizing bias because all of them are unbiased. All aggregation can do in this situation is reduce the variance of the aggregate prediction (Wallsten & Diederich, 2001). Maximizing this term is far more interesting. The right-hand side of (4) becomes arbitrarily large as the average squared bias of the judges becomes large with the squared bias of the judge average remaining small or zero. In this case, all judges are (possibly greatly) biased in their individual predictions, yet the average of their predictions is very close to the true criterion. Put in the terminology of Larrick and Soll (2006), the individual judge predictions "bracket" the true criterion mean, falling, more or less, both above and below μ_y. Here, the individual predictions systematically fall either above or below μ_y, but when averaged, this individual bias is cancelled and the crowd prediction is wise.

Unweighted Average Versus Selecting an Individual According to Their Skill

In the previous section, we considered the simple case where one does not discriminate between the individual judges a priori. Prior work has demonstrated that for intellective tasks without demonstrable solutions, groups are often poor at identifying the highest performing member (Henry, 1995). While informative, this case may not be general. Often, we do have information on the prior performance of judges; that is, their skill at a particular prediction task, for example Cooke's method (Cooke, 1991). In this section, we compare different aggregation weights, w_i, to a randomly selected individual such that the probability of selecting the i^{th} DM is proportional to that judge's skill, defined as the correlation of his or her prediction with the criterion, Y. Let all judges' skill be non-negative, $r_{xiy} \geq 0, \forall i \in \{1, 2, \ldots, N\}$, and let p_i be defined as follows:

$$p_i = \frac{r_{xi,y}}{\sum_{i=1}^{N} r_{xi,y}}. \tag{5}$$

Clearly, the higher the correlation of an individual's predictions with the criterion, the more likely that individual will be selected. If all DM predictions are equally correlated with Y then this choice of P will reduce to selecting a DM uniformly at random, which has already been shown in Result 1 to be inferior, on average, to the case where C is the unweighted average. At the other extreme, if only a single DM's predictions correlate with Y then that DM will be chosen with probability one, and will, most likely, outperform the unweighted average of the crowd.

Let C be defined according to the simple unweighted average and let P be defined according to (5). Applying Proposition 1 and rearranging terms gives us the following result.

Corollary 1. Let $w_i = \frac{1}{N}, \forall i \in \{1, 2, \ldots, N\}$ and let p_i be defined as in Equation (5). Assume that all variables are standardized such that $\sigma_{xi,y} = r_{xi,y}$ and $\sigma_{xi,xj} = r_{xi,xj}, \forall i, j$. Then a wisdom of the crowd effect holds if, and only if,

$$MSE_{crowd} - \sum_{i=1}^{N} \frac{r_{xi,y}}{\sum_{i=1}^{N} r_{xi,y}} MSE_{x_i} \leq 2\left(Mean(r_{xiy}) - \sum_{i=1}^{N} \frac{r_{xi,y}^2}{\sum_{i=1}^{N} r_{xi,y}} \right), \quad (6)$$

where MSE_{x_i} is the mean squared error for the i^{th} DM prediction distribution and MSE_{crowd} is the mean squared error for the crowd prediction, C.

Mean squared error is equivalent to the sum of the prediction distribution's squared bias (with respect to μ_y) and its variance. Examining the right-hand side of Inequality (6), we see that the crowd prediction benefits when skill is evenly distributed among the DMs; in other words, when all DMs are "equally good." The left-hand side of Inequality (6) indicates that for the crowd to do well, the most highly skilled DMs (those with the highest correlations with the criterion) should also have the largest biases.

One DM Doesn't Follow the "Herd": Unbiased Case

Consider the case of a defector, "one DM who doesn't follow the herd" model assuming that C is defined by the unweighted group average and P is defined as in (5). In this case, we will assume that there are N-many judges with $N - 1$ DMs who positively correlate with the criterion, Y, and each other at the value ψ. This group of $N - 1$ DMs represents the "herd." The remaining DM is the "dissident" and correlates with the criterion Y at ϕ. To ensure the positive semidefiniteness of the intercorrelation matrix between judges and the criterion, we will also assume that the dissident correlates with the herd judges at ψ. For now we assume that all N-many DMs are unbiased in their predictions, that is, $\mu_{xi} = \mu_y, \forall i \in \{1, 2, \ldots, N\}$. We will consider biased cases in subsequent sections.

Clearly, when $\psi = 0$ and ϕ is large, we have a group of uncorrelated DMs and only the dissident DM has any skill at predicting the criterion variable.[1] Under this set of assumptions, the dissident is selected with probability 1 under P, and will be more accurate, in expectation, than the group aggregate C. At the other extreme, if $\psi = \phi$ we have a group of equally skilled DMs who are equally correlated with one another. Under this condition, P, as defined in (5), reduces to selecting one of the DMs with equal probability and will always do worse, on average, than the unweighted aggregate, C, by Result 1. To investigate this relationship, we consider three different levels of "skill" on the part of the dissident, $\phi = .95$ (high), $\phi = .70$ (medium) and $\phi = .40$ (low), and vary the ratio $\frac{\psi}{\phi}$ from 0 to 1. As an application of Corollary 1, let LHS be the left-hand side of Inequality (6), similarly, let RHS be the right-hand side of Inequality (6). Thus, the value $R = RHS/LHS$ is the ratio of individual expected loss to crowd expected loss. For ease of presentation, our results are in terms of R on a logarithmic scale, denoted $\log(R)$. $\text{Log}(R)$ provides a continuous measure of the wisdom of the crowd effect with positive (negative) numbers indicating the crowd is expected to be more (less) accurate than an individual chosen at random. When $\log(R) = 0$, the expected loss of the crowd is equal to the individual expected loss. Figure 1 plots $\log(R)$ as a function of the ratio

[1] As N increases without bound, ψ equals the minimal correlation between the dissident and herd DMs that guarantees that the resulting correlation matrix is positive semidefinite.

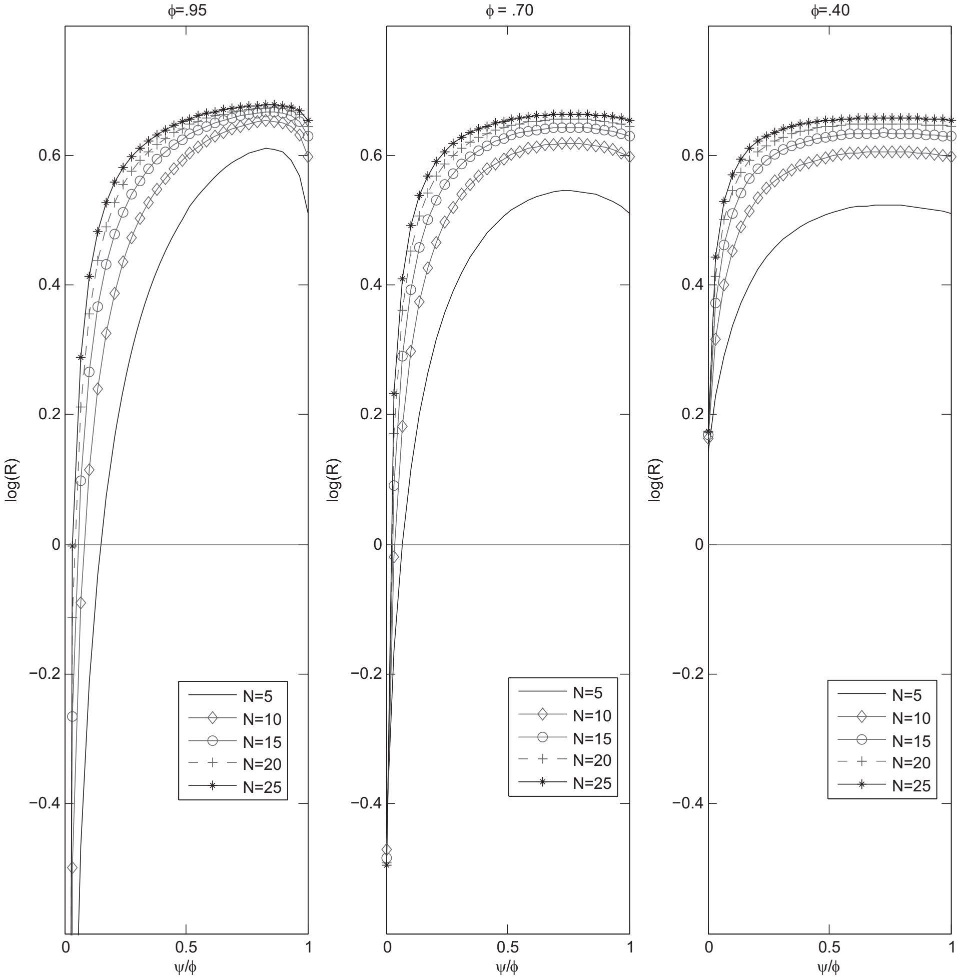

Figure 1. This figure plots log(*R*) as a function of the ratio $\frac{\psi}{\phi}$ for $\phi = .95, .70, .40$. For each value of ϕ, log(*R*) values are plotting for five samples sizes, $N = 5, 10, 15, 20, 25$. The left-hand graph corresponds to $\phi = .95$, the middle graph corresponds to $\phi = .70$, and the right-hand graph corresponds to $\phi = .40$. The line, $\log(R) = 0$, is plotted for reference. All points above this line represent cases where the crowd's performance is superior.

$\frac{\psi}{\phi}$ for group sizes N = 5,10,15,20,25 separately for the different skill levels of the dissident. The line denoting equal accuracy of the crowd and a randomly selected individual is plotted for reference.

Note that the simple unweighted average, *C*, performs quite favorably compared with an individual selected at random according to (5), even in the case where the dissident has a high correlation with the criterion. Selecting an individual according to (5) outperforms the crowd in this case only when the herd is very weakly correlated with the criterion, for example, the point at which *C* outperforms *P* for $N = 15$ under $\phi = .95$ occurs when $\psi = .052$. The size of the groups plays a large role in determining the point at which the wisdom of the crowd effect emerges. As expected, the larger the group size the more pronounced the wisdom of the crowd effect and the smaller the value of $\frac{\psi}{\phi}$

at which it emerges. There is also a strong effect of the skill level of the dissident. As ϕ decreases, the favorable range of the ratio $\frac{\psi}{\phi}$ under P becomes quite small and, eventually, vanishes ($\phi = .40$). In this case, even if one could select deterministically the best member of the group, their modest correlation with the criterion would not offset the reduction in sampling error by incorporating an unweighted average of the rest of the DMs. To summarize, unless one could nearly deterministically select the best member of the group, who must be *highly* skilled, a simple unweighted group average will, on average, prevail.

It is interesting that the performance of C versus P under these assumptions is highly nonlinear. Under the extreme case of $\frac{\psi}{\phi} = 1$, a wisdom of the crowd effect is guaranteed to occur by Result 1, yet this is not the condition that yields maximal log(R) values. Across all three conditions, the largest values of log(R) occur when the herd is modestly, but not maximally, correlated with the criterion. Under these values of $\frac{\psi}{\phi}$, the dissident has a reasonable chance of *not* being selected with the remaining group members having relatively smaller prediction correlations with the criterion. Yet, there is a large amount of information present in the group as a whole, as measured by small judge intercorrelation, so the prediction of C will likely perform extremely well.

One DM Doesn't Follow the "Herd": Biased Case

Let us return to the "one DM doesn't follow the herd" model analyzed above but allow the DMs in the herd to be biased in their predictions. Intuitively, the wisdom of the crowd effect, as defined in Proposition 1, should depend not just on the magnitude of the DM biases but also on their relative *configuration*. For example, a herd in which the DM biases are equally likely to be above/below μ_y will likely result in different log(R) values than a herd in which all DM biases are in the same direction.

We consider two cases. In the symmetric case the prediction biases of the herd DMs are no more likely to be above than below μ_y with the dissident DM as the sole member whose predictions are unbiased. Table 1 displays the μ_X values for this model under the symmetry condition for the five group sizes, $N = 5, 10, 15, 20, 25$. Recall that because μ_y is defined to be zero, it suffices to specify μ_X to model bias. As before, we assume that all prediction values are standardized so that values of μ_X can be interpreted as bias in units of standard deviations. As shown in Table 1, the number of DMs in each group with positive or negative biases are roughly equal and symmetric with respect to bias magnitude. As group size increases, the magnitude of the biases also increases, with a maximal bias of plus or minus two standard deviations.

Table 1
Bias Configurations for the Five Hypothetical Groups Under the Symmetry Condition

Number of decision makers (DMs)	Possible bias values								
	0	.5	−.5	1	−1	1.5	−1.5	2	−2
	Counts per group								
$N = 5$	1	1	1	1	1	0	0	0	0
$N = 10$	1	3	2	2	2	0	0	0	0
$N = 15$	1	3	3	3	3	1	1	0	0
$N = 20$	1	4	3	3	3	3	3	0	0
$N = 25$	1	3	3	3	3	3	3	3	3

Note. The columns indicate the possible bias values we consider, as well as the number of DMs in each group with that bias level. For example, the group with 5 DMs has one unbiased DM, with the remaining 4 DMs having bias levels of .5, −.5, 1, and −1.

Figure 2 displays log(R) values as a function of the ratio $\frac{\psi}{\phi}$ under the symmetry condition. All other assumptions are identical to the analysis in the previous section. The log(R) values are much larger in this condition than the completely unbiased case previously examined, likewise, the size of the group has a more pronounced effect on how "wise" a group is, with larger group values resulting in larger log(R) values. In other words, given that the dissident is the only unbiased DM in the group, selecting an individual probabilistically incurs a much higher penalty. However, from the perspective of the crowd prediction C, the bias penalty is averaged out, because the biases of the individual members are symmetric about μ_Y, similar to the bracketing effect of Larrick and Soll (2006). In this scenario, aggregation can only help to lower prediction variance, hence the more extreme wisdom of the crowd effect.

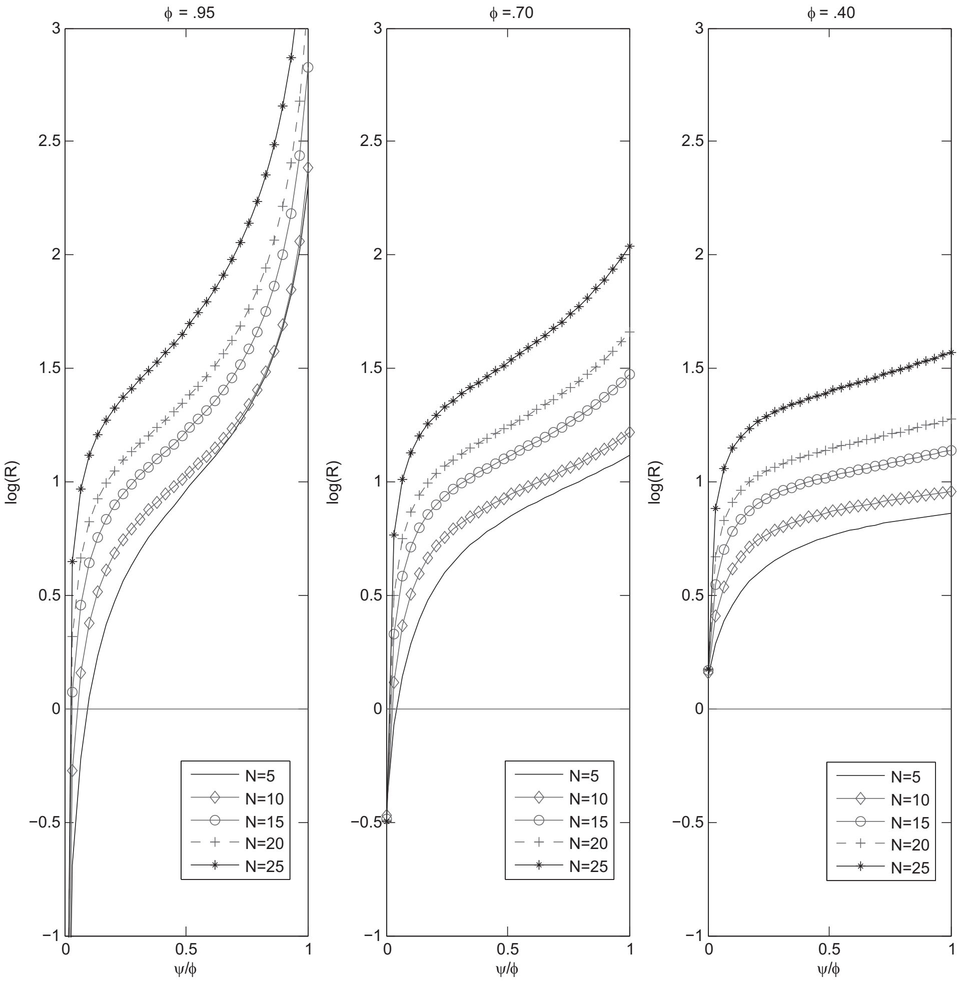

Figure 2. This figure plots log(R) as a function of the ratio $\frac{\psi}{\phi}$ for ϕ = .95, .70, .40. For each value of ϕ, log(R) values are plotting for five samples sizes, N = 5, 10, 15, 20, 25. The left-hand graph corresponds to ϕ = .95, the middle graph corresponds to ϕ = .70, and the right-hand graph corresponds to ϕ = .40. The line, log(R) = 0, is plotted for reference. All points above this line represent cases where the crowd's performance is superior.

Next, we examine another version of the model with identical assumptions except that the bias configuration of the DM predictions is asymmetric. For example, all judges systematically overestimate the probability of a rare event, such as the probability of a high intensity earthquake. We set the μ_X vectors equal to those defined in Table 1, with the exception that we consider the absolute values of all entries for all μ_X. For this model, all nondissident DMs are systematically positively biased in their predictions with respect to μ_y with bias values ranging from .5 to 2 standard deviations.

Figure 3 displays these log(R) values as a function of $\frac{\psi}{\phi}$ for $\phi = .95, .70, .40$. As expected, the wisdom of the crowd effect is less extreme than in the symmetry condition. Although these log(R) values are smaller than in the symmetry condition, the overall magnitudes of the log(R) values and their relationships to

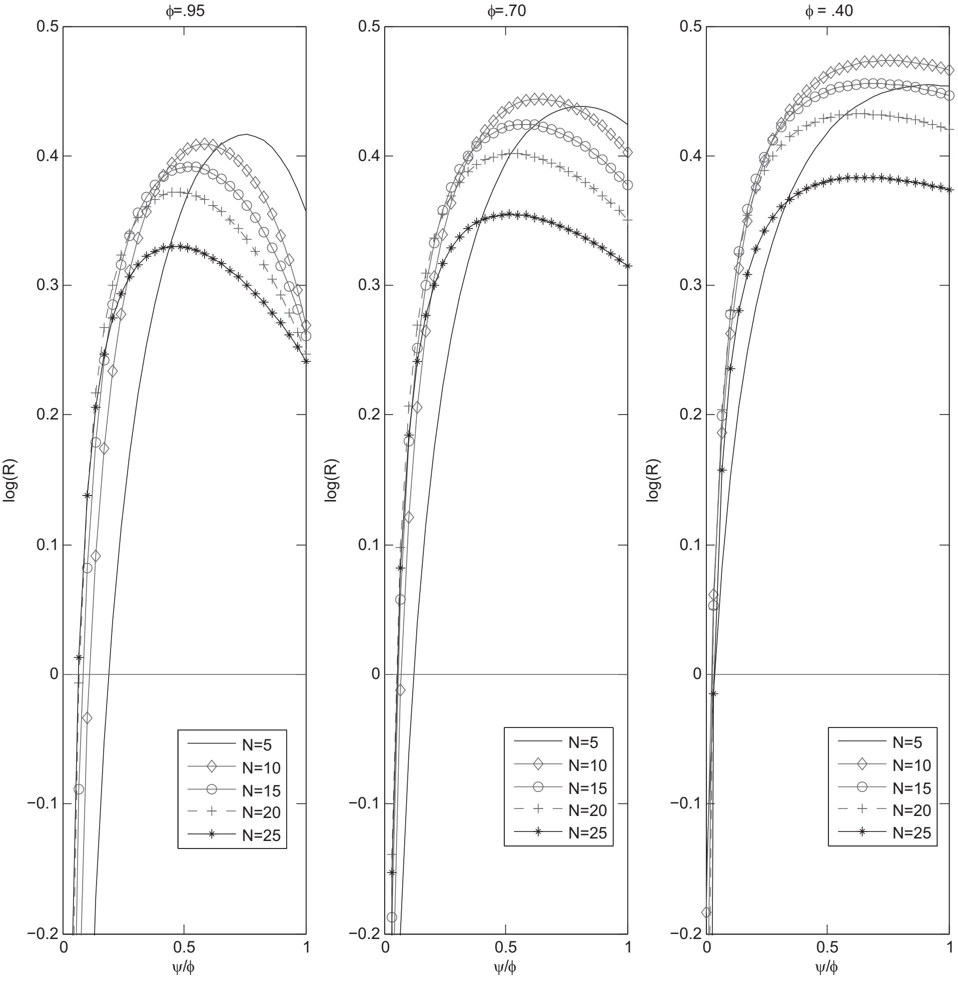

Figure 3. This figure plots log(R) as a function of the ratio $\frac{\psi}{\phi}$ for $\phi = .95, .70, .40$. For each value of ϕ, log(R) values are plotting for five samples sizes, $N = 5, 10, 15, 20, 25$. The left-hand graph corresponds to $\phi = .95$, the middle graph corresponds to $\phi = .70$, and the right-hand graph corresponds to $\phi = .40$. The line, log(R) $= 0$, is plotted for reference. All points above this line represent cases where the crowd's performance is superior.

the ratio $\frac{\psi}{\phi}$ are similar to that of the completely unbiased condition. This result speaks strongly to the general robustness of the unweighted average. Even in the face of highly and unidirectionally biased members, it is still often preferable to simply average the members as opposed to selecting a single, best-performing unbiased one.

Applications of the Model to Real Data

In this section, we reanalyze data from two papers that investigated the wisdom of crowds using two different types of tasks (trivia questions and sports betting) with different types of data (continuous estimates and dichotomous choices). The first analysis applies our framework to the data from Vul and Pashler (2008) to estimate the expected loss of a crowd versus an individual selected at random. This analysis illustrates how our theory could be extended to multiple prediction/estimation tasks. The second analysis applies our framework to the data from Simmons et al. (2011) to estimate the wisdom of a crowd whose members repeatedly predicted sports outcomes against a point handicap. We use our framework to highlight the impact of member bias induced by the four different experimental conditions and demonstrate how bias impacts the performance of a crowd.

Estimation of Expected Loss

Recall that Inequality (2), our criteria for a crowd to be wise, provides a breakdown of expected squared loss from a crowd versus a randomly chosen individual, and is as follows,

$$(\mu_X'\boldsymbol{w} - \mu_y)^2 + \boldsymbol{w}'\Sigma_{XX}\boldsymbol{w} - 2\boldsymbol{w}'\sigma_{\boldsymbol{xy}} + \sigma_y^2 \leq \sum_{i=1}^{N} p_i\left[(\mu_{xi} - \mu_y)^2 + \sigma_{xi}^2 - 2\sigma_{y,xi} + \sigma_y^2\right].$$

The left-hand side (LHS) of this inequality is a linear combination of (a) the crowd level bias, (b) covariances among crowd members, and (c) the crowd covariance with criterion, and (d) the variance of the criterion. The right-hand side (RHS) of Equation 2 is the linear combination of (a) the bias of each individual's estimates, (b) the variance of each individual's estimates, and (c) the covariance of each individual's estimates with the criterion, and (d) the variance of the citerion. For each dataset, we estimate each of the components of LHS and RHS. As in the previous section, we will evaluate crowd wisdom via log(R) where R = RHS/LHS, which we estimate for each dataset. We compare the estimate of log(R) to a previously established measure of crowd performance: the percentage of individuals that are less accurate than the crowd (Simmons et al., 2011). This percentage is determined by comparing the mean-squared error (*MSE*) of each individual with the *MSE* of the crowd. It is important to note that by moving from the theory to empirical data, we must estimate the parameters of interest and, subsequently, log(R). Hence, we also need to be concerned with sampling error with respect to parameter estimation. To impose as few assumptions as possible on the empirical data, we carried out a jackknife procedure (Miller, 1974) to estimate this variability. Alternative methods could also be used, for example, Bayesian estimation, with additional distributional assumptions on DM behavior.

Reanalysis of Vul and Pashler (2008)

The theory we have developed in previous sections was defined over a single, abstract prediction task. To illustrate how our theory could be applied to the more general case of multiple tasks, we consider experimental data from Vul & Pashler (2008). To be clear, we are not proposing a theory of crowd wisdom across multiple tasks per se, rather we are suggesting one possible method of extending our approach. Vul and Pashler ran a study with multiple tasks in the form of 8 trivia questions. Our analysis of this data set will require additional assumptions on the questions. Given the similarity of the questions, we considered these questions as a sample of questions drawn from a universe Y of questions that could have been selected. We introduce a new index, j, for estimating the parameters from this population of questions. Considering the questions as a random sample from Y allows this collection of questions to be treated as a random variable with mean, μ_y, and variance, $\hat{\sigma}_y^2$.

Hence, we are making inferences at the level of the population of possible questions. The biases and covariances of the DMs are defined at the level of the random variable Y. We do not assume that DMs have stationary biases with respect to each trivia question; rather, each DM has a bias with respect to the average answer from the universe of possible trivia questions, μ_y. All questions from Vul and Pashler were on a similar scale (responses from 0–100).

Vul and Pashler (2008) used data from $N = 428$ subjects who provided estimates (from 0 to 100) to $J = 8$ questions. Each subject provided two responses (immediately; delayed by three weeks); the current analysis uses only the immediate response data. Each of the subjects, $i = 1, \ldots 428$, produced judgments for each of the questions, $j = 1, \ldots 8$, denoted x_{ij}. The answers to the 8 questions are denoted as y_j. Vul and Pashler tested the wisdom of the crowd by comparing individual versus group mean squared error (*MSE*) across the 8 questions.

Inequality (2) requires an estimate of the mean and variance of the criterion Y that are computed as the sample mean and sample variance of the 8 true answers given by,

$$\hat{\mu}_y = \bar{y} = \frac{1}{8}\sum_{j=1}^{8} y_j = 32.64,$$

and

$$\hat{\sigma}_y^2 = \frac{1}{(8-1)}\sum_{j=1}^{8}(y_j - \bar{y})^2 = 556.11.$$

Next, we estimate the mean judgment from each individual and the covariance between the judgments produced by all pairs of individuals as,

$$\hat{\mu}_{x_i} = \bar{x}_i = \frac{1}{8}\sum_{j=1}^{8} x_{ij} \text{ for } i = 1, \ldots, N,$$

and

$$\hat{\Sigma}_{x_i x_k} = \frac{1}{8-1}\sum_{j=1}^{8}(x_{i,j} - \bar{x}_i)(x_{k,j} - \bar{x}_k).$$

Finally, the covariances of the judgments with criterion variable are computed in the same way for each individual using the 8 judgments and answers:

$$\hat{\sigma}_{x_i y} = \frac{1}{8-1}\sum_{j=1}^{8}(x_{i,j} - \bar{x}_i)(y_j - \bar{y}).$$

Results

We computed the estimates for the LHS and RHS of Inequality (2) and compared the measure log(R) with the commonly used post hoc method of computing the percent of individuals beat by the crowd in Table 2. We manipulate the definition of the crowds' judgment by (a) creating several subgroups of the original crowd and (b) applying different weighting criteria for the crowd prediction. The first row is a comparison of the

Table 2
Estimates for Expected Loss of Crowd Versus a Randomly Selected Individual

	Expected loss estimate			
Weights (w)	LHS (Crowd)	RHS (Individual)	log(R) (*SE*)	Proportion of individuals beat by the crowd
Equal weights				
100% of crowd	131.23	608.58	1.53 (0.07)	0.96
Most valid individual	112.18	608.58	1.69 (0.08)	0.99
Most valid 5%	60.54	608.58	2.31 (0.04)	1.00
Most valid 25%	52.95	608.58	2.44 (0.07)	1.00
Most valid 50%	57.36	608.58	2.36 (0.09)	1.00
Least valid 50%	299.25	608.58	0.71 (0.06)	0.78
Least valid 25%	429.05	608.58	0.35 (0.06)	0.62
Least valid 5%	722.55	608.58	−0.17 (0.04)	0.30
Least valid individual	1251.11	608.58	−0.72 (0.04)	0.08
Unequal weights				
Proportional to validity	87.12	608.58	1.94 (0.07)	1.00
Inversely proportional to validity	261.96	608.58	0.84 (0.06)	0.83

Note. Estimates of log(R) are accompanied by *SE*s produced from the jackknife procedure.

equally weighted crowd against individuals drawn randomly with equal probability. The next eight rows in Table 2 show the results of producing subsets of the original crowd by equally weighting the judgments produced by subsets of individuals based on their validity. For example, the *most valid individual* crowd is created by giving only the most valid crowd member a weight of 1, and putting zero weight on the remainder of the crowd. *The most valid 50%* crowd was created by weighting equally those ranked in the top 50% based on validity and giving a weight of zero to the bottom 50%. Each crowd in Table 2 is compared with the expected loss of randomly selecting an individual from the entire sample. The bottom two rows apply unequal weights to the entire crowd that are proportional to each individual's validity.

The columns of Table 2 provide estimates of expected loss (in the first two columns), the measure of crowd performance, log(R), in the third column, and the percent of individuals beat by the crowd in the last column. We also include the standard error of the estimate of log(R) computed using the Jackknife procedure.[2] The values of log(R) correspond nicely to the percent of individuals beat by the crowd (Pearson $r = 93$; Kendall $\tau = .94$). Table 3 presents a breakdown of the expected loss estimate based on (a) crowd bias, (b) crowd covariance, and (c) crowd covariance with criterion. This breakdown shows the source of changes in the estimates of the crowd expected loss. We can see that the marginal improvement of the *most valid individual* is explained by the very high crowd covariance term despite having lower bias and higher covariance with the criterion.

The values from this example in Table 2 show that the crowd's expected loss only exceeds the individual expected loss in very extreme cases (i.e., the crowd weighting includes the 5% least valid individuals). These results demonstrate the robustness of the wisdom of this particular crowd, even when applying weights to the crowd that are inversely related to validity.

Reanalysis of Simmons et al. (2011)

We now analyze a data set from a study which suggested that the crowd is not wise, and performs worse than a large majority of individuals. Simmons et al. (2011) hypothesized that systematic bias in individual's judgments can potentially cause the crowd to be unwise even when all conditions that typically foster wisdom of the crowds hold. To test this hypothesis, Simmons et al. designed a series of experiments which use a point spread betting market; previous research suggests that crowds in this context may not be wise (Kahneman & Frederick, 2002; Simmons & Nelson, 2006) because of individuals having a tendency to bet on favorites over underdogs despite the fact that point spreads attempt to produce even odds for underdogs and favorites (Levitt, 2004; Simmons & Nelson, 2006).

Simmons et al. (2011) ran an experiment where they accentuated the effect of this bias to choose the favorite by having subjects bet on point spreads that were systematically shifted (relative to Las Vegas point spreads) to make the underdog team have better odds of winning. The control condition required subjects to choose which team they believed would win against the point spread (labeled *choice condition*). The authors attempted to shift the amount of bias for choosing the favorite by three experimental manipulations. The *warned choice condition* warned each individual that the point spreads have been set incorrectly such that betting on the underdog team has the better odds of winning. In the *estimate condition* subjects did not bet against the point spread, but instead provided an estimate of the final score of the game. The estimate is then compared with the point spread to infer a choice against the point spread. This method is predicted to reduce bias by shifting the response mode away from the choice between the favorite and underdog (which is systematically biased toward choosing the favorite) to estimating the number of points in which they expect the favorite to win. Finally, subjects in the *choice/estimate condition* predicted the winner against the point spread and then provided an estimate of the final score. The choices for this condition are also inferred from the estimates. The data consist of $N = 178$ individuals (choice: $n = 43$; warned choice: $n = 39$; estimate: $n = 45$; choice/estimate: $n = 51$) betting on 226 games over the

[2] Given J observations, the jackknife procedure that we employed computes J-many estimates of log(R) after eliminating the j^{th} observation ($j = 1, \ldots J$). The J estimates are used to compute the standard error. See Miller (1974) for more details.

Table 3
Break Down of Crowd Expected Loss Estimate

Weights (w)	Bias	Crowd covariance	Covariance with criterion
Equal weights			
100% crowd	19.19	339.76	783.84
Most valid individual	0.58	1,041.84	1,486.35
Most valid 5%	0.63	878.50	1,374.70
Most valid 25%	8.44	678.01	1,189.61
Most valid 50%	8.70	562.95	1,070.41
Least valid 50%	33.79	206.62	497.28
Least valid 25%	40.08	144.04	311.18
Least valid 5%	39.20	87.67	−39.56
Least valid individual	168.68	181.70	−344.63
Unequal weights			
Proportional to validity	14.31	442.53	925.84
Inversely proportional to validity	29.54	209.83	533.52

course of 17 weeks (number of games per week varied).

Simmons et al. analysis. Details of the original analysis conducted by Simmons et al. are found in Tables 2–4 of Simmons et al. (2011). Their results show that in both choice conditions (regardless of warning), the crowd is biased and picks the favorite far too often (Simmons et al. Table 2), and as a result, the crowd predictions have fewer wins (Simmons et al. Table 3) and outperform a very small percentage of individuals (Simmons et al. Table 4). They report that the percent of individuals beat by the crowd are 7% for choice, 0% for warned choice, 57.8% for estimate, and 35.2% for choice/estimate.[3] These results suggest that the crowds in the choice and warned choice conditions are highly biased and as a result, *not wise*.

Expected loss analysis. We apply the expected loss metric to test crowd performance in each of the conditions outlined above. This methodology has the unique advantage of showing the contribution of bias to the performance of a crowd, a main objective of Simmons et al. (2011). We take a slightly different approach to the data analysis that fits the statistical assumptions of our framework more closely by computing our estimates on the percent of favorites chosen in each of the 17 weeks of betting. This is a continuous measure with a mean and variance that can be estimated across the 17 weeks of betting. Each individual judgment is the percent of choices for the favorite made for each of the 17 weeks of betting, denoted x_{ij}, and the criterion is the percent of times the favorite actually wins in each of the 17 weeks, denoted y_j. Some subjects were missing a large number of estimates, so we included only subjects who had missing data for less than 50% of the bets to ensure that all interindividual covariances could be computed. Our sample used $N = 164$, eliminating only 14 individuals.

The parameters in Inequality (2) are estimated by modeling the true percent of favorites winning against a point spread as a stochastic process drawn from a distribution with a fixed mean and variance for each week. The estimate of the mean and variance of the criterion are computed as the sample mean and sample variance of the true weekly percent of favorites winning. The subjects' mean judgment is computed as the sample mean of their judgments and the individual variance and covariance are computed as the sample covariance matrix between the 164 individuals. Finally, the validity is computed as the covariance between the individual judgments of the percent of favorites winning each week and the actual percent of favorites winning each week. These estimates are used to compute the LHS and RHS of Inequality (2), the measure of performance, log(R), the *MSE* between the judged and true

[3] The percentages differ depending on the method used to compute them, we report the results based on the Simmons et al. counting/median method.

Table 4
Computation of Expected Loss Using Equal Weights and Equal Probabilities for Each Individual

Condition	Crowd loss	Individual loss	log(R) (*SE*)	Proportion of individuals beat by the crowd
Choice	0.07	0.11	0.40 (0.01)	0.71
Warned choice	0.07	0.10	0.40 (0.01)	0.66
Estimate	0.02	0.08	1.16 (0.02)	1.00
Choice/estimate	0.03	0.08	1.06 (0.01)	0.96

Note. Estimates of log(R) are accompanied by *SE*s produced from the jackknife procedure.

percent of favorites winning each week and the percent of individuals beat by the crowd.

Results. Table 4 presents the expected loss for an equally weighted crowd versus randomly choosing an individual from the crowd, in each of the four experimental conditions. Our analysis supports the hypothesis of Simmons et al. by demonstrating that the choice and warned choice condition crowds performed worse than the estimate and choice/estimate condition crowds. However, our results also contradict the Simmons et al. results in that crowds have lower expected loss than a randomly selected individual and the crowd outperforms more than 50% of individuals in all of the conditions. In other words, all four conditions produced wise crowds based on expected loss. Finally, Table 5 presents the breakdown of the expected loss term for the crowd. The inferiority of the choice and warned choice conditions is driven by a much larger bias term, as predicted by Simmons et al.

Why do the two methods produce different results? There is a large discrepancy between the percent of individuals outperformed by the crowd calculated by Simmons et al. (2011) (Table 2), and the percentages obtained by our analysis (Table 5). This difference can be attributed to the different metrics used to evaluate crowd performance. The Simmons et al. method for generating crowd prediction is based on averaging all individual predictions for each individual game. A crowd choice for the favorite is produced when more than 50% of the individuals chose the favorite for that game and a crowd choice for the underdog is produced when more than 50% of the individuals chose the underdog for that game. This is a “majority choice rule” that is sensitive only to the average being above/below a threshold, but insensitive to the magnitude of the distances from the threshold.

Table 5
Breakdown of the Crowd Expected Loss Estimate

Condition	Crowd bias	Crowd covariance	Crowd covariance with criterion
Choice	0.047	0.002	−0.005
Warned choice	0.043	0.002	−0.004
Estimate	0.001	0.001	−0.003
Choice/estimate	0.001	0.002	−0.003

Figure 4 plots the proportion of individuals beat by the crowd for all possible majority choice rules. The wisdom of the crowd changes by varying the majority choice rule from 0% to 100% of choices for the favorite required to indicate a crowd choice for the favorite. For example, at the 0% choice rule, the crowd chooses the favorite for each game and performs at the base rate level of 43% correct choices (i.e., 43% of the favorites win in this experiment). At the 100% choice rule, the crowd chooses the underdog for each game, and performs at the base rate level of 57% correct. Figure 4 clearly shows that any choice rule above 60% produces a crowd that beats more than 50% of the individual members for all conditions. While Figure 4 can reveal that the top two panels exhibit more bias by shifting the step function to the right, it does not definitively reveal if any of the crowds are more or less wise.

Extensions to Small Groups Versus the Crowd

Our definition and analysis has, thus far, been restricted to the case of comparing a group prediction to that of a randomly selected individual. The framework itself could be generalized in a few directions with very minor modi-

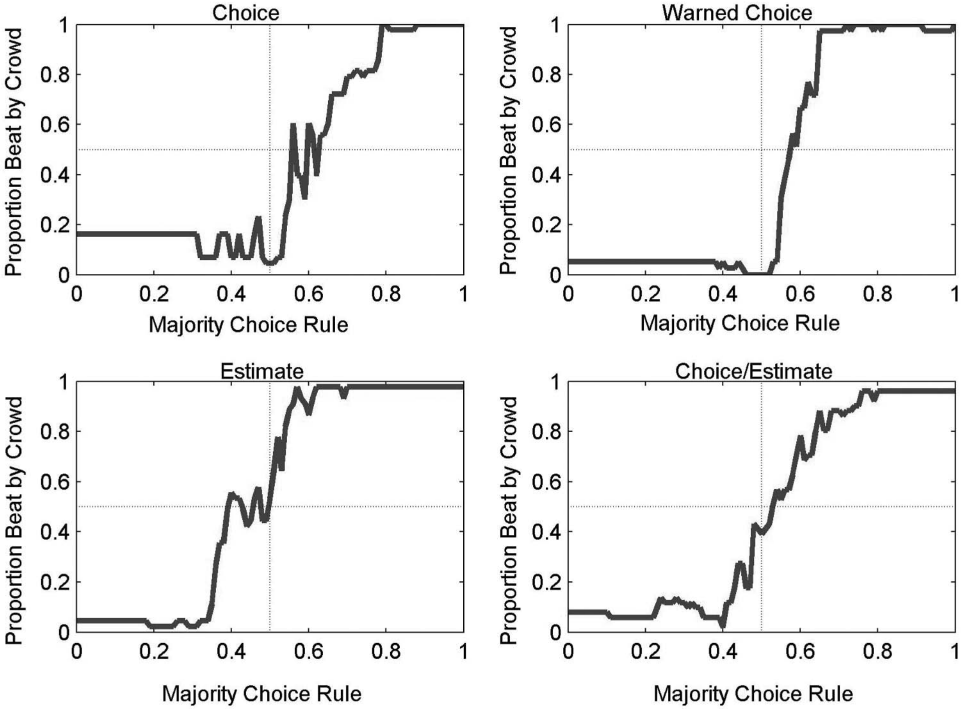

Figure 4. The proportion of individuals beat by the crowd for majority choice decision rules ranging from 0% to 100%.

fications. For example, we could consider the predictive accuracy of an aggregate of a small group of talented decision makers compared with the overall crowd. Certainly, a small group comprised of experts has the potential to outperform a larger crowd comprised of less talented members. Yet, for the small group of experts, it is reasonable to ask if the relative boost in accuracy for predicting the expected value of Y would outweigh the potentially greater gains in variance reduction by incorporating a larger number of less-talented group members. Also, performance for the small group could be further diminished if the expert predictions are highly correlated (e.g., Broomell & Budescu, 2009).

We could examine such cases by defining a new set of weights, $w_j^*, j \in \{1, 2, \ldots, N\}$, that correspond to the aggregate weighting of a small group of experts. For simplicity, we will consider the case of a small group of k-many experts such that these k-many experts are members of the larger crowd. Let the larger crowd's prediction random variable, C, be defined by the weighting scheme $w_i, i \in \{1, 2, \ldots, N\}$. The small group weighting scheme, $w_j^*, j \in \{1, 2, \ldots, N\}$, will necessarily have $N - k$ many zero weights, corresponding to the $N - k$ many individuals that are not members of the small group of experts. Let C^* be the prediction random variable defined by the weights, $w_j^*, j \in \{1, 2, \ldots, N\}$. Following a similar structure to Inequality (1), we could use the inequality, $E[(C^* - Y)^2] \leq E[C - Y)^2]$, as our definition of "small group wisdom." Given a set of crowd and small group weights, one could examine whether, and to what extent, the resulting inequality holds:

$$(\mu_X'\boldsymbol{w}^* - \mu_y)^2 + \boldsymbol{w}^{*\prime}\Sigma_{XX}\boldsymbol{w}^* - 2\boldsymbol{w}^{*\prime}\sigma_{\boldsymbol{xy}} \leq (\mu_X'\boldsymbol{w} - \mu_y)^2 + \boldsymbol{w}'\Sigma_{XX}\boldsymbol{w} - 2\boldsymbol{w}'\sigma_{\boldsymbol{xy}},$$

where $\boldsymbol{w}^*$ is the $N \times 1$ vector of weights, $w_j^*, j \in \{1, 2, \ldots, N\}$, defining C^*.

The weights determining the small group, $w_j^*, j \in \{1, 2, \ldots, N\}$, could be determined either deterministically or as a function of the correlation with *Y* (as in the individual selection mechanism in previous sections) or through some other process. For example, Budescu and Chen (2012) consider the case of weighting only the upper 50% of judges who make a positive contribution to the crowd (roughly the top 50%). As in the previous section, the deviation in accuracy for one set of weights over the other could be evaluated by taking the natural logarithm of the ratio between the left- and right-hand sides of the above inequality. This result allows us, a priori, to evaluate the effects of varying the intercorrelation among the small group DMs, the biases of their predictions, and the number of small group members.

Discussion

We have presented a precise definition of the "wisdom of the crowds" effect as well as a mathematical framework in which to evaluate it. We define a crowd as wise if a linear combination of member predictions is, on average, closer to the criterion value than the prediction of a single member who is selected according to a prespecified probability distribution. Our definition can be simply stated as an inequality (Proposition 1).

Given the popularity and ubiquity of the "wisdom of the crowds" (over 9 million hits on Google) one may be tempted to downplay the importance of this contribution. However, it is important to realize that in the vast majority of instances the effect is either not precisely defined or not defined at all. In fact, one could say that, like obscenity (Jacobellis v. Ohio, 1964) it is easy to recognize wisdom of crowds when seeing it, but rather hard to define it. In particular, the appropriate way to assess the crowd's performance is not clear because it is not obvious what is the proper comparative benchmark. Larrick et al. (2012) took an important first step in this direction by comparing the mean of the judges with the mean judge. Our article extends and generalizes this definition and illustrates its application theoretically and empirically.

Under this definition, we can specify boundary conditions on the wisdom of certain methods of crowd aggregation. Analyzing special cases of the framework, including different rules for combining judgments, different rules for selecting individuals against which to compare the crowd's judgment, cases of biased crowd members, and correlated crowd member judgments, we confirm that even a simple crowd average is robustly wise. Indeed, for large groups, nearly deterministic selection of a highly skilled individual DM is necessary before a crowd average is unwise. Of course, as Result 1 shows, there always exists a wise aggregation rule for every individual selection rule.

An advantage of our approach is that it can predict when a crowd will be wise prior to any data collection. In this manner, our framework can guide the a priori construction of an optimal (i.e., maximally wise) group. Because our framework can accommodate a wide set of constraints, for example, various bias configurations and essentially any pattern of interjudge correlations, it can be tailored for particular problems and environments that do not fit the classic conditions for crowd wisdom. Also, our definition of a crowd prediction is sufficiently general and flexible to accommodate robust aggregation measures based on trimmed or Windsorized means (e.g., Jose & Winkler, 2008), or medians (Hora, Fransen, Hawkins, & Susel, 2012).

Our general results are limited to the use of the squared error accuracy metric. Future work could consider alternative accuracy metrics, such as average absolute accuracy. This would require additional assumptions on the prediction distributions, but, in principle, our general approach could be extended to any well-defined accuracy metric. In addition, one could consider alternative generalizations, such as comparing the crowd performance with the best performing individual.

One perhaps surprising conclusion that emerges is that, contrary to the extant literature that uses the case of uncorrelated judges as the baseline (e.g., Clemen & Winkler, 1986; Hogarth, 1978), we find that a group is wisest, all things equal, when it is maximally "diverse" in that its members are as negatively correlated as possible. Though we begin with different motivations and use different mathematics, this result confirms earlier literature suggesting that diverse groups perform better (see Hong & Page, 2004). Why is diversity so important? A helpful analogy is to think of a group like a financial portfolio whose members are assets. It is useful to hedge one's bets by holding some assets that are negatively correlated with the rest of the portfolio, so that there are some positive returns when other assets perform poorly. Similarly, we find that wise groups should include some judges who predict better when others falter. In large groups, there are considerable mathematical constraints on how negatively correlated judges can be with each other. In these cases, the rule reduces to maximal performance when all judges are uncorrelated, which, under normality assumptions, implies statistical independence.

When applying our theory, it is important to distinguish between judges being independent and their judgments being uncorrelated. The crowd wisdom literature stresses the importance of independence—having the judges generate predictions without consulting, conferring and communicating—but this does not imply that their quantitative predictions will be uncorrelated. Indeed, practically all the empirical literature shows that experts in all domains are highly and positively correlated (Ashton, 1986; Clemen & Winkler, 1986; Winkler, 1971; Winkler & Poses, 1993). Broomell and Budescu (2009) describe the sources of these interjudge correlations, such as access to common information, intercorrelated cues, and similar training of the experts. Broomell and Budescu go on to illustrate how unlikely it is to find uncorrelated judges. Our framework is well suited to modeling such situations as it naturally accommodates correlated judges.

In light of these constraints, the relevance of skill-diversity trade-offs becomes apparent. Having skilled members in the group is important, but in the presence of some skilled members, it becomes more important to add members with truly different perspectives and/or access to other sources of information. Diversity of this sort is highly valuable to crowd wisdom. Our framework offers a systematic method of investigating the precise conditions under which a crowd is no longer wise. This allows us to answer questions of the form: Given a specified level of intercorrelations among the group members, how much member bias can be tolerated before the group is no longer wise (or vice versa)? A numerical example is helpful to illustrate. Consider a group with five unbiased members, who predictions highly intercorrelate with one another at .7. Suppose these five members are all skilled with a correlation of .5 with the criterion (assume the criterion random variable has a variance equal to 1). For this group, adding another member with identical attributes will create a six member group with an expected squared error value of 22. However, adding a less skilled member who correlates with the criterion at.1, but who is also less correlated with the other group members at .2, will yield a six member group with an expected squared error value of 17.8. For this example, adding a much less skilled member who created more diversity in the crowd yielded a more accurate crowd than adding another, much more skilled group member.

In our analyses, we found that the direction, pattern, and magnitude of individual biases all played a role in determining crowd wisdom. However, the overall effect of crowd wisdom was surprisingly robust to individual bias overall. In other words, unless one could identify nearly deterministically the best individual, who must be quite skilled (high correlation with the criterion), one is still better off using an unweighted aggregate. We confirmed these results with a reanalysis of an empirical study (Simmons et al., 2011) in which participants made systematically biased predictions.

Given our results, we conclude that, in general, extraordinary evidence is needed to justify choosing an expert's judgment over the aggregate of a crowd.

References

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic.

Ashton, R. H. (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes, 38,* 405–414.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics* (Vol. *1,* 2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Broomell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika, 74,* 531–553.

Budescu, D. V., & Chen, E. (2012). Identifying expertise and using it to extract the Wisdom of the Crowds. Submitted for publication.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5,* 559–583.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics, 4,* 39–46.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford, United Kingdom: Oxford University Press.

Dawes, R. M. (1970). *An inequality concerning correlation of composites vs. composites of correlations*. Eugene, OR: Oregon Research Institute Methodological Note.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84,* 158–172.

Galton, F. (1907). Vox populi. *Nature, 75,* 450–451.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121,* 149–167.

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes, 62,* 190–197.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21,* 40–46.

Hong, L., & Page, S. E. (2004). Groups of diverse problem-solvers can outperform groups of high-ability problem-solvers. *Proceedings of the National Academy of Sciences, USA, 101,* 16385–16389.

Hora, S. C., Fransen, B. R., Hawkins, N., & Susel, I. (2012, October). *Median aggregation of probabilistic judgments*. Paper presented at INFORMS meeting, Phoenix, AZ.

Jacobellis v. Ohio, 378 U.S. 184 (1964).

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting, 24,* 163–169.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffen, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, United Kingdom: Cambridge University Press.

Krause, J., Ruxton, G. D., & Krause, S. (2009). Swarm intelligence in animals and humans. *Trends in Ecology and Evolution, 25,* 28–34.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology: Social psychology and decision making. Philadelphia,* PA: Psychology Press.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52,* 111–127.

Laughlin, P. R. (1996). Group decision making and collective induction. In J. Davis & E. Witte (Eds.), *Understanding group behavior: Vol. 1. Consensual action by small groups* (pp. 61–80). Mahwah, NJ: Erlbaum.

Lee, M. D., Steyvers, M., de Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science, 4*, 151–163.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York, NY: Springer-Verlag.

Levitt, S. D. (2004). Why are gambling markets organized so differently from financial markets? *Economic Journal*, *114*, 223–246.

Linstone, H. A., & Turoff, M. (1975). *The Delphi Method: Techniques and applications*. Reading, MA: Addison Wesley.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, USA, 108,* 9020–9025.

Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgments of magnitude. In *Social computing, behavioral-cultural modeling and prediction* (pp. 236–243). Springer Berlin Heidelberg.

Miller, R. G. (1974). The jackknife-a review. *Biometrika, 61*, 1–15.

Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135,* 409–428.

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research, 38*, 1–15.

Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109,* 524.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19,* 645–647.

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences, 41,* 1–18.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association, 66,* 675–685.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science, 27,* 479–488.

Winkler, R. L., & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science, 39,* 1526–1543.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science, 36,* 452–470.

(*Appendix follows*)

Appendix

Proof of Result 1

The result follows if we show that the inequality from Proposition 1 always holds when $w_i = p_i, \forall i \in \{1, 2, \ldots, N\}$. Recall that we assume $w_i \geq 0, \forall i \in \{1, 2, \ldots, N\}$, and that $\sum_{i=1}^{N} w_i = 1$. More precisely, we demonstrate that the following inequality always holds:

$$\left(\sum_{i=1}^{N} w_i\mu_{xi} - \mu_y\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} w_iw_j\sigma_{xi,xj} - 2\sum_{i=1}^{N} w_i\sigma_{xi,y} + \sigma_y^2 \leq \sum_{i=1}^{N} w_i\left[(\mu_{xi} - \mu_y)^2 + \sigma_{xi}^2 - 2\sigma_{xi,y} + \sigma_y^2\right].$$

Expanding terms and simplifying gives,

$$\left(\sum_{i=1}^{N} w_i\mu_{xi} - \mu_y\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} w_iw_j\sigma_{xi,xj} \leq \sum_{i=1}^{N} w_i(\mu_{xi} - \mu_y)^2 + \sum_{i=1}^{N} w_i\sigma_{xi}^2,$$

which holds if, and only if,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_iw_j\sigma_{xi,xj} - \sum_{i=1}^{N} w_i\sigma_{xi}^2 \leq \sum_{i=1}^{N} w_i(\mu_{xi} - \mu_y)^2 - \left(\sum_{i=1}^{N} w_i\mu_{xi} - \mu_y\right)^2.$$

The right-hand side of the above inequality is always non-negative by Jensen's inequality. Hence, we need only show that the left-hand side of the above inequality is non-positive and the result will follow. The left-hand side of the above inequality being non-positive is equivalent to the following,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_iw_j\sigma_{xi,xj} \leq \sum_{i=1}^{N} w_i\sigma_{xi}^2,$$

which holds if, and only if,

$$\sum_{\forall(i,j),i\neq j} w_iw_j\sigma_{xi,xj} \leq \sum_{i=1}^{N} w_i(1 - w_i)\sigma_{xi}^2.$$

To prove that the above inequality always holds, consider the maximal sum on the left-hand side. Because Σ_{XX} is positive semi-definite, $|\sigma_{xi,xj}|$ is bounded above by $\frac{1}{2}(\sigma_{xi}^2 + \sigma_{xj}^2)$ for all $(i, j) \in \{1, 2, \ldots, N\}^2, i \neq j$. Substituting this upper bound for all $\sigma_{xi,xj}$ terms gives the following,

$$\frac{1}{2}\left[\sum_{\forall(i,j),i\neq j} w_iw_j\sigma_{xi}^2 + \sum_{\forall(i,j),i\neq j} w_iw_j\sigma_{xj}^2\right] \leq \sum_{i=1}^{N} w_i(1 - w_i)\sigma_{xi}^2,$$

which, by symmetry of Σ_{XX}, equals the following,

$$\sum_{\forall(i,j),i\neq j} w_iw_j\sigma_{xi}^2 \leq \sum_{i=1}^{N} w_i(1 - w_i)\sigma_{xi}^2.$$

(*Appendix continues*)

Finally, by expanding the terms on the left-hand side we will demonstrate that $\sum_{i \neq j} w_i w_j \sigma_{xi}^2 = \sum_{i=1}^{N} w_i(1 - w_i)\sigma_{xi}^2$ and that the previous inequality is, in fact, an equality. Expanding the left-hand side, we obtain

$$\begin{aligned}\sum_{\forall(i,j), i \neq j} w_i w_j \sigma_{xi}^2 = {} & \left(w_1 w_2 \sigma_{x1}^2 + w_1 w_3 \sigma_{x1}^2 + \ \ldots \ + w_1 w_N \sigma_{x1}^2\right) + \left(w_2 w_1 \sigma_{x2}^2 + w_2 w_3 \sigma_{x2}^2 + \ \ldots \right. \\ & \left. + w_2 w_N \sigma_{x2}^2\right) + \ \ldots \ + \left(w_N w_1 \sigma_{xN}^2 + w_N w_2 \sigma_{xN}^2 + \ \ldots \right. \\ & \left. + w_N w_{N-1} \sigma_{xN}^2\right), \\ = {} & \sigma_{x1}^2 w_1 \left(\sum_{i=1}^{N} w_i - w_1\right) + \sigma_{x2}^2 w_2 \left(\sum_{i=1}^{N} w_i - w_2\right) + \ \ldots \ + \sigma_{xN}^2 w_N \left(\sum_{i=1}^{N} w_i - w_N\right), \\ = {} & \sum_{i=1}^{N} w_i(1 - w_i)\sigma_{xi}^2,\end{aligned}$$

and the main result follows, thus completing the proof.

Received March 15, 2013
Revision received July 23, 2013
Accepted August 20, 2013 ■