Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Wisdom of Crowds in Matters of Taste

Johannes Müller-Trede, Shoham Choshen-Hillel, Meir Barneron, Ilan Yaniv

To cite this article:
Johannes Müller-Trede, Shoham Choshen-Hillel, Meir Barneron, Ilan Yaniv (2018) The Wisdom of Crowds in Matters of Taste. Management Science 64(4):1779-1803. https://doi.org/10.1287/mnsc.2016.2660

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 64, No. 4, April 2018, pp. 1779–1803
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Wisdom of Crowds in Matters of Taste

Johannes Müller-Trede,[a] Shoham Choshen-Hillel,[b] Meir Barneron,[c] Ilan Yaniv[c]

[a] Rady School of Management, University of California, San Diego, La Jolla, California 92093; [b] Jerusalem School of Business Administration, Hebrew University of Jerusalem, 9190501 Jerusalem, Israel; [c] Department of Psychology and the Federmann Center for Study of Rationality, Hebrew University of Jerusalem, 9190501 Jerusalem, Israel

Contact: jmullertrede@ucsd.edu (JM-T); shoham@huji.ac.il (SC-H); meir.barneron@gmail.com (MB); ilan.yaniv@huji.ac.il (IY)

Received: September 7, 2014
Revised: August 6, 2015; February 23, 2016
Accepted: August 7, 2016
Published Online in Articles in Advance: February 17, 2017

https://doi.org/10.1287/mnsc.2016.2660

Abstract. Decision makers can often improve the accuracy of their judgments on factual matters by consulting "crowds" of others for their respective opinions. In this article, we investigate whether decision makers could similarly draw on crowds to improve the accuracy of their judgments about their own tastes and hedonic experiences. We present a theoretical model that states that accuracy gains from consulting a crowd's judgments of taste depend on the interplay among taste discrimination, crowd diversity, and the similarity between the crowd's preferences and those of the decision maker. The model also delineates the boundary conditions for such "crowd wisdom." Evidence supporting our hypotheses was found in two laboratory studies in which decision makers made judgments about their own enjoyment of musical pieces and short films. Our findings suggest that although different people may have different preferences and inclinations, their judgments of taste can benefit from the wisdom of crowds.

History: Accepted by Teck-Hua Ho, judgment and decision making.
Funding: This research was supported by grants from the Israeli Science Foundation [327/10] and the Chicago Wisdom Project [John Templeton Foundation, Subaward FP050019-B].
Supplemental Material: Data are available at https://doi.org/10.1287/mnsc.2016.2660.

Keywords: advice • combining forecasts • decision making • prediction • tastes • wisdom of crowds

1. Introduction

Aggregates of multiple judgments from different sources are often more accurate and reliable than individual judgments (Armstrong 2001, Stewart 2001, Surowiecki 2005). This phenomenon is known as the *wisdom of crowds*. First documented in the scientific literature by Galton (1907), crowd wisdom and the benefits of aggregation have been demonstrated in a variety of domains from economic and business forecasting to judgment estimation problems (Ariely et al. 2000, Armstrong 2001, Budescu and Chen 2015, Clemen and Winkler 1999, Gigone and Hastie 1997, Timmermann 2006, Yaniv 2004. See Larrick and Soll 2006 for a historical review). For example, the average of several forecasts of a country's gross domestic product (GDP) growth rate is often more accurate than most of the individual forecasts.

Why are aggregate judgments often so accurate? Loosely put, different individuals tend to make different kinds of judgment errors, which cancel each other out when the judgments are aggregated. The error of the average judgment consequently tends to be smaller than the error of a randomly chosen individual judgment. Theoretical work has formalized this finding for independent judgments that are relatively unbiased (Einhorn et al. 1977, Wallsten and Diederich 2001) or are subject to heterogeneous biases (Davis-Stober et al. 2014). In addition, empirical research has shown that averaging in its simplest form is often as accurate a method for aggregating judgments as more complex weighting schemes (Makridakis and Hibon 2000, Makridakis and Winkler 1983; see also Armstrong 2001, Dawes 1979). This makes the wisdom of crowds of great practical use, since simple averages can be computed even when information about the relative quality of the different sources is not available.

The existing research has thus documented the potential, robustness, and applicability of the wisdom of crowds in matters of fact (see also Bates and Granger 1969, Brooks et al. 2015, Broomell and Budescu 2009, Clemen 1989, Gino and Moore 2007, Harvey and Fischer 1997, Hertwig 2012, Herzog and Hertwig 2009, Winkler and Makridakis 1983, Yaniv 2004). The present research investigates whether the benefits of crowd wisdom could also be obtained in matters of taste. Specifically, we tested the idea that one may be able to improve the accuracy of one's personal judgments of taste—estimates of how much one would endorse, like, or enjoy different outcomes, stimuli, or events—by averaging other individuals' judgments of their respective tastes. Since "rational choice involves . . . a guess about uncertain future preferences" (March 1978, p. 587), it is important to predict one's own hedonic and affective reactions accurately. Yet research in experimental psychology and behavioral economics has documented pervasive inaccuracies in

people's "guesses" about their future preferences (e.g., Gilbert 2006, Kahneman and Thaler 2006, Loewenstein and Schkade 1999). We aim to establish the conditions under which decision makers could enhance the accuracy of such judgments by consulting other people. It is not a foregone conclusion that one could gain accuracy by combining different individuals' judgments in this context, because different people often have different tastes.

According to the logic of crowd wisdom, a moviegoer could consult other individuals and combine their opinions to improve her estimates of facts such as a particular film's gross revenue on its opening weekend. Could the moviegoer also improve her estimate of how much she would enjoy the film by asking other individuals for their personal expectations and opinions? The two prediction problems are conceptually different. When predicting the film's gross revenue, different individuals estimate the same criterion value. By contrast, when predicting the pleasure each of them will derive from the film, different individuals estimate potentially different criterion values. This example highlights theoretical and practical questions concerning the possible benefits of the wisdom of crowds in matters of taste.

In this article, we show that despite the personal nature of tastes, it is frequently possible to exploit the wisdom of crowds to improve the accuracy of judgments on inclination and preference. We offer three main contributions. First, we provide a theoretical model that allows us to identify when and why simple averages of others' judgments can accurately predict a decision maker's own tastes. Second, we report two laboratory studies that tested and confirmed the model's predictions. Third, the existing research on advice taking and aggregation in matters of taste has focused exclusively on "asymmetric" decision settings in which the decision maker has less information about the stimuli than the individuals in the crowd (who usually have full information in the form of firsthand experience; Eggleston et al. 2015, Gilbert et al. 2009, Yaniv et al. 2011). Our approach is more general: we show that in predicting tastes (as in predicting facts), crowds can also be wise in "symmetric" settings in which the decision maker and the crowd must rely on equally limited information about a stimulus.[1] The moviegoer in the preceding paragraph, for example, could benefit not only from the impressions of others who have already seen a particular film but also from others' expectations for a film that has not yet been released.

Our theoretical framework differs from conventional models of the wisdom of crowds in judgments of fact (e.g., Davis-Stober et al. 2014, Wallsten and Diederich 2001) in allowing different individuals to estimate different criterion values. In our model, how much different people's tastes differ ("taste diversity") and how closely they resemble the decision maker's tastes ("taste similarity") emerge as key determinants of the wisdom of crowds. According to the model, potential accuracy gains from averaging a crowd's judgments of taste are greatest when the crowd consists of individuals whose tastes resemble the decision maker's but are otherwise maximally diverse—that is, *dis*similar to each other. The third key determinant of crowd accuracy in our model is the degree to which the decision maker and the people in the crowd can make discriminative and reliable judgments about their preferences ("taste discrimination"). Formally, discrimination is captured by the judgments' signal-to-noise ratio; substantively, it depends on the individuals' familiarity with the stimuli, their expertise, and the quality of the information available to them. The model predicts that while potential accuracy gains from averaging judgments of taste generally tend to be larger when individuals in the crowd discriminate better, the benefits of such discrimination should be particularly pronounced when taste similarity is high rather than low. In other words, a crowd is wisest when those individuals whose tastes most resemble the decision maker's also make the most discriminating judgments.

We tested the predictions of our theoretical model in two laboratory studies. In the first study, participants listened to musical pieces from a broad range of genres and rated how much they liked each piece and how familiar they were with it. In the second study, participants first viewed clips from a series of short films and were asked to forecast how much they would enjoy seeing the complete films. A week later, they watched the complete (five- to seven-minute) short films and then rated how much they actually enjoyed each one. In both studies, simple averages of other participants' judgments could be leveraged to accurately predict a target participant's tastes. The benefits of averaging obtained for both judgments based on full information about the stimuli (Studies 1 and 2) and ones based on limited information (i.e., the clips in Study 2). In judgments based on such limited information, the crowd even outperformed the participants' own judgments (Study 2). The laboratory data also support our model's predictions concerning the way taste similarity, diversity, and discrimination affect crowd wisdom in matters of taste.

The remainder of this article is organized as follows. In Section 2, we develop the concepts laid out in this introduction more formally, and we illustrate several key results numerically. In Section 3, we report the results of the two laboratory studies. Finally, in Section 4, we discuss our findings in relation to theories of advice taking, as well as their managerial and marketing implications.

2. Modeling Judgments of Taste and Their Aggregation

The framework for the wisdom of crowds in matters of taste proposed in this section considers various individuals who each make judgments concerning their respective hedonic reactions to, or satisfaction with, one or more stimuli. Accuracy is defined as the distance between individuals' judgments and their personal criterion values. This setup allows us to investigate the performance of "crowd judgments"—linear combinations of several individuals' judgments—in predicting the tastes of a target individual (the "decision maker"). A careful examination of the factors that determine the accuracy of crowd judgments can shed light on the conditions under which these judgments can outperform various accuracy benchmarks, including the decision maker's own judgments.

2.1. Simple Crowd Judgments

Consider a framework in which $i = 1, \ldots, N$ individuals each make judgments $x_{i,s}$ about their respective personal satisfaction value $v_{i,s}$ in response to a set of $s = 1, \ldots, S$ stimuli. The individual-specific subscript on $v_{i,s}$ implies that, unlike standard models of judgment (e.g., Davis-Stober et al. 2014, Wallsten and Diederich 2001), our framework allows criterion values to differ across individuals. Note that the framework can accommodate both symmetric decision settings in which all individuals have access to the same information and asymmetric decision settings in which some individuals know more than others (including the special case in which some individuals know some of their own satisfaction values; i.e., $x_{i,s} = v_{i,s}$ for some i, s).

We define a *simple crowd judgment* C as the arithmetic mean of a collection of individual judgments: $C = 1/N\Sigma_i x_i$. Models of judgment based on simple averages and other improper linear models have proven to be remarkably accurate and robust in other contexts (Dawes 1979, Makridakis and Winkler 1983). In the present context, the appeal of simplicity is even greater since judgments of taste are often expressed in qualitative terms requiring (potentially crude) transformations into quantitative predictions. We therefore focus mainly on simple crowd judgments rather than other linear combinations.[2]

Next, we require an appropriate standard of comparison for evaluating judgment accuracy. This reflects an important difference between conventional prediction problems involving facts and the taste prediction problem we are considering. In factual judgments, the wisdom of crowds may be assessed by computing the distance of a crowd judgment from the criterion (i.e., the true value) and comparing it to the distance of the prediction of a randomly sampled individual from the same criterion (Davis-Stober et al. 2014, Larrick et al. 2012). By contrast, in judgments of taste, where each individual holds his or her own criterion value, crowd wisdom must be assessed separately for each individual vis-à-vis the corresponding personal criteria. We speak of a *wisdom-of-crowds effect* when a crowd judgment outperforms an appropriate benchmark in predicting an individual's personal criterion values. The benchmarks we consider range from the facile (e.g., "Does the crowd judgment outperform a randomly chosen individual's judgments in predicting a target individual's satisfaction values?") to the demanding (e.g., "Does the crowd judgment outperform the target individual's own judgments in predicting his or her satisfaction values?").

Finally, we measure accuracy in terms of the mean squared error (MSE) between the judgments and the respective (individual-specific) criterion values. The (in)accuracy of a simple crowd judgment in predicting the satisfaction values $v_{D,s}$ the decision maker D derives from the stimuli is thus defined as

$$\begin{aligned} \mathrm{MSE}_{C,D} &= 1/S\Sigma_s[(C_s - v_{D,s})^2] \\ &= 1/S\Sigma_s[(1/N\Sigma_i x_{i,s} - v_{D,s})^2]. \end{aligned} \quad (1)$$

As detailed below, the MSE holds the distinct advantage over other accuracy measures that it can be decomposed into meaningful components and thus provides insights into the sources of inaccuracy. In addition, the MSE penalizes extreme errors, which is desirable because large errors are disproportionally more likely to affect the ranking of the stimuli implied by the judgments than are smaller ones. This implied ranking can be important, since in many choice settings only the highest-ranked option is ultimately chosen (Harrison and March 1984, March 1978, Smith and Winkler 2006).[3]

2.2. Decomposing Prediction Accuracy in Matters of Taste

What determines the accuracy of crowd judgments in matters of taste? We begin by addressing this question from a statistical perspective, and we decompose the MSE in Equation (1) into several different sources of error. For example, judgments may be inaccurate because they fail to rank the stimuli correctly, resulting in a poor linear correspondence between judgments and outcomes. Or judgments may systematically fall above or below the criterion values, resulting in substantial bias. Such differences in the sources of inaccuracy shed light on the nature of the benefits of relying on crowd judgments, as we shall see below.

Several decompositions of the MSE have been suggested (see Gigone and Hastie 1997, Lee and Yates 1992, Murphy 1988, Stewart 1990). We adopt a modification of a decomposition proposed by Theil (1966), which allows us to identify three components that are psychologically meaningful and relevant to the study of the

benefits of the crowd wisdom. For any predictor $\hat{y}$ of a criterion y, Theil's (1966) decomposition is given by

$$\mathrm{E}[(\hat{y}-y)^2] = (\mathrm{E}[\hat{y}]-\mathrm{E}[y])^2 + (\sigma_{\hat{y}}-\sigma_y)^2 + 2(1-\rho_{\hat{y}})\sigma_{\hat{y}}\sigma_y,$$

where $\rho_{\hat{y}}$ is the correlation between judgments and criteria or achievement correlation (see also Lee and Yates 1992). Applying this decomposition to Equation (1), we obtain

$$\mathrm{MSE}_{C,D} = (M_C - M_{vD})^2 + (\sigma_C - \sigma_{vD})^2 + 2(1-\rho_{C,vD})\sigma_C\sigma_{vD},$$

where M_C is the mean of the crowd judgments across the stimuli, M_{vD} is the mean of the decision maker's satisfaction values, σ_C and σ_{vD} are the corresponding standard deviations, and $\rho_{C,vD}$ is the correlation between the crowd judgments and the decision maker's satisfaction values across the stimuli.

From left to right, the decomposition's three components capture the following three sources of error. The first term captures *bias*—that is, to what extent the judgments are systematically higher or lower than the criteria. The second term captures *variability*—that is, to what extent the variability of the judgments is higher or lower than that of the criteria. Finally, the third term captures the (lack of) *linear correspondence*—that is, the extent to which there is a linear relation between judgments and criteria. The decomposition is valuable because it allows us to assess the accuracy of crowd judgments for each of these separate aspects.

A slight modification of the decomposition further facilitates its interpretation in the present context. As noted by Gigone and Hastie (1997), the error-minimizing relation between $\sigma_{\hat{y}}$ and σ_y is given by $\sigma_{\hat{y}} = \rho_{\hat{y}}\sigma_y$, because judgments should regress to the mean whenever they are not perfectly correlated with the criteria (i.e., when $\rho_{\hat{y}} < 1$). We therefore modify the decomposition so that its terms can be interpreted as deviations from their respective optima:

$$\mathrm{MSE}_{C,D} = (M_C - M_{vD})^2 + (\sigma_C - \rho_{C,vD}\sigma_{vD})^2 + (1-\rho^2_{C,vD})\sigma^2_{vD}. \quad (2)$$

The modified decomposition has two advantages. First, the modified variability term can be interpreted as "variability bias," as it captures the degree to which the variability of the crowd judgment deviates from the optimal degree of regression to the mean implied by the strength of its linear correspondence with the criteria. Second, the modified correspondence term does not depend on the variability of the crowd judgment. This facilitates comparisons between predictors—any differences between predictors in the correspondence term reflect differences between their achievement correlations. We rely on this property later on when we compare the accuracy of different crowd judgments as well as individuals' self-predictions.

2.3. A Simple Model of Judgments of Taste

We now complement the statistical decomposition of the MSE with a simple model of judgments of taste. To derive the model's key predictions, it is sufficient to consider the case in which $i = 1, \ldots, N$ individuals make judgments about their own satisfaction with a single stimulus. We model both the judgments x_i and the satisfaction values v_i as random variables with finite means and variances. In analogy to Equations (1) and (2), we consider the crowd judgment's expected squared error. That is, we consider the following expression:

$$\mathrm{E}[(C-v_D)^2] = (\mathrm{E}[C]-\mathrm{E}[v])^2 + (\sigma_C - \rho_{C,vD}\sigma_{vD})^2 + (1-\rho^2_{C,vD})\sigma^2_{vD}.$$

Our model is based on the assumption that the judgments x_i include a systematic and a random element. Conceptually, this assumption draws on the classic random utility model (Böckenholt 2006, Thurstone 1927). We further assume that the judgments' systematic element is given by the satisfaction values v_i, so that $x_i = v_i + e_i$, where e_i is an independent, unbiased error component ($\mathrm{E}[e_i] = 0$ for all i, and $\mathrm{Var}[e_i] = \sigma^2_{ei}$). In addition, we make the convenience assumption that the expected value of the satisfaction values $\mathrm{E}[v_i]$ is the same for all i.[4] All of our modeling assumptions (both conceptual and convenience) are discussed in more detail in Section 2.5. As we next detail, the particular assumptions we make allow us to identify several key variables that are psychologically meaningful and that jointly determine the wisdom of crowds in matters of taste.

From a mathematical standpoint, our assumptions allow us to simplify the above expression. Consider, for example, the covariance between the crowd judgment and the decision maker's satisfaction values (which determines the correlation $\rho_{C,vD}$). When judgments are modeled as the sum of the associated satisfaction values and an independent error term, this covariance is readily calculated as

$$\begin{aligned}\mathrm{Cov}[1/N\Sigma_i x_i, v_D] &= \mathrm{Cov}[1/N\Sigma_i(v_i+e_i), v_D] \\ &= 1/N\Sigma\rho_{vi,vD}\sigma_{vi}\sigma_{vD}.\end{aligned}$$

These and similar calculations based on the properties of the expected value operators and variance operators allow us to express the expected squared error of the crowd judgment exclusively in terms of correlations and (co)variances (for details, see the appendix):

$$\mathrm{E}[(C-v)^2] = -2/N\Sigma\rho_{vi,vD}\sigma_{vi}\sigma_{vD} + 1/N^2\Sigma\Sigma\rho_{vi,vj}\sigma_{vi}\sigma_{vj} + 1/N^2\Sigma\Sigma\rho_{ei,ej}\sigma_{ei}\sigma_{ej} + \sigma^2_{vD}. \quad (3)$$

Equation (3) captures our model's major insights and pinpoints the factors that determine the accuracy

of crowd judgments. These include the *taste similarity* between the people in the crowd and the decision maker, captured by the correlation between their respective satisfaction values $\rho_{vi,vD}$, and the crowd's *taste diversity*, captured by the correlation between its people's satisfaction values $\rho_{vi,vj}$. They also include the *crowd size* N, as well as the individuals' *taste discrimination* and the crowd's *error diversity*, captured by the signal-to-noise ratios $\sigma^2_{vi}/\sigma^2_{ei}$ and the correlations between the individuals' judgment errors $\rho_{ei,ej}$, respectively. The following paragraphs analyze each of these factors in more detail. To simplify the exposition, crowds do not include the decision maker (i.e., $i, j \neq D$) unless otherwise indicated.

2.3.1. Taste Similarity. The first term on the right-hand side of Equation (3) shows that the MSE of crowd judgments decreases in the correlation $\rho_{vi,vD}$ between the decision maker's satisfaction values and those of the people in the crowd. This *taste similarity* between the decision maker and the crowd emerges as a first determinant of the wisdom of crowds in matters of taste. That taste similarity is beneficial is straightforward: as similarity between individuals increases, the differences between their criterion values decrease. The taste prediction problem then increasingly comes to resemble a factual prediction problem in which a single criterion value is shared by all individuals, and the standard results for the wisdom of crowds (in matters of fact) apply (cf. Davis-Stober et al. 2014, Larrick et al. 2012).

The role played by taste similarity in matters of preference thus resembles that played by expertise in matters of fact (e.g., Budescu and Chen 2015, Davis-Stober et al. 2014). The two concepts, however, while related, are not equivalent. Loosely put, expertise captures an individual's ability to consistently and reliably predict a particular quantity of interest; formally, it is often defined as the correlation between a set of judgments and a set of criteria (Davis-Stober et al. 2014, Hogarth 1978; see also Broomell and Budescu 2009, Einhorn 1974, Weiss and Shanteau 2003 on defining expertise.) By contrast, taste similarity merely captures the congruence between the quantities that are of interest to different individuals; formally, it is defined as the correlation between two sets of criteria. In our model of individuals making judgments about their own preferences, the correlations between sets of judgments and sets of criteria are bounded by the correlations between sets of criteria: for $\sigma^2_{ei} > 0$, $|\rho_{xi,vD}| < |\rho_{vi,vD}|$. In other words, our model asserts that expertise in matters of taste is a function of both taste similarity and taste discrimination, defined below: being an "expert" for another's tastes requires both congruent criteria and the ability to accurately predict these criteria.[5]

2.3.2. Taste Diversity. Next, consider the second term on the right-hand side of Equation (3). Other things being equal, the MSE of crowd judgments increases in the correlation $\rho_{vi,vj}$ between the satisfaction values of the people in the crowd. This correlation term, which we refer to as *taste diversity*, is the second determinant of the accuracy of crowd judgments in matters of taste. At first, its effect may seem counterintuitive: a crowd is most accurate when the pairwise correlations between the individuals' satisfaction values are as low as possible, ideally even negative. The effect becomes more intuitive when viewed through the lens of redundancy. The higher the correlations between different individuals' satisfaction values, the more redundant the values, and less could be gained from averaging them. Recent work shows how similar diversity considerations play a role in the analysis of predictions in the factual domain (Davis-Stober et al. 2014). Our model yields the novel conclusion that in the domain of taste, diversity is beneficial not only for individual predictions but also for the (different) *criteria* that different individuals aim to predict. In other words, holding similarity to the decision maker constant, crowds including individuals with heterogeneous tastes are wiser than crowds including individuals with homogeneous tastes.

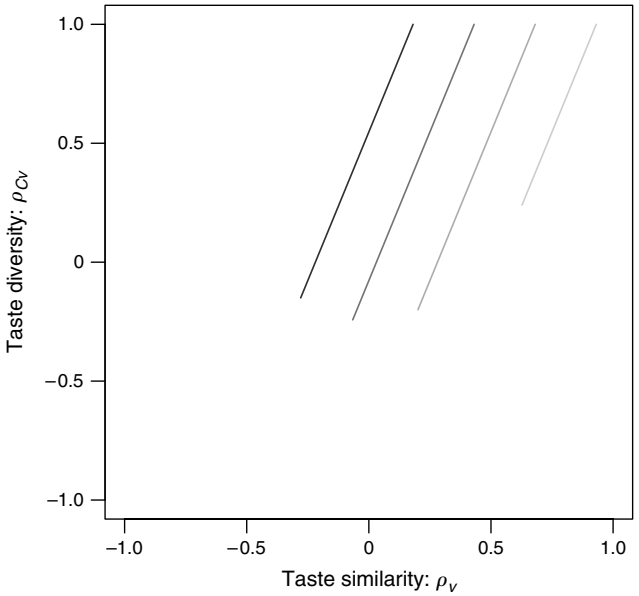
Now consider taste diversity and taste similarity jointly. For the crowd judgment to be accurate, the individuals in the crowd should resemble the target individual in their tastes and yet differ from one another. But when all the individuals resemble the decision maker, how diverse can they be? Accuracy gains from diversity constrain potential accuracy gains from taste similarity, and vice versa. In mathematical terms, these constraints are imposed by the positive semidefiniteness of the covariance matrix.

The following example explores these relations between taste similarity, taste diversity, and the accuracy of the crowd judgment. It considers a crowd of moderate size with $N = 10$ individuals. To isolate the effects of similarity $\rho_{vi,vD}$ and diversity $\rho_{vi,vj}$, the example assumes high taste discrimination, discussed below, so that $\sigma^2_{vD} = \sigma^2_{vi} = 1$ and $\sigma^2_{ei} = 0.1$ for all i, and sets error diversity, also discussed below, to $\rho_{ei,ej} = \rho_{Ce} = 0.5$ for all $i, j \neq D$. Since the judgments are assumed to be highly discriminative, Example 1 may be thought of as a stylized representation of a decision setting in which the individuals in the crowd base their judgments on firsthand experience, as in the asymmetric settings studied by Gilbert et al. (2009) and Yaniv et al. (2011).

Example 1 (Taste Similarity and Diversity). Let $\rho_{vi,vD}=\rho_v$ and $\rho_{vi,vj} = \rho_{Cv}$ for all $i, j \neq D$. Figure 1 shows the trade-off between taste similarity and taste diversity in simple crowd judgments.

The lines in Figure 1 are "iso-MSE lines"; that is, each line represents combinations of parameter values for ρ_v and ρ_{Cv} that result in an identical MSE. Darker

Figure 1. Accuracy of Crowd Judgments as a Function of Taste Similarity ρ_v and Taste Diversity ρ_{Cv}: Iso-MSE Curves of Simple Crowd Judgments for Crowds of Size 10



Note. Darker shades of gray indicate larger MSEs.

shades of gray indicate larger MSEs: lines farther to the right represent more accurate crowd judgments (e.g., for the lightest line, $\text{MSE}_{C,D} = 0.2$, and for the darkest line, $\text{MSE}_{C,D} = 1.4$). Importantly, the slopes of the lines capture the trade-off between the two factors. Note that higher values of ρ_v represent greater taste similarity, whereas higher values of ρ_{Cv} represent *less* taste diversity. Consequently, all iso-MSE lines have positive slopes.

The steepness of the lines implies that taste similarity has a greater effect on the accuracy of the crowd judgments than taste diversity (i.e., their slope is greater than 1). Nonetheless, the effect of taste diversity is sizable. For example, the MSE of a crowd of 10 heterogeneous individuals ($\rho_{Cv} = 0.24$) with tastes moderately similar to the decision maker's ($\rho_v = 0.63$) is the same as the MSE of a crowd of 10 homogeneous individuals ($\rho_{Cv} = 0.93$) whose tastes are identical to the decision maker's ($\rho_v = 1$). Both taste similarity and taste diversity can increase the accuracy of crowd judgments.

At the same time, Figure 1 illustrates how the two factors constrain one another: the iso-MSE lines "break off" for low values of taste diversity. As a result, the lines do not reach a sizable blank region at the bottom of Figure 1. This breaking off, which occurs sooner (i.e., at higher levels of diversity) for more extreme levels of taste similarity, represents the constraints imposed by the positive semidefiniteness of the correlation matrix. In intuitive terms, individuals who are very dissimilar in their tastes cannot concurrently all be very similar to (or very dissimilar from) the decision maker. These constraints, which are powerful for the moderately sized crowds of $N = 10$ shown in Figure 1, are even more restrictive for larger crowds.

Finally, recall that Example 1 may be thought of as representing a decision setting in which individuals in the crowd base their judgments on firsthand experience. The trade-off between taste similarity and taste diversity exists even when the crowd judgments are based on full information. In fact, our model asserts that the trade-off is inherent to all crowd judgments in matters of taste, regardless of the quality of the information they are based on. In what follows, we show that taste similarity also affects the relation between information quality and judgment accuracy. To do so, we first introduce a third determinant of the accuracy of crowd judgments: taste discrimination.

2.3.3. Taste Discrimination. According to Equation (3), the MSE of crowd judgments depends on a number of variance terms. The magnitude of the effect of taste diversity, for example, depends on the standard deviations of the satisfaction values σ_{vi} of the people in the crowd (second term). The MSE also depends on the standard deviation of the judgment error σ_{ei} and on the standard deviation of the decision maker's satisfaction values σ_{vD}. We summarize these effects as *taste discrimination*, defined by the signal-to-noise ratios $\sigma^2_{vi}/\sigma^2_{ei}$ and $\sigma^2_{vD}/\sigma^2_{eD}$.

For individuals with poor taste discrimination, making confident, reliable judgments about their preferences is difficult. Poor discrimination may occur because an individual lacks the knowledge or the familiarity required to distinguish between the stimuli, resulting in a low value for the "differentiation component" σ_v. Alternatively, poor discrimination may occur when judgments are subject to large errors, resulting in a large "noise component" σ_e. In defining taste discrimination in terms of these two components, we build on previous work on judgmental discriminability (Linville and Fischer 2004, Yaniv et al. 2011). Equation (3) shows how each component affects the accuracy of the crowd judgment.

Consider the decision maker first. His or her ability to discriminate affects the accuracy of the crowd judgment via the differentiation component σ_{vD}, which enters Equation (3) twice. In conjunction with taste similarity, it reduces the MSE (first term), but importantly, it also directly contributes to the MSE (last term). The intuition for the latter effect is that as decision makers discriminate more confidently among the stimuli, their preferences become more nuanced and fine grained. Other things being equal, this makes their satisfaction values more difficult to predict, which limits the usefulness of crowd judgments.[6]

The taste discrimination of the individuals in the crowd affects the accuracy of the crowd judgment via both its noise component σ_{ei} and its differentiation

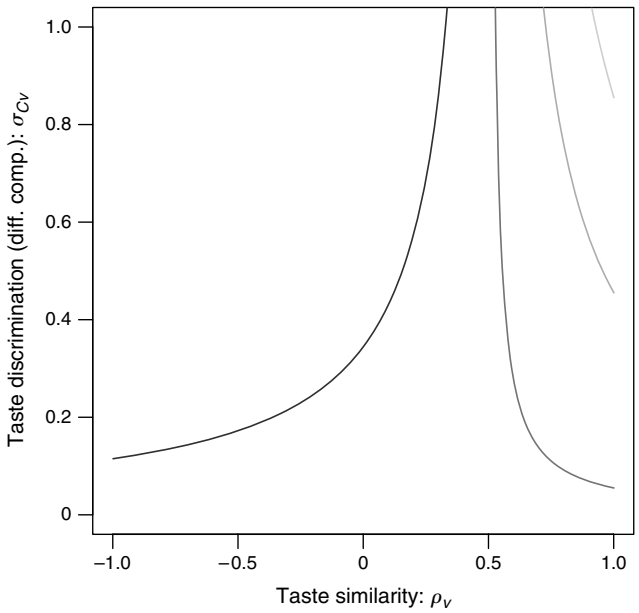
component σ_{vi}. First, the noise component σ_{ei} enters in the third term on the right-hand side of Equation (3). Its implication is straightforward: the less noise in the individuals' judgments concerning their own personal preferences, the more accurate the crowd judgment. Next, consider the differentiation component σ_{vi}. Its effect is subtler, as it enters the MSE of crowd judgments negatively in conjunction with taste similarity (first term) and positively in conjunction with taste diversity (second term). The resulting net effect of σ_{vi} on the MSE can be negative or positive. Other things being equal, the greater the taste similarity between an individual in the crowd and the decision maker, the more favorable the effect of σ_{vi} on the MSE. Increasing the crowd's taste discrimination (e.g., by providing individuals in the crowd with additional information) thus has complex effects on judgment accuracy that interact with taste similarity. Intuitively speaking, when individuals in the crowd share the decision maker's tastes, their judgments will be more useful when they reflect nuances and subtle differences among the stimuli. If, on the other hand, the other individuals' preferences do not align with the decision maker's, the nuances and differences reflected in their judgments merely add uninformative variability to the crowd judgment from the decision maker's perspective.

Our second example explores these relations between the differentiation component of taste discrimination, taste similarity, and the accuracy of crowd judgments in more detail. Again, we consider moderately sized crowds with $N = 10$, and as before, we make auxiliary assumptions about the remaining parameters to isolate the effect and set $\sigma^2_{vD} = 1$, $\sigma^2_{ei} = 0.1$ for all i and $\rho_{ei,ej} = \rho_{Ce} = 0.5$ for all $i, j \neq D$. We also assume $\rho_{vi,vj} = 1$ for all $i \neq D$ to make the constraints imposed by the positive semidefiniteness of the correlation matrix nonbinding over the range of ρ_v.[7]

Example 2 (Taste Discrimination and Similarity). Let $\rho_{vi,vD} = \rho_v$ and $\sigma^2_{vi}/\sigma^2_{ei} = \sigma^2_{Cv}/\sigma^2_{Ce}$ for all $i, j \neq D$. Figure 2 shows the relation between the differentiation component of taste discrimination and taste similarity in simple crowd judgments.

As in Figure 1, the curves in Figure 2 are iso-MSE curves representing the combinations of parameter values that yield a particular MSE, with darker grays indicating larger MSEs. Figure 2 shows that greater discrimination often reduces the MSE. For a wide range of parameter values in the region of moderate-to-high taste similarity, the iso-MSE curves slope downward. In other words, more discriminating crowds yield more accurate crowd judgments. Moreover, the curves flatten as similarity increases: the higher the taste similarity, the larger the gains from increasing taste discrimination. When taste similarity is sufficiently low, however, greater discrimination can increase the MSE. This effect is illustrated by the leftmost iso-MSE curve, which is upward-sloping. In this region in the parameter space, increasing discrimination while holding taste similarity constant leads to less accurate crowd judgments. Intuitively, individuals whose judgments predict a set of criteria that bears little or no resemblance to the decision maker's criteria can impair the accuracy of the crowd judgment by supplying reliable judgments ruled by these "alien" criteria.

Figure 2. Accuracy of Crowd Judgments as a Function of Taste Similarity ρ_v and the Differentiation Component of Taste Discrimination σ_{Cv}: Iso-MSE Curves of Simple Crowd Judgments for Crowds of Size 10



Note. Darker shades of gray indicate larger MSEs.

The final effect of the crowds' ability to discriminate on the accuracy of the crowd judgment combines the effect of the differentiation component σ_{vi} explored in Example 2 with that of the noise component σ_{ei} discussed previously. Recall that decreasing σ_{ei} decreases the MSE of the crowd judgment, regardless of taste similarity. This will attenuate the combined effect compared with the pattern in Figure 2. Nonetheless, our model asserts that in predicting matters of taste, the potential benefits of providing crowds with more information are conditional on taste similarity. As a direct consequence, it also asserts that the benefits of familiarity and firsthand experience will be moderated by taste similarity. We describe our empirical test of these assertions in Section 3 below, after discussing the remaining two parameters in Equation (3), error diversity and crowd size.

2.3.4. Error Diversity. The third term on the right-hand side of Equation (3) shows that the effects of (the noise component of) taste discrimination are qualified

by the correlation between the judgment errors $\rho_{ei,ej}$. This correlation captures the crowds' *error diversity*. Similar to taste diversity, error diversity increases the accuracy of crowd judgments: crowd judgments are most accurate when the correlations between the judgment errors of the people in the crowd are as low as possible, even negative. Moreover, unlike the case of taste diversity, the only meaningful constraint that the positive semidefiniteness of the correlation matrix imposes on judgment errors is that these correlations cannot be negative, on average, for large crowds (cf. Davis-Stober et al. 2014). Put differently, there are no benefits to similarities in judgment errors, and the benefits of error diversity thus do not trade off with any benefits derived from similarity (whereas the benefits from taste diversity do trade off with the benefits from taste similarity). The more diverse the individuals' judgment errors, the more accurate the crowd judgment.

2.3.5. Crowd Size. Finally, consider the relation between the *crowd size* N and the accuracy of crowd judgments. We discuss this parameter last because the effects of crowd size are difficult to isolate: almost all terms in Equation (3) feature N. An upside to this, however, is that analyzing the marginal effects of crowd size by comparing the MSEs of simple crowd judgments of size N and $N+1$ allows us to summarize many of the effects we have discussed thus far.

Adding a new ($N+1$th) individual to the crowd can increase the accuracy of the crowd judgment if the taste similarity between the new individual and the decision maker is larger than the average similarity between the N individuals already in the crowd and the decision maker (i.e., the mean correlation between the satisfaction values of these N individuals and those of the decision maker). This is not sufficient, however—accuracy gains from increasing taste similarity can be offset if adding a new individual makes the crowd less diverse in taste (and) or error. Conversely, it is not necessary, either—if adding a new individual increases taste or error diversity, this can lead to an accuracy gain even if the taste similarity between the new individual and the decision maker is smaller than the average taste similarity between the individuals already in the crowd and the decision maker. The relation between size and accuracy is consequently not necessarily monotonic: larger crowds are not always more accurate than smaller crowds. These results are reminiscent of Hogarth's (1978) analysis of crowd size, expertise, and interjudge correlations as determinants of the accuracy gains from combining factual judgments (see the appendix for details).

2.4. Optimality: MSE-Minimizing Weights

Before moving on to the experimental tests of the model, we briefly consider crowd judgments that are not based on simple averages. In the appendix, under the auxiliary assumption that $\mathrm{E}[v_i] = 0$ for all i, we derive the optimal linear weighting of predictions that would minimize the MSE of crowd judgments of the form $\Sigma_i w_i x_i$. We show that the optimal (MSE-minimizing) weights w_i^* that the decision maker should place on the different judgments are of the form $w_i^* = a\rho_{vi,vD} + b$, where $\rho_{vi,vD}$ is the taste similarity between the individual i and the decision maker, a is a term that is inversely related to the variance of i's judgments, and b is a term that accounts for the crowd's overall correlation structure (i.e., its taste and error diversity). Other things being equal, the optimal weights therefore increase in taste similarity: the decision maker should place larger weights on the judgments of similar others than on those of dissimilar others. This result also implies that individuals can benefit from combining their own judgments of taste with the crowd judgment, and they should often place more weight on their own judgments than on others' (since individuals are "maximally similar" to themselves; i.e., $\rho_{vD,vD} = 1$). This conclusion contrasts with the finding that individuals should usually not place more weight on their own judgment than on other people's judgments when predicting objective facts (e.g., Soll and Larrick 2009, Yaniv 2004).

2.5. Modeling Assumptions

We conclude this section with a brief discussion of the assumptions underlying our model. These include two substantive and two convenience assumptions.

First, our method implicitly assumes that judgments of taste are measurable on an interval scale (or that utility is "cardinal"). In other words, it assumes that degrees of differences between judgments (and, in turn, their means and variances) are meaningful. Modern utility theory and stochastic choice models derived from it are often interpreted in terms of the less restrictive assumption that utility is measurable on an ordinal scale (see, e.g., Luce and Suppes 1965, Manski 1977, Roberts 1985). The cardinality assumption is central to our work and to other efforts to understand judgments in matters of taste. Similar assumptions are made in research programs exploring utility predictions and process interpretations of utility theory (e.g., Eggleston et al. 2015; Gilbert and Wilson 2007; Kahneman and Snell 1990, 1992; Kahneman et al. 1997; Loewenstein and Schkade 1999) and subjective well-being measured as happiness or life satisfaction (e.g., Deaton 2008, Diener and Diener 1996, Stevenson and Wolfers 2008).

Second, we distinguish between a taste component and an independent, additive error component in judgments. Assuming an additive relation between the two components greatly facilitates the derivation of the results in Section 2.3 (because of the linearity of

the expectation operator). Further assuming independence between the two components similarly simplifies the algebra. The general pattern of our results, however, would remain qualitatively unchanged if this assumption were relaxed. A positive correlation between the taste and the error component, for example, would reduce the potential for wisdom-of-crowds effects, but it would not eliminate them. We also believe that in many contexts the independence assumption is descriptively plausible, as individuals with similar tastes need not necessarily make similar judgment errors.

Third, the assumption that judgments are unbiased is largely a convenience assumption. In light of the existing experimental evidence for cognitive bias in predicting affective reactions (cf. Kahneman and Thaler 2006, Wilson and Gilbert 2005), we note that the framework could readily accommodate biased judgments. It is straightforward to show that heterogeneous biases would be attenuated by the averaging process underlying crowd judgments. Cognitive biases shared by multiple individuals, by contrast, could persist even in averages, and they would decrease the accuracy of the crowd judgment.

The fourth and final assumption is that that the expected value of the satisfaction values $\mathrm{E}[v_i]$ is the same for all i. Primarily a convenience assumption, it rules out scale differences, i.e., individual differences in the overall propensity to enjoy the stimuli. Its main implication is that crowd judgments are unbiased, paralleling the unbiasedness assumption for individual judgments. If this assumption were to be relaxed, decision makers with different propensities to enjoy the stimuli would have to first normalize each other's judgments for these to be useful. (For deriving the optimal weights in the appendix, we further normalize $\mathrm{E}[v_i]$ to 0; the interpretation remains unchanged.) Our framework could readily be extended to accommodate this case.

3. Empirical Evidence for the Wisdom of Crowds in Matters of Taste

In this section, we report two laboratory studies designed to test the predictions of our analytical framework. In Study 1, participants rated how much they enjoyed several pieces of music after listening to them. In this setting, we tested how well the judgments made by a "crowd" of several participants could predict a target participant's judgments of taste. In the first study, the crowd thus had access to fairly reliable information, as in many "asymmetric" decision settings (i.e., the participants experienced the stimuli before making their judgments). In Study 2, participants first viewed brief excerpts from a series of short films and were asked to forecast how much they would enjoy seeing the complete films. A week later, they watched the complete (five- to seven-minute) films and then rated how much they actually enjoyed each of them. We assessed the predictive accuracy of crowd judgments based on full information (as in Study 1) as well as ones based on limited information (i.e., the brief excerpts). This study thus goes beyond Study 1 by allowing us to assess crowd wisdom in a symmetric setting in which the crowd and the decision maker based their judgments on equally limited information. The data from both studies support the predictions of our theoretical analysis for similarity, diversity, and discrimination.

3.1. Study 1: Music

The participants in Study 1 listened to musical pieces and then rated how much they liked each piece and how familiar they were with it. All judgments in Study 1 were thus based on fairly reliable information, and we treat them as approximating the participants' satisfaction values (i.e., in terms of our framework, we assume $\sigma_e^2 \approx 0$). In this setting, we assessed how well crowd judgments predicted a target individual's liking for the music. Critically, the study included a large number of pieces selected from a broad range of genres, each one paired with a second, similar piece. This design feature allowed us to reliably estimate the similarity between different participants on a subset of the musical pieces and then use the remaining pieces to test the predictions of our theoretical framework for the effects of taste similarity. Finally, the participants' familiarity ratings enabled us to investigate how taste discrimination affected the accuracy of crowd judgments.

3.1.1. Method

Participants. One hundred and eight undergraduate students participated in the study in exchange for the equivalent of $5. Four of them failed to provide answers to some of our questions and were excluded from all analyses.

Materials. The stimuli consisted of 11 pairs of musical pieces—that is, 22 pieces in total. Each pair consisted of two works by the same composer or performer (e.g., two orchestral pieces by Johann Sebastian Bach, two songs by Bob Dylan from the same album). The stimuli were selected from a wide range of musical styles to create meaningful variability in the participants' familiarity with and liking for the music. Among other styles, they included classical music, African folk music, and national and international pop.

Procedure. The participants were run in groups of 10–18, individually seated in a large, quiet classroom. The experimenter distributed response sheets and explained the task. Participants were told they would be asked to judge how much they liked each piece and how familiar they were with it. To help them interpret

the response scales for liking and familiarity (ranging from 1 ("not at all") to 10 ("very much")), they first listened to several 20-second clips from a diverse selection of unrelated musical pieces. The music, which was prerecorded on a CD, was played from the front of the classroom, using a CD player with large loudspeakers. After this warm-up phase, participants listened to one-minute excerpts from each of the 22 musical pieces. The pieces were presented in a fixed order and in two sets that each featured one piece from each of the 11 pairs. The order was the same for all subjects and across sets (e.g., the first Bob Dylan song was the third piece in the first set, and the second Bob Dylan song was the third piece in the second set). After listening to each excerpt, participants were asked to rate it on four scales. The first two questions targeted enjoyment: participants were asked, "How much did you like the music excerpt?" and "How likely would you be to listen to this piece in your free time?" The other two questions targeted familiarity. Participants were asked, "How familiar are you with this specific piece of music?" and "How familiar are you with this musical genre?" All four questions used the 10-point scales described above (see Loewenstein and Schkade 1999 for a discussion of methods for measuring subjective experiences). After they rated the final piece of music, the participants were thanked and paid for their participation. The study did not include any measures or conditions that are not reported.

3.1.2. Results. Consider first participants' *enjoyment scores*, calculated as the average of their two enjoyment judgments (which were correlated at mean $r = 0.85$ at the participant level). As stated above, we treat these scores as approximating participants' satisfaction values.[8] Enjoyment scores varied considerably, both within and across participants. At the participant level, the mean enjoyment score across the 22 musical pieces ranged from 3.2 to 6.8 (grand mean = 5.0). The median of the standard deviations associated with these means was 3.2, showing substantial within-participant variation in tastes (i.e., individual participants liked particular pieces more than others). In addition, for each possible pair of participants, we computed the correlation between their enjoyment scores across all musical pieces. On average, these pairwise correlations were low (mean $r = 0.06$, SD = 0.25, 95% confidence interval (CI) between 0.04 and 0.08), which suggests that tastes also varied a great deal across participants and that there were no universal standards for judging the music.[9]

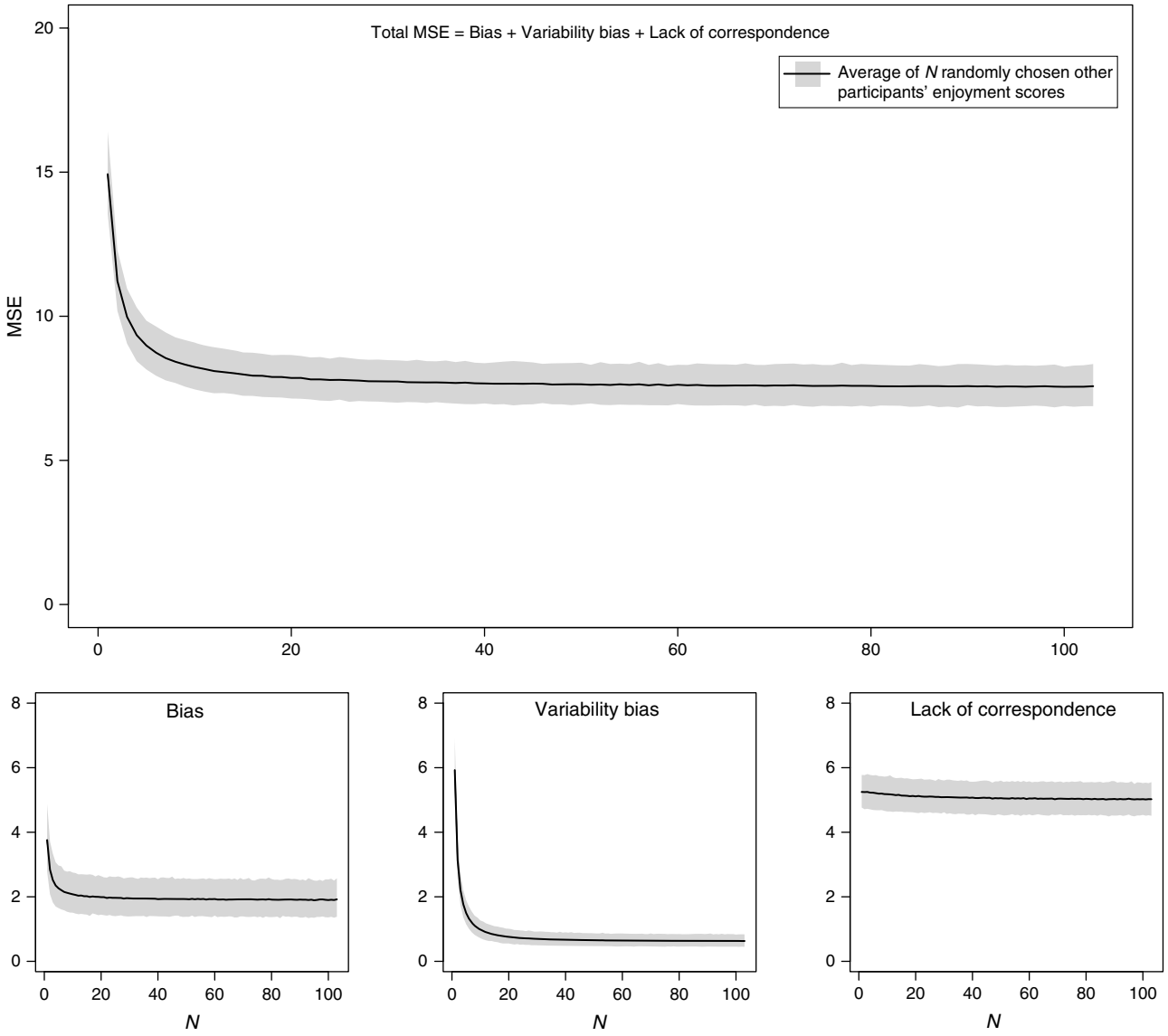
Similarly, we calculated *familiarity scores* as the average of the two familiarity ratings (which correlated at mean $r = 0.66$ at the participant level). These scores, too, exhibited substantial within- and between-participant heterogeneity. The mean familiarity scores across the 22 musical pieces ranged from 2.1 to 7.8 (grand mean = 4.9), and the median of the standard deviations associated with these means was 2.7. The participant-level pairwise correlations of the scores were also quite low (mean $r = 0.26$, SD = 0.25, 95% CI between 0.22 and 0.30). This suggests that while participants were more familiar with some pieces of music than with others, familiarity scores still varied considerably.

Crowd Wisdom: Randomly Sampled Participants. We now examine our participants' enjoyment scores for the music to assess a basic tenet of crowd wisdom: Could our participants use one another's personal judgments to predict their own musical tastes? To address this question, we created "crowds" of various sizes by randomly sampling participants, and we averaged their enjoyment scores to predict each participant's scores. According to our theoretical framework, the accuracy of such crowd judgments should increase in the size of the crowd, rapidly at first and then more slowly as the crowd size grows further. In particular, we predict a basic wisdom-of-crowds effect in which average scores of random samples of participants should outperform the scores of a single randomly sampled participant.

We employed a bootstrap method to calculate crowd judgments and their accuracy (e.g., Davison and Hinkley 1997). First, we sampled participants with replacement from the original data set to construct a bootstrap sample. Subsequently, we matched each participant in the bootstrap sample with N other participants, sampled without replacement from the original data set, and calculated the average enjoyment score for each musical piece across these N participants. For each bootstrap sample, we then computed the accuracy of the crowd judgment as the MSE between each participant's enjoyment score for each piece and the corresponding average score of the N other participants he or she had been matched with.[10] Finally, we repeated this procedure for 2,000 bootstrap samples and for N ranging from 1 to 103. The main panel in Figure 3 shows the average MSE across the 2,000 bootstrap samples as well as the 95% bootstrap percentile confidence intervals, i.e., the 2.5th and the 97.5th percentiles of the distribution of bootstrap estimates.[11] In addition, several key crowd judgments are summarized in Table 1.

Figure 3 shows that averaging other people's judgments about their respective preferences can indeed be beneficial in predicting musical taste. Its extremes illustrate the predicted wisdom-of-crowds effect: at an MSE of 14.9, the enjoyment score of a single random participant was a substantially less accurate predictor than the average enjoyment score of all other participants (MSE = 7.5; see Table 1). Figure 3 also reveals that the accuracy gains from combining judgments decreased rapidly as crowd size increased. Moderately sized crowds (i.e., between 5 and 15 participants) performed on par with much larger crowds. This result

Figure 3. Study 1: Accuracy of Randomly Selected Crowds in Predicting a Target Participant's Enjoyment Scores



Note. Shown are the bootstrap estimates of the mean (line) and 95% confidence intervals (shaded areas).

mirrors similar findings for factual judgments discussed below. It is also consistent with our theoretical framework, according to which additional participants should only produce accuracy gains if they increase diversity or if their tastes are more similar to the target participant's than the tastes of the other participants already in the crowd (see Section 2.3). Neither is likely for large crowds of randomly chosen participants.

The three smaller panels in the bottom of Figure 3 decompose the MSE into its three components discussed in Section 2.2: bias, variability bias, and error resulting from a lack of linear correspondence. In estimating these components, we employed the same bootstrap method described above. Again, the panels show average values across the 2,000 bootstrap samples as well as the 2.5th and the 97.5th percentiles of the distributions of bootstrap estimates. The decomposition reveals that the crowd judgment's advantage lies in reducing variability bias and, to a lesser extent, bias, whereas the error resulting from a lack of correspondence was hardly affected by the averaging procedure (see also Table 1). In other words, crowd judgments of others sampled at random induce a beneficial regression to the mean in predictions and reduce systematic error from over- and underpredicting. Finally, the reductions in the MSE's individual components obeyed the same pattern of rapidly decreasing change with increasing crowd size that characterized the total MSE.

Crowd Wisdom: Taste Similarity. Having established these important basic findings, we now test our

Table 1. Accuracy of Key Crowd Judgments Based on Enjoyment Scores in Study 1

			95% CI		MSE decomposition		
Predictor	MSE	SD	Low	High	Bias	Variability bias	Correspondence
Crowd enjoyment scores							
One randomly selected other	14.9	7.5	13.5	16.4	3.8	5.9	5.2
Ten randomly selected others	8.2	4.1	7.5	9.0	2.1	1.0	5.2
One most similar other	9.9	5.9	8.8	11.1	3.0	2.6	4.3
Ten most similar others	5.8	3.4	5.2	6.5	1.8	0.4	3.6
All others	7.5	3.6	6.8	8.2	1.9	0.6	5.0

Notes. All statistics were calculated using the bootstrap method described in Section 3.1. "Bias," "Variability bias," and "Correspondence" refer to the three components of the decomposition in Equation (2) in Section 2.2. Any discrepancies between the MSE and the sum of its three components reflect rounding errors.

similarity hypothesis: according to our model, the benefits of averaging should be greater for those who share similar tastes. To test this, we calculated a second set of crowd judgments based on crowds of participants selected for their similarity to the target participant. These lend themselves to more stringent tests of wisdom-of-crowds effects: will a single participant whose tastes resemble the target participant's be more or less accurate in predicting the latter's enjoyment scores, for example, than the crowd of all other participants? And how will small crowds of similar participants fare in comparison?

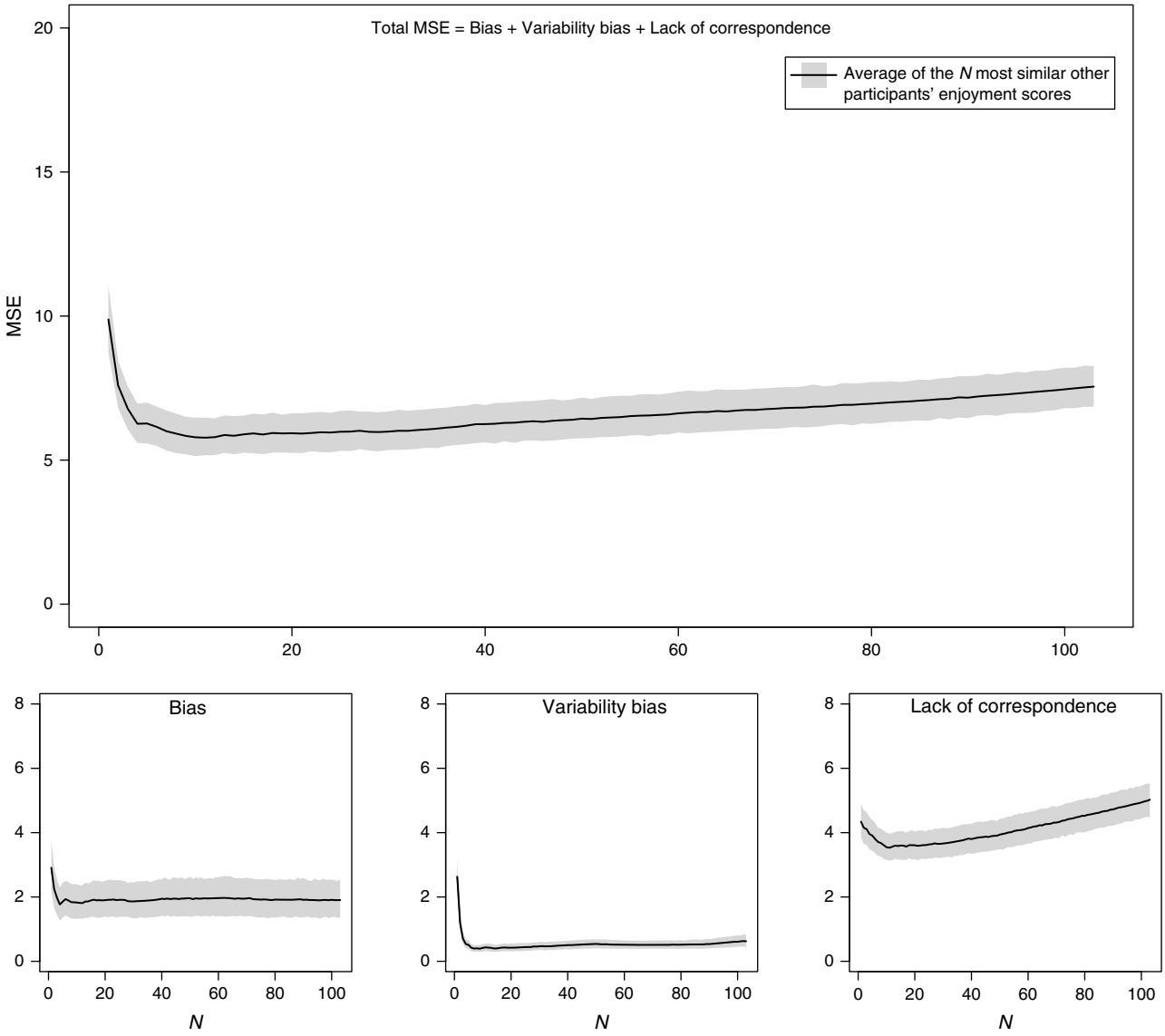
As in our theoretical model, we defined the taste similarity between two participants as the correlation between their respective enjoyment scores. As described above, the 22 stimuli consisted of two matched sets of 11 musical pieces each, selected to maximize resemblance across the two sets and diversity in musical styles within each set. This design feature allowed us to conduct all analyses involving taste similarity in out-of-sample analyses: the enjoyment scores for the first set of 11 musical pieces were used to estimate taste similarity, and the enjoyment scores for the second set were used to evaluate predictive accuracy.

Again, we used a bootstrap method to calculate crowd judgments and their accuracy. First, we sampled participants with replacement from the original data set to construct a bootstrap sample. Subsequently, we matched each participant in the bootstrap sample with those N other participants in the original data set whose enjoyment scores for the first set of music yielded the highest pairwise correlations with the target participant's scores, and we calculated the average enjoyment score across these "N most similar" participants. As before, we then computed how accurately these crowd judgments predicted each participant's enjoyment score for each piece. Finally, we repeated this procedure for 2,000 bootstrap samples and for N ranging from 1 to 103. The main panel in Figure 4 shows the average MSE across the 2,000 bootstrap samples as well as the 2.5th and the 97.5th percentiles of the distribution of bootstrap estimates, and the small panels show the same estimates for its three components. Again, several key crowd judgments are summarized in Table 1.

Comparing Figure 4 with Figure 3 is instructive. Both figures clearly demonstrate a pronounced initial effect of averaging on accuracy with diminishing yields as crowd size increases. Yet there are subtle and important differences between them. First, consider the overall levels of MSE. The similar crowds in Figure 4 exhibited consistently lower MSEs than the randomly selected crowds in Figure 3. For $N = 1$, for example, the enjoyment scores of the most similar participant (MSE = 9.9) were substantially more accurate than those of a randomly chosen participant (MSE = 14.9; see Table 1). Notably, however, we again found a wisdom-of-crowds effect: the enjoyment scores of the most similar participant were outperformed by the crowd of all participants (MSE = 7.5; see Table 1). Second, the MSE in Figure 4 decreases with crowd size up to about $N = 10$ and then increases, resulting in a U-shaped curve. Crowd judgments that included the most similar participants thus produced sizable accuracy gains, and small crowds of similar participants offered the most accurate judgments overall (see Table 1). These gains diminished when more participants (who are estimated to be less similar to the decision maker by construction) were added to the crowd.

Finally, the MSE decomposition reveals that the benefits of similarity can be attributed mainly to improvements in the judgments' linear correspondence (see also Table 1). Recall that in Figure 3, the linear correspondence for randomly selected crowds was barely affected by averaging. The improvements in linear correspondence for similar crowds in Figure 4 suggest that such crowds are informative not only in that they reduce excessive variability but also in a more fundamental sense. Similar crowds can help participants improve their ability to predict their own preferential ranking of the stimuli. Taken together, these findings confirm the hypothesis that taste similarity enhances the wisdom of crowds in matters of taste.

Figure 4. Study 1: Accuracy of Similar Crowds in Predicting a Target Participant's Enjoyment Scores



Note. Shown are the bootstrap estimates of the mean (line) and 95% confidence intervals (shaded areas).

Taste Discrimination: Familiarity. Next we consider the effects of familiarity. Our model asserts that the benefits of the wisdom of crowds should be particularly pronounced when the decision maker's taste discrimination is low. Presumably, people make less discriminative judgments when evaluating music they are less familiar with (e.g., folk music from a remote culture) than when evaluating music they are more familiar with (e.g., local pop music). Crowds should then be "wiser" for unfamiliar than for familiar music.

We first verified that taste discrimination was higher for familiar music. To this end, we calculated how well each participant's enjoyment scores for the musical pieces in the first set predicted his or her scores for the corresponding pieces in the second set (MSE = 5.7, SD = 1.7, 95% CI between 5.0 and 6.4). If taste discrimination is higher for familiar than for unfamiliar music, then this within-participant squared error should decrease in familiarity scores. We estimated a linear mixed model based on all 104 participants' enjoyment scores for the 22 pieces (N = 2,288), with their familiarity scores as the predictor variable. The results supported our hypothesis: on average, a one-point increase in a participant's familiarity score for a particular piece of music decreased the squared error for the piece by approximately one-third (b = −0.32, SE = 0.13, 95% CI between −0.64 and −0.04).[12] Across the 10-point scale of familiarity scores, the analysis suggests that the expected within-participant squared error declined from 6.9 for highly unfamiliar music (i.e., rated 1) to 4.0 for highly familiar music (i.e., music rated 10).

These analyses confirm the hypothesized relation between familiarity and taste discrimination. According to our model, this should have implications for the efficacy of the wisdom of crowds. Specifically, crowd judgments should be more useful in predicting one's preferences for unfamiliar than for familiar music. We tested this hypothesis in a linear mixed model with participants' familiarity scores as the response variable and crowd judgments based on all other participants' enjoyment scores as the predictor variable ($N = 2{,}288$). It found clear support in the data: on average, a one-point increase in a participant's familiarity score for a particular piece of music increased the crowd judgment's squared error by approximately two-thirds ($b = 0.70$, SE = 0.10, 95% CI between 0.52 and 0.91). Across the full range of familiarity scores, this translates to sizable differences: the analysis suggests that the squared error of the crowd judgment ranged from 4.9 for the most unfamiliar music to 11.2 for the most familiar. Overall, we find strong support for the hypothesis that crowd judgments are particularly beneficial in the context of (less certain) preferences for less familiar music.

3.1.3. Discussion. Study 1 demonstrates that other people's judgments about their personal musical preferences can be valuable in predicting one's own musical tastes. Crowd judgments for musical pieces (i.e., averages of others' enjoyment scores) were predictive of target participants' enjoyment scores. In line with our theoretical framework, the effects were strongest for groups of participants who shared a target participant's tastes and when the music was unfamiliar for the target participant. Our theoretical framework also makes the novel prediction that crowd judgments can be useful even when based on randomly chosen others, especially in predicting tastes for unfamiliar music. This prediction, too, was borne out by the data. In summary, our findings suggest that although different people have different tastes, crowds can be wise in matters of taste.

3.2. Study 2: Short Films

Study 2 replicated our principal findings in a different domain, that of short films. It also introduced an important methodological change compared with the first study. Participants in Study 2 were asked to forecast their future enjoyment of short films based on some limited information provided to them about a week before they actually watched the films. The study thus featured two types of judgments: "enjoyment forecasts" based on limited information and "enjoyment ratings" based on full information. Comparisons between the two types of judgments allowed us to assess whether participants were more or less accurate than the crowd in predicting their own tastes. This design feature also allowed us to directly assess the effect of additional information on taste discrimination and to examine crowd wisdom in a "symmetric" setting in which the decision maker and the crowd alike base their judgments on limited information about the stimuli. Finally, we took advantage of Study 2's design to investigate yet another aspect of behavioral importance, and we surveyed our participants' intuitions regarding the predictive accuracy of crowd judgments based on limited information. This survey provided insights into people's awareness of the potential benefits of aggregating others' opinions in matters of taste.

3.2.1. Method

Participants. Sixty-six undergraduate students participated in the study. Four of them were excluded from all the analyses because they failed to answer some or all of the questions; the final sample included 24 males and 38 females. Participants received partial course credit or the equivalent of $7 for the two sessions.

Materials. We obtained 21 short films, each less than eight minutes in length, from the websites FILMSshort (http://filmsshort.com, accessed December 2011) and Online Short Films (http://onlineshortfilms.net, accessed December 2011). Participants in a pilot study rated how much they enjoyed each of the 21 films on 100-point scales. Seven films that produced mean ratings close to the center of the scale and exhibited considerable variability at the participant level were included in the main study. With these inclusion criteria, we aimed to eliminate (or at least minimize) quality differences between the films and foster the wide range of individual differences in tastes required to test our hypotheses.

Procedure. The procedure included two sessions. In the first session, participants were presented with 10-second excerpts from each of the seven short films. After viewing each excerpt, participants were asked to indicate how much they thought they would enjoy the full-length version of the film. In the second session, conducted about one week later, participants watched the full-length version of each of the seven films and rated their enjoyment (immediately after viewing each of them). Participants were run individually in a computerized laboratory. They watched all the films on PCs equipped with headphones.

The details of the procedure were as follows. In the first session, participants were informed that the study would involve making judgments about short films and that there would be two sessions. They were further informed that in the first session, they would watch several brief excerpts taken from the short films and would later watch the full-length films in the second session. To familiarize the participants with the kinds of film used in the study, they were shown two other full-length short films (about five minutes each) at the beginning of the first session. These two films were selected to anchor the end-points of the enjoyment scale (i.e., one was among the highest-ranked and

the other among the lowest-ranked films in the pilot study). Participants were then shown the series of 10-second clips, one from each of the seven films, in a randomized order. All participants viewed the same clips. These were taken from the beginning of each of the full films and did not include identifying information such as the title or names. Each clip was accompanied by a short label describing its genre (e.g., comedy, animation). After viewing each clip, the participants were asked to predict how much they would enjoy watching the full-length short film a week later, on a scale that ranged from 0 ("I do not expect to enjoy the film at all") to 100 ("I expect to enjoy the film a lot"). Participants were also asked whether they had seen any of the films before.[13]

At the end of the first session, the participants were asked to provide demographic information. They were also asked to judge how accurate each of the following would be in predicting their own future enjoyment of the full versions of the films: (i) the judgments they had just made based on the limited information provided by the clips, (ii) a randomly selected other participant's judgments about his or her respective enjoyment based on the same limited information, and (iii) the average of all other participants' judgments of their respective enjoyment, again based on the same limited information.

In the second session that took place a week later, participants were shown the full-length versions of the seven short films in a randomized order. At the end of each film, they were asked to indicate how much they enjoyed the film on a scale that ranged from 0 ("I did not enjoy the film at all") to 100 ("I enjoyed the film a lot"). At the end of the session, the participants were thanked for their participation and paid or awarded their course credits. The study did not include any measures or conditions that are not reported.

3.2.2. Results. In analogy to Study 1, we treated participants' enjoyment ratings based on reliable information (elicited in the second session) as approximating their satisfaction values (i.e., in terms of our model, we assume $\sigma_e^2 \approx 0$ for these judgments). Mean enjoyment ratings for the seven films ranged from 37.5 to 64.0 (grand mean = 51.7). As in our first study, the ratings varied a great deal within and across participants. The median within-participant standard deviation of the ratings was 26.6, and the average pairwise correlation of enjoyment ratings between participants was fairly low (mean $r = 0.16$, SD = 0.39, 95% CI between 0.10 and 0.23). These results suggest that there were no universal norms for judging the films.

Taste Discrimination: Self-Predictions. Next, we tested the accuracy of our participants' enjoyment forecasts based on limited information (elicited in the first session) as predictors of their own enjoyment ratings. In other words, we used the judgments based on the excerpts to predict the judgments based on full information about the films. This allowed us to quantify the noise component of taste discrimination (see Section 2.3). As before, the MSE was our principal accuracy measure, and we calculated the MSE between each participant's enjoyment forecasts and ratings across the seven films.[14] The average of the participant-level MSE was 1,212 (see Table 2). Its square root, which translates the MSE back to the scale used to elicit the ratings, was 34.8, about a third of the scale. This confirms that we successfully created a setting in which taste discrimination was quite low—in other words, it was difficult for participants to forecast their preferences accurately. As noted before, low taste discrimination increases the potential benefits of the wisdom of crowds (see Section 2.3).

We also calculated two additional accuracy measures at the participant level that allowed us to decompose the MSE as laid out in Section 2.2 (see also Table 2). The average achievement correlation r_a between the participants' forecasts and their actual enjoyment ratings was relatively low, at 0.27 (SD = 0.41, 95% CI between 0.17 and 0.37), and on average, the corresponding component quantifying the lack of linear correspondence accounted for 44% of the MSE. The mean prediction error or bias showed that, on average, participants underestimated their enjoyment ratings by 5.7 points (SD = 14.3, 95% CI between −9.3 and −2.3), which accounted for 19% of the MSE. The remaining 36% of the MSE resulted from variability bias—that is, the participants' failure to regress their forecasts sufficiently to the mean, given the difficulty of making accurate forecasts (reflected in the low achievement correlations).

Crowd Wisdom: Self-Predictions and Randomly Sampled Participants. We now turn to assessing the wisdom of crowds in these data. Could our participants have made more accurate forecasts by taking others' opinions into account? To answer this question, we compare the MSE of participants' own enjoyment forecasts with the MSEs associated with various crowd judgments. These include both averages of random samples of enjoyment *ratings* made by other participants after watching the full-length films and averages of enjoyment *forecasts* made by participants after viewing the brief excerpts. According to our theoretical model, a participant's own forecasts should predict his or her enjoyment ratings more accurately than the forecasts of another participant chosen at random. At the same time, since making discriminative judgments was relatively difficult, we predict a wisdom-of-crowds effect: the average forecast of a (sufficiently large) random sample of other participants should be a more accurate predictor of a participant's actual enjoyment

Table 2. Accuracy of Participants' Own Enjoyment Forecasts and of Key Crowd Judgments Based on Either Enjoyment Forecasts or Enjoyment Ratings in Study 2

			95% CI		MSE decomposition		
Predictor	MSE	SD	Low	High	Bias	Variability bias	Correspondence
Self-forecasts	1,212	799	1,015	1,410	233	440	539
Crowd enjoyment forecasts							
One randomly selected other	1,629	917	1,404	1,872	326	720	583
Ten randomly selected others	965	545	821	1,110	231	149	585
One most similar other	1,329	652	1,169	1,502	362	442	524
Ten most similar others	818	467	701	936	232	96	491
All others	906	514	769	1,055	221	107	578
Crowd enjoyment ratings							
One randomly selected other	1,558	834	1,369	1,764	379	609	569
Ten randomly selected others	856	452	745	978	209	113	534
One most similar other	577	349	491	675	265	134	179
Ten most similar others	393	280	327	462	200	73	120
All others	778	413	674	883	190	86	503

Notes. All statistics were calculated using the bootstrap method described in Section 3.1. "Bias," "Variability bias," and "Correspondence" refer to the three components of the decomposition in Equation (2) in Section 2.2. Any discrepancies between the MSE and the sum of its three components reflect rounding errors.

than his or her own forecasts. Averaging the participants' enjoyment ratings based on the full-length films, instead of their forecasts based on the excerpts, should yield an even more accurate predictor. For a crowd that is not selected on the basis of its similarity to the target participant (such as one created by random sampling), however, the advantage of using ratings rather than forecasts should be limited.

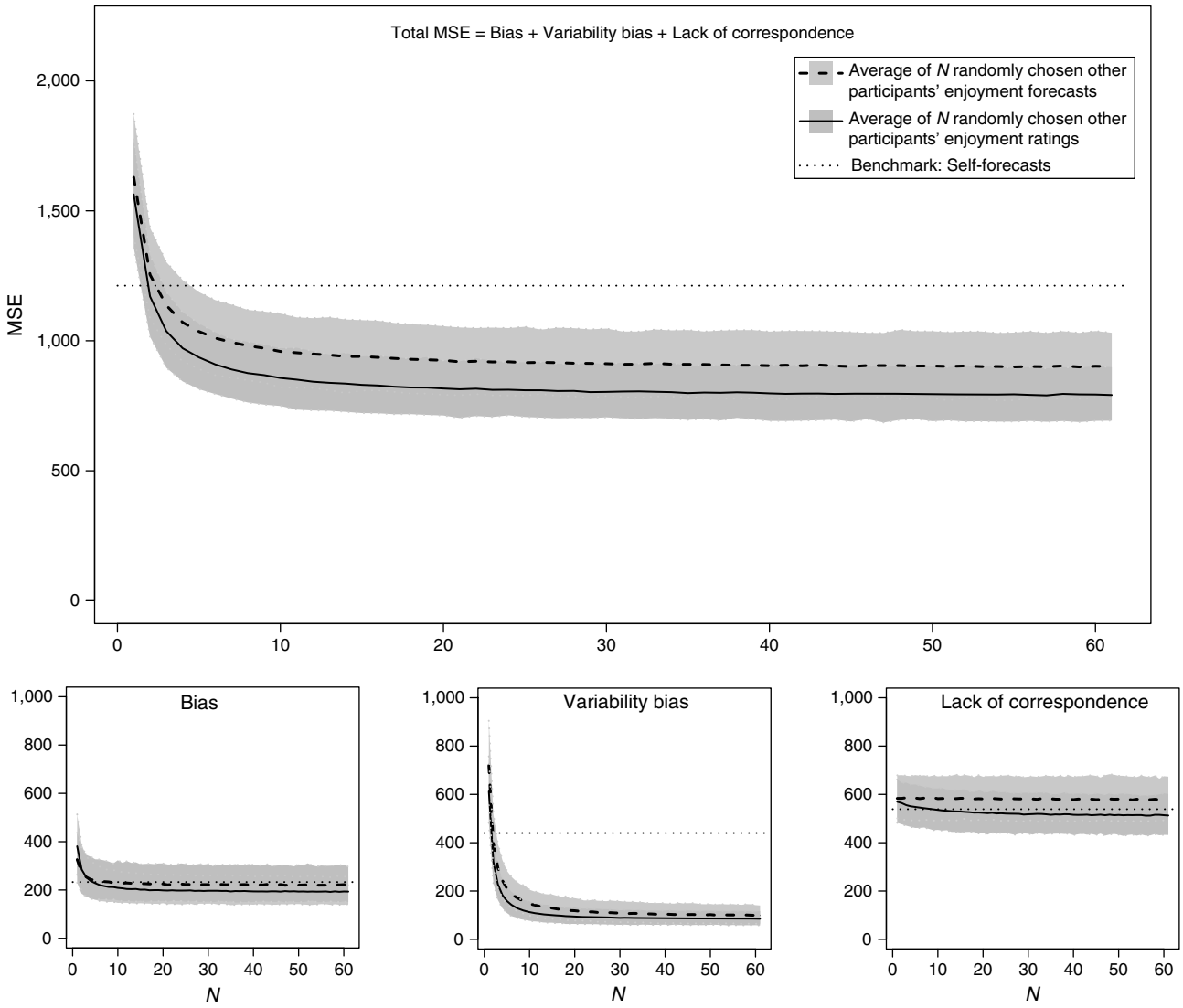
We calculated the accuracy of the crowd judgments with the same bootstrap method as in Study 1 (see Section 3.1). Several key crowd judgments are summarized in Table 2. In addition, the main panel in Figure 5 presents the average MSE of the different crowd judgments across 2,000 bootstrap samples as well as the 95% percentile confidence intervals, i.e., the 2.5th and the 97.5th percentiles of the distribution of bootstrap estimates. Wisdom-of-crowds effects are evident in films, as they were in music.

Consider first the "wisdom" of the crowd's enjoyment *forecasts*. Although taste discrimination was low, a participant's own forecasts were substantially more accurate (MSE = 1,212) than those of a randomly chosen participant (MSE = 1,629; see Table 2). But participants' forecasts of their own enjoyment were outperformed by crowd forecasts, e.g., the film-level averages of all other participants' forecasts (MSE = 906). Crowd judgments based on enjoyment *ratings* tended to be even more accurate. In particular, the enjoyment ratings of a single randomly chosen participant (MSE = 1,558) were less accurate than participants' own self-forecasts but more accurate than the forecasts of a randomly chosen participant. Again, averaging several participants' judgments produced accuracy gains; the MSE of the film-level averages of all other participants' enjoyment ratings was 778. In line with our model, this crowd judgment proved to be the most accurate predictor, but it was only moderately more accurate than the crowd's forecasts.

Finally, Figure 5 also replicates two other findings of Study 1. First, accuracy gains from combining judgments decreased rapidly as crowd size increased. As in the first study, and as predicted by our model (see Section 2.3), small crowds of 5–15 participants performed as well as much larger crowds. Second, decomposing the MSE reveals that the accuracy gains from averaging were largely due to reductions in variability bias. In other words, crowd judgments regressed to the mean, which is beneficial when taste discrimination is low. Crowd judgments also had the advantage of reducing bias. The linear correspondence between crowd judgments and participants' enjoyment ratings, by contrast, was approximately the same for crowds of any size. Crowd enjoyment ratings performed on par with the participants' own forecasts on this component of the MSE; crowd forecasts performed slightly worse. Reliance on crowd judgments of other participants sampled at random thus sometimes yielded a less accurate ranking of the stimuli in exchange for greatly reducing the average distance between the judgments and the criterion values (e.g., in the case of crowd judgments based on forecasts). We return to this point below.

Crowd Wisdom: Combining One's Own Forecasts with the Crowd's. Our model asserts that with symmetric information, people should rely more strongly on their own enjoyment forecasts than on other people's (see Section 2.4). We tested this hypothesis by comparing the accuracy of crowd judgments that either did or did

Figure 5. Study 2: Accuracy of Randomly Selected Crowds in Predicting a Target Participant's Enjoyment Ratings



Note. Shown are the bootstrap estimates of the means (lines) and 95% confidence intervals (shaded areas).

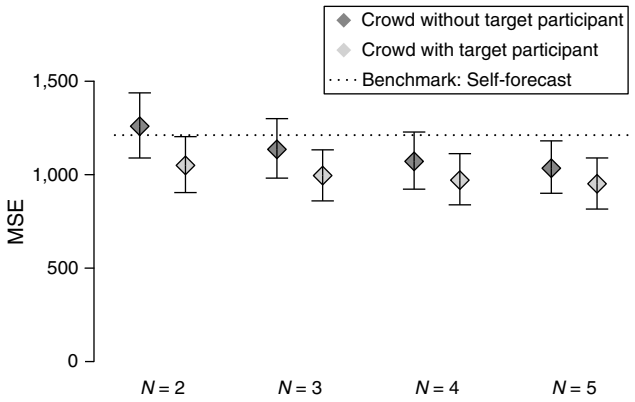
not include the target participant's own forecasts. (All crowd judgments considered thus far did not include the target participants' own forecasts.)

Again, we employed the bootstrap methodology described in Section 3.1 to estimate the accuracy of the various crowd judgments (including or excluding the target participant). The pertinent comparison pits the average enjoyment forecast of a crowd of N randomly sampled participants (without the target participant) against the average enjoyment forecast of a crowd of $N-1$ other participants sampled at random plus the target participant. Figure 6 shows the average MSEs for small crowds of sizes 2–5, based on 2,000 bootstrap samples and the corresponding 95% bootstrap confidence intervals.

Figure 6 shows that, as predicted by our model, crowd judgments that included the target participant's own forecasts predicted his or her enjoyment ratings more accurately than crowd judgments that did not include them. The effect was obtained for all crowd sizes considered. Crowd judgments that included a participant's own forecasts were also more accurate than these same forecasts on their own, providing yet another illustration of a wisdom-of-crowds effect. Finally, the comparative advantage of crowds that included the target participant's own forecast decreased with crowd size, simply because the relative impact of any single judgment on the crowd judgment decreases with N.[15]

Crowd Wisdom: Taste Similarity. We now discuss the effects of taste similarity. According to our model, the benefits of averaging should be greater for those who share similar tastes (see Section 2.3). As in Study 1, we calculated various crowd judgments that draw on the

Figure 6. Study 2: Accuracy of Enjoyment Forecasts of Randomly Selected Crowds in Predicting a Target Participant's Enjoyment Ratings (MSE)



Note. Shown are the bootstrap estimates of the means (diamonds) and 95% confidence intervals (bars).

judgments of participants selected for their similarity to a target participant.

Study 2 goes beyond the first study in allowing us to analyze the interplay between taste discrimination and taste similarity. In parallel to our results on crowds of randomly sampled participants, we computed averages of similar participants' enjoyment *ratings* (as in Study 1) and averages of similar participants' enjoyment *forecasts* (going beyond Study 1). The use of full-length short films in Study 2, however, required us to reduce the number of stimuli compared with the first study (because of their length), thereby reducing the information available for estimating similarity. As a result, it was not possible to reliably estimate similarity on subsets of the films. We thus resorted to estimating similarity by using the enjoyment ratings for the seven films, and assessing the accuracy of crowd judgments on the same seven films. With this exception, we employed the same procedure as in the first study: similarity was defined as the correlation between participants' enjoyment ratings (i.e., their judgments based on full information), and we used the bootstrap method described in Section 3.1 to compute the various crowd judgments and their accuracy.

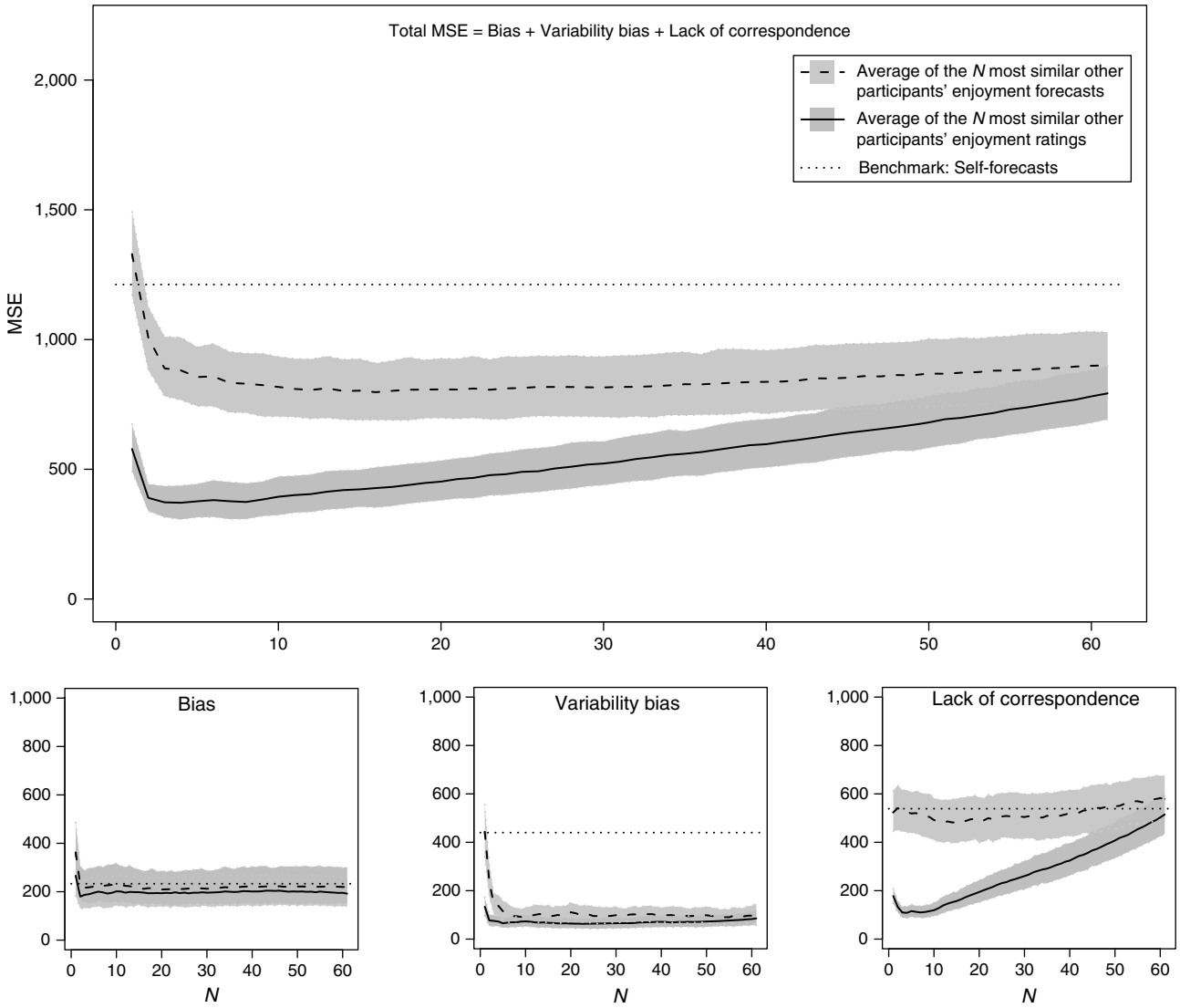
Figure 7 displays the results from this analysis, and estimates for key crowd judgments can be found in Table 2. Consider first the average enjoyment forecasts for crowds including participants selected on the basis of their similarity to a target participant. The average forecasts of such crowds showed sizable accuracy gains. These gains first diminished as crowd size increased and less similar participants were added to the crowd, and later turned to accuracy losses as similarity decreased further. This U-shaped relation between MSE and function of crowd size is predicted by our framework (and was also obtained in Study 1; see Figure 4). Next, consider the crowd's enjoyment ratings. Here, the U shape was even more pronounced. Moreover, Figure 7 reveals dramatic differences in the accuracy gains obtained from the forecasts of a similar crowd (based on brief excerpts) and from their ratings (based on complete films). As noted earlier, this was not the case for randomly sampled crowds, whose enjoyment ratings were only moderately more accurate than their enjoyment forecasts (see Figure 5). Taken together, this is precisely the pattern of results predicted by our theoretical model, which holds that increasing discriminability by providing more information should be particularly beneficial when taste similarity is high (see Section 2.3). While effect sizes should be interpreted with caution because of the in-sample nature of this analysis (i.e., they would likely be smaller if similarity were estimated out of sample), the observed interaction between firsthand experience with the stimuli and taste similarity provides compelling support for our model.

Finally, the three panels in the bottom of Figure 7 represent the three components of the MSE discussed in Section 2.2: bias, variability bias, and error resulting from a lack of linear correspondence. For the first two components, crowd judgments drawing on participants selected on the basis of their similarity to a target participant behaved much like other crowd judgments, yielding a small improvement in bias and a substantial improvement in variability bias (compare with Figures 3–5). Importantly, Figure 7 also reveals improvements in linear correspondence compared to participants' self-forecasts, even for the crowd judgments based on limited information (see also Table 2). This provides further evidence that similar crowds are uniquely "wise" in affording target participants gains in predicting their preferential ranking of the stimuli.

Crowd Wisdom: Lay Beliefs. Finally, were our participants aware of the potential for crowd wisdom in predicting matters of taste based on limited information? As noted, we surveyed our participants' intuitions about the accuracy of their own and others' enjoyment forecasts. The participants correctly expected their own enjoyment forecasts (M = 68.8 on a 100-point scale, SD = 19.1, 95% CI between 64.1 and 73.8) to be more accurate predictors of their enjoyment ratings than the forecasts of a randomly selected participant (M = 52.0, SD = 20.3, 95% CI between 46.9 and 56.8). At the same time, they incorrectly expected their own enjoyment forecasts to also be more accurate than the average of all participants' enjoyment forecasts (M = 57.0, SD = 20.1, 95% CI between 52.0 and 62.0). Our participants thus appeared largely unaware of the potential gains from relying on crowd judgments based on limited information.

3.2.3. Discussion. Study 2, similar to Study 1, showed that other people's judgments can be valuable in

Figure 7. Study 2: Accuracy of Similar Crowds in Predicting a Target Participant's Enjoyment Ratings



Note. Shown are the bootstrap estimates of the means (lines) and 95% confidence intervals (shaded areas).

predicting personal tastes. Study 2 also bolstered the evidence for the effect of taste similarity (between the individuals in the crowd and the decision maker) on the benefits of the wisdom of crowds. Furthermore, the design of Study 2 allowed us to evaluate the wisdom of crowds in a symmetric setting, in which the decision maker and the crowd based their judgments on equally limited information. This is particularly interesting since past studies have focused on asymmetric settings—that is, settings in which the decision maker relies on limited information, whereas the crowd can draw on more complete, firsthand information about the object of the recommendation (e.g., a movie, ski resort, or restaurant). In other words, in past studies, decision makers were forecasting their expected enjoyment, whereas individuals in the crowd reported their actual enjoyment. Study 2 allowed us to compare the accuracy of crowd judgments based on forecasts and those based on reports of actual enjoyment.

In line with our theoretical model, crowd judgments based on full information were more accurate than those based on limited information. Moreover, as predicted by our model, the benefits of additional information depended heavily on the taste similarity between the decision maker and the crowd. Under conditions of high taste similarity, the benefits of crowd wisdom were more pronounced in asymmetric settings (in which crowds have access to full information) than in symmetric settings (in which they relied on the same limited information as the decision makers). Under conditions of low similarity, crowd judgments based on full information had little advantage over those based on limited information.

4. General Discussion

In this article, we have investigated whether and when decision makers can draw on the "wisdom of crowds" to accurately predict their hedonic reactions and subjective experiences. In this domain, the efficacy of relying on other people's opinions cannot be taken for granted, since tastes differ from one person to another. Our findings suggest that crowds can nonetheless confer "wise" advice in matters of taste. In two laboratory studies, averages of other participants' judgments of taste could be leveraged to enhance decision makers' accuracy in predicting their enjoyment of musical pieces (Study 1) and short films (Study 2). Crowd judgments could benefit decision makers in asymmetric settings in which individuals in the crowd had access to reliable information (e.g., firsthand experience). Crowd judgments were useful even in symmetric settings in which the crowd relied on limited information just like the decision maker. These findings are remarkable since the participants in our studies did not predict a set of common criterion values (as in predicting factual matters) but their own personal criterion values (i.e., each participant predicted how much he or she would enjoy each stimulus).

Indeed, the theoretical model developed in this article emphasizes the importance of taste similarity in aggregating judgments of taste. Individuals with similar tastes share similar criterion values, and they can usually benefit more from one another's opinions than individuals with dissimilar tastes.[16] This prediction was confirmed in our empirical analyses. The role of taste similarity in judging tastes resembles the role of expertise in judging facts (Broomell and Budescu 2009, Budescu and Chen 2015, Davis-Stober et al. 2014). Yet our model also emphasizes the importance of taste diversity. A crowd of several individuals is "wisest" in judging matters of taste when the individuals' tastes resemble the decision maker's but are otherwise maximally diverse—that is, dissimilar from one another. This parallels recent findings on the benefits of diversity in judging facts (Davis-Stober et al. 2014), problem solving (Hong and Page 2004), and innovation economics (van den Bergh 2008). It explains why even crowd judgments based on randomly selected participants were useful for predicting a decision maker's tastes in our studies.

Our theoretical model also delineates the boundary conditions of such "crowd wisdom," highlighting the role of taste discrimination. An individual's ability to discriminate accurately and confidently among stimuli depends on factors such as his or her familiarity with the stimuli and the reliability of the information available. In our model, decision makers who make highly discriminative judgments stand to gain little from relying on crowd judgments; doing so could even affect them adversely. In parallel, crowds of people who make highly discriminative judgments tend to be particularly useful to decision makers, although this effect is moderated by the taste similarity between the decision maker and the people in the crowd. When similarity is low, our model predicts that crowd judgments based on limited information can be almost as useful as well-informed crowd judgments. Again, our empirical analyses confirmed these predictions. This explains why we were able to observe the wisdom of crowds in symmetric decision settings, going beyond the asymmetric settings studied in the past (e.g., Eggleston et al. 2015, Gilbert et al. 2009, Yaniv et al. 2011).

The remainder of this discussion is organized as follows. First, we discuss our model and findings in relation to theories of the wisdom of crowds and of advice taking in factual matters. Second, we connect our findings on lay intuitions to previous work on intuitions about averaging and the wisdom of crowds. Third, we consider the practical implications of the present research for business and management. We conclude by situating our work within the broader context of research on judging subjective experiences.

4.1. Crowd Wisdom in Tastes and Facts

Throughout this article, we have pointed to various conceptual resemblances and differences between the wisdom of crowds in matters of taste and in factual matters. We now discuss our findings in relation to several key results in existing treatments of crowd wisdom in matters of fact. Specifically, we examine the role of the information structure of the stimuli and of the social environment for judgments of taste, and we highlight parallels to the literature on small, "smart" crowds.

Consider first the information structure of the stimuli. Our model of judgments focuses on allowing criterion values to differ across individuals and forgoes incorporating the informational structure of the environment in favor of preserving parsimony. In an insightful analysis based on Brunswik's lens model, Broomell and Budescu (2009) argue that when different people base their judgments on the same informational cues, their judgments will often be markedly correlated (see also Endnote 5). Their analysis is highly general and may be applied to judgment problems involving tastes.[17] In particular, Broomell and Budescu (2009) show that when cues are highly correlated with one another, high interjudge correlations are inevitable even when different individuals weigh the cues differently. In our studies, we observed substantial heterogeneity in participants' tastes and that participants' judgments did not, on average, correlate strongly. In light of Broomell and Budescu's analysis, this suggests that the intercue correlations in our studies were at most moderate. Applying their lens model analysis to our setting further suggests that people who share

similar tastes (i.e., similar criterion values) may also attend to similar cues. We did not attempt to quantify or measure the cues that characterize the stimuli in our studies (music and films), so our data cannot speak to this conjecture. We believe, however, that the relations between stimulus-specific informational cues and individual-specific criterion values in judgment problems involving matters of taste are of considerable theoretical interest, and that they merit further research.

Next, consider the information structure of the social environment. Our model suggests that decision makers who look for advice from others' judgments of taste face a formidable inference problem: to determine what weight to place on advisors' opinions, decision makers have to assess each advisor's taste similarity and taste discrimination. In other words, a decision maker needs to assess not only an advisor's reliability (as in matters of fact) but also whether the advisor's criterion values are relevant to the decision maker's own tastes. This can be difficult, especially when the number of judgments available for each advisor is limited (cf. Analytis et al. 2015). Indeed, repeated interactions with potential advisors might be required to generate the rich information the decision maker needs to reliably assess the advisors' taste similarity and discrimination. Absent such rich information, decision makers may discount others' opinions (see Harvey and Fischer 1997, Soll and Larrick 2009, Yaniv 2004, Yaniv and Kleinberger 2000 for related discussions on the use of advice on matters of fact).

Finally, our findings suggest that a fairly small number of opinions can be sufficient to reap the benefits of the wisdom of crowds in matters of taste. In our laboratory data, the marginal benefits of additional opinions quickly diminished, and virtually all of the improvement was realized with as few as 10–15 advisory opinions. Our theoretical model holds that in judgment problems involving matters of taste, high-performing small crowds are characterized by a favorable tradeoff between taste similarity and diversity. This echoes the findings on the wisdom of small crowds in the literature on judgments in matters of fact (Budescu and Chen 2015, Davis-Stober et al. 2014, Goldstein et al. 2014, Hogarth 1978, Jose et al. 2014, Mannes et al. 2014, Yaniv and Milyavsky 2007).

4.2. Lay Intuitions About Aggregating Judgments of Taste

Recall that we also surveyed our participants' opinions on the benefit of aggregating judgments of taste in symmetric settings (Study 2). This survey provides some insight into laypeople's beliefs about the potential benefits of using the wisdom of crowds in domains involving personal preference. When queried about the usefulness of others' judgments based on the same limited information that they themselves had access to, our participants appeared largely unaware of the potential for crowd wisdom. Yet the popularity of Internet resources publishing crowdsourced reviews such as Yelp and TripAdvisor shows that in asymmetric settings, people often gladly take the judgments of better-informed advisors into account.[18] Whether people are willing to take others' opinions into account in matters of taste thus appears to depend on the presence or the absence of an information asymmetry: the divergence between our survey results and Yelp's popularity suggests that people attribute the usefulness of others' reviews largely to the additional information therein (usually obtained from personal experience), which is lacking in symmetric settings. It also points to a "misappreciation of the averaging principle" (Larrick and Soll 2006)—the lack of an intuitive understanding of the benefits of judgment aggregation, a well-documented phenomenon in the realm of factual judgments (see also Mannes 2009, Soll 1999, Soll and Larrick 2009). In the present context, the subjective nature of tastes may render the benefits of averaging even less intuitive, since people may view tastes as personal and varied. Consequently, they may assume that judgments of taste made by strangers with limited information (and combinations thereof) should have little bearing on their own hedonic reactions, and that their own impressions should supersede others' opinions for such judgments.

By contrast, people have more accurate intuitions about the role of taste similarity. Previous research has shown that decision makers rely on perceived similarity when considering others people's opinions in matters of taste (Gino et al. 2009, Yaniv et al. 2011). In Yaniv et al. (2011), for example, participants heeded the counsel of others who had displayed similar preferences on previous occasions (e.g., in choosing what music to listen to, participants tended to follow the advice of a person who had previously expressed similar musical tastes). Future research could target people's intuitions about the relations among key concepts in our analysis, such as diversity, discrimination, and similarity, and their joint effect on the usefulness of averaging.

4.3. Business and Managerial Implications

Understanding the nature of recommendations in matters of taste and their influence on decision makers is of considerable practical importance. Online retailers such as Amazon and Netflix use recommender systems to predict their customers' preferences and adjust their website contents accordingly. Our results provide psychological foundations for understanding algorithms used in recommenders based on analyses of similarities among consumers, such as collaborative filtering (cf. Ansari et al. 2000, Koren and Bell 2011). Our results also imply that to fully capitalize on the wisdom of crowds, recommender systems should strive to

simultaneously optimize crowd diversity. Finally, they suggest that recommender systems could also operate profitably in symmetric settings in which crowds and decision makers alike have only preliminary, limited information.

Other professionals in the business of predicting preferences face similar challenges. Consider designers, writers, and others in the creative industries, who rely on their personal tastes, feelings, and artistic insights to compose and shape their creations. These professionals need to predict their own preferences as well as the preferences of their potential customers. Our results on judgments of taste based on limited information provide support for piloting, suggesting that opinions based on drafts, samples, or sketches can help these professionals make more accurate taste predictions. Interestingly, drawing on the wisdom of crowds might help them predict their own tastes in addition to those of their customers.

More generally, our findings may shed new light on the old adage, "De gustibus non est disputandum" (In matters of taste, there can be no disputes). While different people are entitled to different tastes, our results imply that ignoring others' opinions in matters of taste may come at a price. This insight, we believe, is particularly relevant to "soft" decision problems (i.e., problems that lack well-defined solution criteria) in organizational settings. In these problems, the involvement of multiple individuals can frequently yield multiple points of view. Our results suggest that there may be wisdom in combining such divergent perspectives. Thus, even when the final decision is made by the individual with the highest standing in the organizational hierarchy, taking different perspectives into account may serve the organization well.

4.4. Concluding Remarks

Our approach draws on the notions of "preference processing" (Harrison and March 1984, March 1978) and "affective forecasting" (Gilbert 2006, Wilson and Gilbert 2005), as well as interpretations of utility theory as a model of hedonic experiences (Kahneman and Thaler 2006, Kahneman et al. 1997). In these research traditions, a decision maker's judgments of his or her subjective experiences play a central role in the descriptive study of choice and in devising prescriptive recommendations for improving choice. The present research investigated whether and when other people's judgments of their respective subjective experiences can also play a role in (good) decision making. Our results suggest they can: under the conditions outlined in this article, decision makers tapping into the wisdom of crowds can improve their predictions of their own subjective experiences and thereby make better choices.

Acknowledgments

The authors thank Robin Hogarth and Ed Vul, who both had an important influence on this work. They are also grateful to David Budescu, Shlomi Sher, George Wu, and Mike Yeomans for their comments and suggestions, to Maxim Milyavsky for his help with Study 1, and to Naomi Goldblum for her editorial assistance.

Appendix

We provide derivations of the results in Section 2.3, in the order that they appear in the text.

Consider first Equation (3). Under the assumptions in Section 2.3, the expected squared error of the crowd judgment can be computed from the basic properties of the expectation and the variance operators:

$$\begin{aligned}
&\mathrm{E}[(C - v_D)^2] \\
&\quad = \mathrm{E}[C^2 + v_D^2 - 2Cv_D] = \mathrm{E}[C^2] + \mathrm{E}[v_D^2] - 2\mathrm{E}[Cv_D] \\
&\quad = \mathrm{E}[C]^2 + \mathrm{Var}[C] + \mathrm{E}[v_D]^2 + \mathrm{Var}[v_D] \\
&\qquad - 2(\mathrm{E}[C]\mathrm{E}[v_D] + \mathrm{Cov}[C, v_D]) \\
&\quad = (\mathrm{E}[C] - \mathrm{E}[v_D])^2 + \mathrm{Var}[C] + \mathrm{Var}[v_D] - 2\mathrm{Cov}[C, v_D] \\
&\quad = \mathrm{Var}[\Sigma 1/N x_i] + \mathrm{Var}[v_D] - 2\mathrm{Cov}[\Sigma 1/N x_i, v_D] \\
&\quad = \mathrm{Var}[\Sigma 1/N(v_i + e_i)] + \mathrm{Var}[v_D] - 2\mathrm{Cov}[\Sigma 1/N(v_i + e_i), v_D] \\
&\quad = -2\Sigma 1/N\sigma_{vi,vD} + 1/N^2\Sigma\Sigma(\sigma_{ei,ej} + \sigma_{vi,vj}) + \sigma_{vD}^2 \\
&\quad = -2/N\Sigma\rho_{vi,vD}\sigma_{vi}\sigma_{vD} + 1/N^2\Sigma\Sigma\rho_{vi,vj}\sigma_{vi}\sigma_{vj} \\
&\qquad + 1/N^2\Sigma\Sigma\rho_{ei,ej}\sigma_{ei}\sigma_{ej} + \sigma_{vD}^2.
\end{aligned}$$

The equality between the third and the fourth line relies on the assumption that $E[v_i]$ is the same for all i.

The expression for the optimal weights in Section 2.4 is obtained by minimizing the following expression for the expected squared error of a crowd judgment with general weights:

$$\min_{wi} E[(\Sigma w_i x_i - v_D)^2] = \min_{wi} E[(\Sigma w_i x_i)^2 - 2(\Sigma w_i x_i)v_D + v_D^2].$$

Partial derivatives with respect to w_i yield the following first-order conditions for all i:

$$\begin{aligned}
&E[2(\Sigma w_j x_j)x_i - 2x_i v_D] = 0 \\
&\quad \Leftrightarrow E[2(\Sigma w_j x_j)x_i] - E[2x_i v_D] = 0 \\
&\quad \Leftrightarrow \Sigma w_j E[x_j x_i] - E[x_i v_D] = 0 \\
&\quad \Leftrightarrow \Sigma w_j \mathrm{Cov}[x_j, x_i] - \mathrm{Cov}[x_i, v_D] = 0,
\end{aligned}$$

where the final equivalence makes use of the assumption that $E[v_i] = 0$ for all i. These first-order conditions represent a system of linear equations. Let $\boldsymbol{\Sigma}_x$ denote the covariance matrix of the predictions, and $\boldsymbol{\sigma}_D = \{\sigma_{x1,vD}, \sigma_{x2,vD}, \ldots, \sigma_{xN,vD}\}^T = \{\sigma_{v1,vD}, \sigma_{v2,vD}, \ldots, \sigma_{vN,vD}\}^T$. Then, using bold type for vectors and matrices, the system can be written as

$$\boldsymbol{\Sigma}_x \mathbf{w} = \boldsymbol{\sigma}_D.$$

Now $\boldsymbol{\Sigma}_x$ can be factorized as

$$\boldsymbol{\Sigma}_x = \mathrm{Diag}(\sigma_{x1}, \sigma_{x2}, \ldots, \sigma_{xN})\mathbf{R}\mathrm{Diag}(\sigma_{x1}, \sigma_{x2}, \ldots, \sigma_{xN}),$$

where $\mathbf{R}$ is the correlation matrix of the predictions, whose qrth entry captures the correlation between the taste predictions of any two individuals q and r. The inverse of the

covariance matrix of the predictions can be computed from this factorization as

$$\mathbf{\Sigma}_x^{-1} = \mathrm{Diag}(\sigma_{x1}, \sigma_{x2}, \ldots, \sigma_{xN})^{-1}\mathbf{R}^{-1}\mathrm{Diag}(\sigma_{x1}, \sigma_{x2}, \ldots, \sigma_{xN})^{-1}.$$

It follows that

$$\mathbf{w} = \mathrm{Diag}(\sigma_{x1}^{-1}, \sigma_{x2}^{-1}, \ldots, \sigma_{xN}^{-1})\mathbf{R}^{-1}\mathrm{Diag}(\sigma_{x1}^{-1}, \sigma_{x2}^{-1}, \ldots, \sigma_{xN}^{-1})\boldsymbol{\sigma}_D,$$

and the elements of the inverse of the correlation matrix can be computed by cofactors as

$$R_{qr}^{-1} = (-1)^{q+r}|\mathbf{R}(\mathbf{q}, \mathbf{r})|/|\mathbf{R}|,$$

where $\mathbf{R}(\mathbf{q}, \mathbf{r})$ is the correlation matrix $\mathbf{R}$ with the qth row and the rth column removed, and the vertical bars denote the determinant. The elements of $\mathbf{\Sigma}_x^{-1}$ are then given by

$$\Sigma_{xqr}^{-1} = R_{qr}^{-1}(\sigma_{xq}\sigma_{xr})^{-1},$$

and for the weights, this implies

$$\begin{aligned} w_i &= \Sigma_j R_{ij}^{-1}(\sigma_{xi}\sigma_{xj})^{-1}\sigma_{vj,vD} \\ &= R_{ii}^{-1}(\sigma_{xi})^{-2}\sigma_{vi,vD} + \Sigma_{j\neq i}R_{ij}^{-1}(\sigma_{xi}\sigma_{xj})^{-1}\sigma_{vj,vD}. \end{aligned}$$

To obtain the expression for the weights in Section 2.4, let $a = R_{ii}^{-1}(\sigma_{xi})^{-2}\sigma_{vi}^2\sigma_{vD}^2$ and $b = \Sigma_{j\neq i}R_{ij}^{-1}(\sigma_{xi}\sigma_{xj})^{-1}\sigma_{vj,vD}$.

Finally, the result for the crowd size described in Section 2.3 is obtained by simply evaluating the expression for the accuracy of the crowd prediction $E[(C - v_D)^2]$ for $i = 1, \ldots, N$ and for $i = 1, \ldots, N, N+1$ (see the derivation of Equation (3)) and comparing:

$$\begin{aligned} &-2/N\Sigma_N\sigma_{vi,vD} + 1/N^2\Sigma\Sigma_N(\sigma_{ei,ej} + \sigma_{vi,vj}) \\ &\geq -2/(N+1)\Sigma_{N+1}\sigma_{vi,vD} + 1/(N+1)^2\Sigma\Sigma_{N+1}(\sigma_{ei,ej} + \sigma_{vi,vj}) \\ &\Leftrightarrow 2/(N+1)\Sigma_{N+1}\sigma_{vi,vD} - 2/N\Sigma_N\sigma_{vi,vD} \\ &\geq 1/(N+1)^2\Sigma\Sigma_{N+1}(\sigma_{ei,ej} + \sigma_{vi,vj}) - 1/N^2\Sigma\Sigma_N(\sigma_{ei,ej} + \sigma_{vi,vj}) \\ &\Leftrightarrow 2[1/(N+1) - 1/N]\Sigma_N\sigma_{vi,vD} + 2/(N+1)\sigma_{vN+1,vD} \\ &\geq 1/(N+1)^2\Sigma\Sigma_{N+1}(\sigma_{ei,ej} + \sigma_{vi,vj}) - 1/N^2\Sigma\Sigma_N(\sigma_{ei,ej} + \sigma_{vi,vj}) \\ &\Leftrightarrow 2/(N+1)[\sigma_{vN+1,vD} - 1/N\Sigma_N\sigma_{vi,vD}] \\ &\geq 1/(N+1)^2\Sigma\Sigma_{N+1}(\sigma_{ei,ej} + \sigma_{vi,vj}) - 1/N^2\Sigma\Sigma_N(\sigma_{ei,ej} + \sigma_{vi,vj}). \end{aligned}$$

Endnotes

[1] Asymmetric settings in which at least one advisor has full information are less interesting in judgments of facts because when all individuals predict the same criterion value, knowledge of the true value is in principle sufficient to "solve" the judgment problem. Research on aggregation and the wisdom of crowds in factual matters consequently focuses on (more) symmetric settings.

[2] More general linear combinations can nonetheless be of interest, particularly since individuals' own judgments may be more predictive of their own tastes than others' judgments, and may thus play a special role. We return to this point below.

[3] Similar arguments have been made in favor of the MSE as an accuracy measure for predictions of facts (e.g., Davis-Stober et al. 2014, Gigone and Hastie 1997). The justifications for using the MSE are just as relevant when criterion values are personal (individual-specific) rather than objective.

[4] The simple model we propose is both more general and more restrictive than contemporary models of factual judgments such as the one proposed by Davis-Stober et al. (2014). By allowing satisfaction values to differ across individuals, our model is more general. At the same time, our modeling assumptions impose additional mathematical structure; in this sense, our model is more restrictive.

[5] A third, related concept is interjudge correlation, i.e., the correlation between sets of judgments. Expert judgments are often highly correlated in matters of fact (e.g., Morris 1986, Winkler 1981). Broomell and Budescu (2009) offer a detailed, insightful analysis of interjudge correlations based on Brunswik's lens model (e.g., Hammond and Stewart 2001). The lens model affords Broomell and Budescu (2009) a more fine-grained view of the relations between judgments, criteria, and the environment. To model judgments of taste, we focus on allowing criteria to differ and forgo modeling the information structure of the environment. We explore the possibility of augmenting our framework with a more detailed model of the environment in the General Discussion.

[6] The accuracy of decision makers' *own* judgments evidently depends on σ_{eD} (a term not in Equation (3) when the crowd does not include the decision maker). Decision makers whose taste discrimination is low can thus benefit greatly from the wisdom of crowds, both because crowd judgments are more useful and because their own judgments are more fallible when their preferences are uncertain.

[7] Assuming that $\rho_{vi,vj} = 1$ ensures that the iso-MSE lines in Figure 2 do not break off similar to those in Figure 1. This is desirable since the constraints that taste diversity imposes on taste similarity are not the focus of this illustration. Equation (3) shows that high values of $\rho_{vi,vj}$ also make the trade-off between similarity and discrimination more favorable toward discrimination, but the overall pattern in Figure 2 is robust to changes in this parameter (i.e., for lower values of $\rho_{vi,vj}$, the curves in Figure 2 are somewhat steeper but otherwise unchanged).

[8] We also analyzed standardized scores (i.e., subtracting participant means from the scores and dividing them by participant standard deviations). This eliminates bias (by construction); all other results remained qualitatively unchanged (not reported).

[9] Throughout, we report bootstrap percentile CIs (cf. Davison and Hinkley 1997, Chap. 5).

[10] We used only the second set of musical pieces to ensure comparability with other analyses reported below that utilize the first set to estimate taste similarity. Analyses of all 22 pieces yield similar results (not reported).

[11] Those bootstrap estimates that involve resampling small crowds of participants exhibit minor sampling variability. We also note that combining scores to create crowd judgments creates dependencies in the data, which could potentially impair the accuracy of the bootstrap method (cf. Davison and Hinkley 1997, Chap. 8). This creates difficulties for other methods of analyzing the data, too.

[12] For regression coefficients, too, we report bootstrap percentile CIs. Reported estimates are based on mixed-effect models with random effects for both participants and musical pieces; other specifications yield very similar results.

[13] Only one participant indicated having seen any (two) of the films before. We did not exclude this participant from our analysis; his inclusion did not affect the overall pattern of results.

[14] In both studies, qualitatively similar results were obtained in terms of mean absolute deviations rather than MSEs (not reported).

[15] Our theoretical results in Section 2.4 suggest that increasing the weight on the target participant's own forecast should further increase predictive accuracy. In an unreported analysis, we used numerical methods to compute the optimal MSE-minimizing weights for small crowds consisting of the target participant as well as two other participants sampled at random. In line with our model, the optimal weight on the target participant's own self-forecast was estimated to be approximately twice as large as the weights on the randomly sampled participants.

[16] Our analysis may, in principle, also be applied to prediction problems involving facts that share this mathematical structure. For example, different analysts may predict the GDP growth rates for different U.S. states. In this context, our findings identify the conditions under which it can be beneficial to average growth rate predictions for different states to predict, say, California's growth rate. Moreover, several forecasts of each quantity of interest may be available in such factual prediction problems (while this is probably not as common in taste prediction). This opens the door to other interesting comparisons. For example, the accuracy of an average of growth rate forecasts for different states could be compared with that of the average of multiple forecasts of California's growth rate. Our model suggests that while the latter benefits from (maximal) similarity, the former may perform surprisingly well if it can capture the benefits of diversity (and if the growth rates are sufficiently correlated).

[17] In particular, those results in Broomell and Budescu (2009) that are not based on assumptions about cue reliabilities apply to matters of taste as well as matters of fact. Those results that do rely on assumptions about cue reliabilities would be more difficult to apply to judgments on matters of taste because cue reliabilities would have to be individual specific (since criterion values are individual specific).

[18] Crowdsourced reviews often combine a factual and a taste component. For example, a restaurant review may cover both the particular spices used in a dish (largely a matter of taste) and their quality (largely a matter of fact). This likely contributes to Yelp's popularity. It also points to the more general issue that many judgments involve both fact and taste. Perhaps the simplest way to accommodate this in our model is via taste similarity (e.g., in judgment problems featuring substantial factual components, personal criteria will become more universally valid, and different people's tastes will appear more similar).

References

Analytis PP, Barkoczi D, Herzog SM (2015) You're special, but it doesn't matter if you're a greenhorn: Social recommender strategies for mere mortals. Noelle DC, Dale R, Warlaumont AS, Yoshimi J, Matlock T, Jennings CD, Maglio PP, eds. *Proc. 37th Annual Meeting Cognitive Sci. Soc.* (Cognitive Science Society, Austin, TX), 1799–1804.

Ansari A, Essegaier S, Kohli R (2000) Internet recommendation systems. *J. Marketing Res.* 37(3):363–375.

Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G (2000) The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych.: Appl.* 6(2):130–147.

Armstrong JS (2001) *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Springer, New York).

Bates JM, Granger CWJ (1969) The combination of forecasts. *Oper. Res. Quart.* 20(4):451–468.

Böckenholt U (2006) Thurstonian-based analyses: Past, present and future utilities. *Psychometrika* 71(4):615–629.

Brooks AW, Gino F, Schweitzer ME (2015) Smart people ask for (my) advice: Seeking advice boosts perceptions of competence. *Management Sci.* 61(6):1421–1435.

Broomell SB, Budescu DV (2009) Why are experts correlated? Decomposing correlations between judges. *Psychometrika* 74(3):531–553.

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Sci.* 61(2):267–280.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79–101.

Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Application* (Cambridge University Press, Cambridge, UK).

Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psychologist* 34(7):571–582.

Deaton A (2008) Income, health, and well-being around the world: Evidence from the Gallup World Poll. *J. Econom. Perspect.* 22(2):53–72.

Diener E, Diener C (1996) Most people are happy. *Psych. Sci.* 7(3): 181–185.

Einhorn HJ (1974) Expert judgment: Some necessary conditions and an example. *J. Appl. Psych.* 59(5):562–571.

Einhorn HJ, Hogarth RM, Klempner E (1977) Quality of group judgment. *Psych. Bull.* 84(1):158–172.

Eggleston CM, Wilson TD, Lee M, Gilbert DT (2015) Predicting what we will like: Asking a stranger can be as good as asking a friend. *Organ. Behav. Human Decision Processes* 128(May):1–10.

Galton F (1907) Vox populi. *Nature* 75(1949):450–451.

Gigone D, Hastie R (1997) Proper analysis of the accuracy of group judgments. *Psych. Bull.* 121(1):149–167.

Gilbert DT (2006) *Stumbling on Happiness* (Random House Publishing, New York).

Gilbert DT, Wilson TD (2007) Prospection: Experiencing the future. *Science* 317(5843):1351–1354.

Gilbert DT, Killingsworth MA, Eyre RN, Wilson TD (2009) The surprising power of neighborly advice. *Science* 323(5921):1617–1619.

Gino F, Moore DA (2007) Effects of task difficulty on use of advice. *J. Behavioral Decision Making* 20(1):21–35.

Gino F, Shang J, Croson R (2009) The impact of information from similar or different advisors on judgment. *Organ. Behav. Human Decision Processes* 108(2):287–302.

Goldstein DG, McAfee RP, Suri S (2014) The wisdom of smaller, smarter crowds. *Proc. 15th ACM Conf. Electronic Commerce (EC '14)* (ACM, New York), 471–488.

Hammond K, Stewart TR (2001) *The Essential Brunswik: Beginnings, Explications, Application* (Oxford University Press, London).

Harrison JR, March JG (1984) Decision making and postdecision surprises. *Admin. Sci. Quart.* 29(1):26–42.

Harvey N, Fischer I (1997) Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Human Decision Processes* 70(2):117–133.

Hertwig R (2012) Tapping into the wisdom of the crowd—With confidence. *Science* 336(6079):303–304.

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psych. Sci.* 20(2):231–237.

Hogarth RM (1978) A note on aggregating opinions. *Organ. Behav. Human Performance* 21(1):40–46.

Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci. USA* 101(46):16385–16389.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463–475.

Kahneman D, Snell J (1990) Predicting utility. Hogarth RM, ed. *Insights in Decision Making* (University of Chicago Press, Chicago), 295–310.

Kahneman D, Snell J (1992) Predicting a changing taste: Do people know what they will like? *J. Behavioral Decision Making* 5(3): 187–200.

Kahneman D, Thaler R (2006) Anomalies: Utility maximization and experienced utility. *J. Econom. Perspect.* 20(1):221–234.

Kahneman D, Wakker PP, Sarin R (1997) Back to Bentham? Explorations of experienced utility. *Quart. J. Econom.* 112(2):375–405.

Koren Y, Bell R (2011) Advances in collaborative filtering. Ricci F, Rokach L, Shapira B, eds. *Recommender Systems Handbook* (Springer, New York), 145–186.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1):111–127.

Larrick RP, Mannes AE, Soll JB (2012) The social psychology of the wisdom of crowds. Krueger JI, ed. *Frontiers of Social Psychology: Social Psychology and Decision Making* (Psychology Press, New York), 227–242.

Lee J, Yates JF (1992) How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psych. Bull.* 112(2):363–377.

Linville PW, Fischer GW (2004) From basketball to business: Expertise, perceived covariation and social judgment. Yzerbyt V, Judd V, Corneille O, eds. *The Psychology of Group Perception: Contributions to the Study of Homogeneity, Entitivity, and Essentialism* (Psychological Press, London), 179–202.

Loewenstein G, Schkade D (1999) Wouldn't it be nice? Predicting future feelings. Diener E, Schwartz N, Kahneman D, eds. *Well-Being: The Foundations of Hedonic Psychology* (Russell Sage Foundation, New York), 85–105.

Luce RD, Suppes P (1965) Preference, utility and subjective probability. Luce RD, Bush RR, Galanter E, eds. *Handbook of Mathematical Psychology* Vol. 3 (John Wiley & Sons, New York), 249–410.

Makridakis S, Hibon M (2000) The M3-Competition: Results, conclusions and implications. *Internat. J. Forecasting* 16(4):451–476.

Makridakis S, Winkler RL (1983) Averages of forecasts: Some empirical results. *Management Sci.* 29(9):987–996.

Mannes AE (2009) Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Sci.* 55(8):1267–1279.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

Manski CF (1977) The structure of random utility models. *Theory Decision* 8(3):229–254.

March JG (1978) Bounded rationality, ambiguity, and the engineering of choice. *Bell J. Econom.* 9(2):587–608.

Morris PA (1986) [Combining probability distributions: A critique and an annotated bibliography]: Comment. *Statist. Sci.* 1(1): 141–144.

Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Rev.* 116(12):2417–2424.

Roberts FS (1985) *Measurement Theory with Applications to Decision Making, Utility and the Social Sciences* (Cambridge University Press, Cambridge, UK).

Smith JE, Winkler RL (2006) The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Sci.* 52(3):311–322.

Soll JB (1999) Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psych.* 38(2):317–346.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *J. Experiment. Psych.: Learn., Memory Cognition* 35(3):780–805.

Stevenson B, Wolfers J (2008) Economic growth and subjective well-being: Reassessing the Easterlin paradox. *Brookings Papers Econom. Activity* 39(1):1–87.

Stewart TR (1990) A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Weather Forecasting* 5(4): 661–666.

Stewart TR (2001) Improving reliability of judgmental forecasts. Armstrong J, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Springer, New York), 81–106.

Surowiecki J (2005) *The Wisdom of Crowds* (Doubleday, New York).

Theil H (1966) *Applied Economic Forecasting* (North Holland, Amsterdam).

Thurstone LL (1927) A law of comparative judgment. *Psych. Rev.* 34(4):273–286.

Timmermann A (2006) Forecast combinations. Elliott G, Granger CWJ, Timmermann A, eds. *Handbook of Economic Forecasting*, Vol. 1 (North Holland, Amsterdam), 135–196.

van den Bergh JC (2008) Optimal diversity: Increasing returns versus recombinant innovation. *J. Econom. Behav. Organ.* 68(3–4): 565–580.

Wallsten TS, Diederich A (2001) Understanding pooled subjective probability estimates. *Math. Soc. Sci.* 41(1):1–18.

Weiss DJ, Shanteau J (2003) The vice of consensus and the virtue of consistency. Shanteau J, Johnson P, Smith C, eds. *Psychological Explorations of Competent Decision Making* (Cambridge University Press, Cambridge, UK), 226–240.

Wilson TD, Gilbert DT (2005) Affective forecasting: Knowing what to want. *Current Directions Psych. Sci.* 14(3):131–134.

Winkler RL (1981) Combining probability distributions from dependent information sources. *Management Sci.* 27(4):479–488.

Winkler RL, Makridakis S (1983) The combination of forecasts. *J. Roy. Statist. Soc. Ser. A* (*General*) 146(2):150–157.

Yaniv I (2004) The benefit of additional opinions. *Current Directions Psych. Sci.* 13(2):75–78.

Yaniv I, Kleinberger E (2000) Advice taking in decision making: Egocentric discounting and reputation formation. *Organ. Behav. Human Decision Processes* 83(2):260–281.

Yaniv I, Milyavsky M (2007) Using advice from multiple sources to revise and improve judgment. *Organ. Behav. Human Decision Processes* 103(1):104–120.

Yaniv I, Choshen-Hillel S, Milyavsky M (2011) Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organ. Behav. Human Decision Processes* 115(1):111–120.