Journal of Business Research 67 (2014) 1771–1779

Contents lists available at ScienceDirect

Journal of Business Research

Are two interviewers better than one?[☆]

Mario Fifić [a,b,*], Gerd Gigerenzer [b]

[a] Grand Valley State University, MI, USA
[b] Max Planck Institute for Human Development Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany

ARTICLE INFO

Article history:
Received 1 April 2013
Received in revised form 1 November 2013
Accepted 1 January 2014
Available online 21 March 2014

Keywords:
Interview validity
Best interviewer
Wisdom of the crowd
Less is more
Free riding

ABSTRACT

How many interviewers per job applicant are necessary for a company to achieve the highest hit rate? Are two better than one? Condorcet's Jury Theorem and the "wisdom of the crowd" suggest that more is better. Under quite general conditions this study shows, surprisingly, that two interviewers are on average not superior to the best interviewer. Adding further interviewers will also not increase the expected collective hit rate when interviewers are homogeneous (i.e., their hits are nested), only doing so when interviewers are heterogeneous (i.e., their hits are not nested). The current study shows how these results depend on the number of interviewers, their expertise, and the chance of free riding, and specify the conditions when "less is more". This analysis suggests that the best policy is to invest resources into improving the quality of the best interviewer rather than distribute these to improve the quality of many interviewers.

Introduction

When consulting firms hire candidates as business consultants, or university departments invite applicants for faculty positions, the final decision is often based on a series of interviews. How many interviewers should be used for each candidate to achieve the best results? At first glance, the answer seems to be: the more, the better. For instance, the Condorcet's Jury Theorem says that the probability of a correct decision between two options increases with the number of decision makers in the group, provided that the individual probabilities of a correct decision are all greater than chance (Condorcet, 1785). Galton's (1907) seminal work on the vox populi appears to suggest the same conclusion, as does Bernoulli's law of large numbers. Modern concepts such as swarm intelligence (Krause & Ruxton, 2002) have led to speculations that if a diverse group can outperform an expert, then even CEOs might be in less demand in the future (Surowiecki, 2004). Do these arguments apply to interviewers as well?

The research reported in this article was motivated by a period in which one of us advised a consulting firm on their recruitment process. The firm has some 10,000 applications per year from young aspirants for over 100 open positions. Its decision-making process was neither fast nor frugal. In a first round, all applicants were evaluated on the basis of their CVs, statements, and letters, and about 500 were selected. In a second round, these selected applicants were flown in, put up in the best five-star hotel in town, and grilled by three interviewers, after which about half of them were eliminated. In a third round, a few weeks later, the remaining applicants were flown in again, put up in elegant suites, and quizzed by three other interviewers. For the final choice, the interviewers met to vote; offers were made to those with the highest number of votes. Millions of dollars were spent on the direct and indirect costs of this process even though the firm had no systematic quality control and kept no electronic records until a few years ago.

Companies around the world depend on interviews as a tool for selecting the best candidates. The consulting firm above represents a typical (but not isolated) case in which the question about the best number of interviewers was never considered.

The validity of an interview is typically defined as how effective a certain method is in finding the best candidates. The validity can be quantified by keeping records of the hired candidates' advances on the corporate ladder. Several meta-analyses have been conducted on a large body of published studies showing that the interview validity coefficient can vary from low-end values of .10 (Dunnette, 1972) and .22 (Hunter & Hunter, 1984) to moderate values of .3 to .6 (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Conshaw, 1988). These numbers are correlation coefficients between the interview test outcomes and criteria — measures of professional success. Meta-analyses show that improving interview validity is possible by controlling for various factors: Among the most important are the amount of structure imposed during the interview (Huffcutt & Arthur, 1994; Schmidt & Zimmerman, 2004), the interviewer selection and training (Conway, Jako, & Goodman, 1995; Huffcutt & Woehr, 1999),

[☆] The authors are grateful for the comments and insights by the members of Adaptive Behavior and Cognition (ABC) group, Max Planck, Berlin, for improving earlier versions of this paper.

[*] Corresponding author at: Grand Valley State University, Psychology Department, One Campus Drive, Allendale, MI 49401, USA.
E-mail addresses: fificm@gvsu.edu (M. Fifić), gigerenzer@mpib-berlin.mpg.de (G. Gigerenzer).

http://dx.doi.org/10.1016/j.jbusres.2014.03.003
0148-2963/© 2014 Elsevier Inc. All rights reserved.

and the method of aggregation of independent reviewers' decisions (Dreher, Ash, & Hancock, 1988). Some meta-analyses indicate the significant impact of the number of interviewers on interview validity. Interview validity appears to improve as the number of interviewers increases (Conway et al., 1995; McDaniel et al., 1994; Schmidt & Zimmerman, 2004; Wiesner & Conshaw, 1988). Thus, in recruitment practice these findings appear to confirm Condorcet's Jury Theorem.

How to choose the number of interviewers?

The observations at the consulting firm led us to ask: Are more interviewers always better? Is there a systematic way to relate the number of interviewers to the resulting quality of the hiring process? The current study focuses here on the question of how many interviewers are needed to select the best m candidates out of a pool of size M, and exclude other goals that are simultaneously pursued in actual recruiting, such as to impress a candidate by an elaborate selection process, or to familiarize the faculty with the candidates. Before answering the question, the authors of the current study first checked whether such an elaborate process is typical in consulting firms. The authors retrieved information on a sample of companies, including 3M, Bain & Co, Booz & Co, Boston Consulting Group, Deloitte, Cargill, McKinsey & Co, PriceWaterhouseCoopers, Thomson, and Thrivent Financial for Lutherans. The number of interview rounds varied between 2 and 5 (on campus or in the company office), and the number of different interviewers per candidate varied between 5 and 11, depending on position (e.g., associate or senior consultant) and company. This informal survey revealed that an elaborate step-wise process is not uncommon, and that multiple interviewers appear to be standard.

A body of research compares collective decision making and individual experts' opinion. Specifically, this research has addressed the effect of how to combine individual votes into a collective vote, from various forms of aggregation such as the majority rule (Arkes, 2003; Hastie & Kameda, 2005; Reimer & Katsikopoulos, 2004; Sorkin, West, & Robinson, 1998) to averaging of individual judgments (Ariely & Levav, 2000; Armstrong, 2001; Clemen, 1989; Clemen & Winkler, 1987; Einhorn, Hogarth, & Klempner, 1977; Gordon, 1924; Hogarth, 1978; Johnson, Budescu, & Wallsten, 2001; Wallsten, Budescu, Erev, & Diederich, 1997; Winkler & Poses, 1993). One conclusion drawn is that more experts do better, consistent with Condorcet's Jury Theorem. Various amendments have been reported, such as that the increase is inversely related to the average inter-correlation among individual experts' opinions (Hogarth, 1978). In other words, adding new experts to an existing group leads to little improvement if the new experts make similar decisions (e.g., Winkler & Clemen, 2004). A second conclusion is that more is not always better. Several researchers note that the best experts in a group sometimes outperform the group's collective score (e.g., Gordon, 1924). In a study of physicians' performance in an intensive care unit (Winkler & Poses, 1993), the best prediction of patients' survival rates was obtained by taking averages of performance in the two best individually performing groups in the hospital rather than in all groups. Likewise, in a study of economists' ability to predict economic growth, the forecasts of economists with the best previous histories of performance were better than a combined group score (Graham, 1996).

Several other studies focusing on the individual measures of interviewer validity show that some interviewers are better than the others in selecting the best candidates (e.g., Dipboye, Gaugler, Hayes, & Parker, 2001; Ghiselli, 1966; Heneman, 1975; Pulakos, Schmitt, Whitney, & Smith, 1996; Yonge, 1956; Zedeck, Tziner, & Middlestadt, 1983). The implication of these studies is that adding more interviewers might harm the selection personnel process, thus potentially implying the contrary to Condorcet's Jury Theorem.

The existence of free riders is a proposed resolution of this apparent contradiction—the phenomenon that with increasing group size, the extent of experts' involvement in the group decreases (Albanese & van Fleet, 1985; Kameda, Tsukasaki, Hastie, & Berg, 2011; Kerr & Tindale, 2004). In consequence, the quality of collective decision making may decline. As the team grows larger, individual experts tend to feel less responsible for collective decision making and invest less in information accrual. Free riding is predicated on the belief that someone else in the team will collect and process the relevant pieces of information. The evidence for free riding has been investigated in criminal law for determining the right jury size, not too big and not too small (Mukhopadhaya, 2003). In organizational economics, some researchers have been argued that larger groups lead individual members to engage less in information acquisition (Holmstrom, 1982). In social psychology, free riding is attributed to individuals' loss of motivation to contribute to social groups (Kerr & Tindale, 2004).

We aim here at a more general analysis of the conditions under which “less is more” in choosing the right number of interviewers, including interviewer characteristics and the free riding phenomenon.

Setting and terminology

In this article, the authors derive a systematic answer to the question of how many interviewers are needed to select the best candidates. To do so, one first needs to define the setting, which the authors model after the situation in many large consulting firms, as described above. The task is to pick the m best candidates out of a pool of size M. The m top candidates are called *targets*. All other candidates are called *non-targets*. Each interviewer i is characterized by a *hit rate* h_i, which the authors define as the relative frequency of correct target identifications among the interviewer's m votes. A hit rate h_i defined here could be interchangeably used with the term *interviewer's selection validity*, as both can be used to measure the efficiency of personnel interview to predict future job performance of hired candidates. For instance, if $m = 10$, a hit rate of .8 means that an interviewer has an expectation (or long-run frequency) of correctly identifying 8 out of the 10 targets, while missing two and voting for two non-targets (false positives). In this setting, interviewers differ in h_i, and the identity of the best interviewer is known (e.g. Dougherty, Ebert, & Callender, 1986; Ghiselli, 1966; Yonge, 1956). In addition, pairs of interviewers can differ in *homogeneity* in judgment (defined below), which reflects the kind of cues they look for and the strategy for processing these cues.

Each interviewer conducts the interview alone and independently votes yes/no for each candidate to be hired (*interviewer independence*), with the constraint that the number of yes-votes equals m. Finally, the votes of the N interviewers are added up to determine who survives to the next round or who will be made an offer (this is called the *majority rule*, as in Condorcet's Jury Theorem). The majority rule specifies that each vote counts equally and the group decision is the tally of votes (Hastie & Kameda, 2005). In case of a tie between candidates, offers will be decided randomly. The ties are candidates who received an equal number of votes, but of whom only a subset can be selected as top m candidates. The resulting hit rate of the N interviewers achieved by applying the majority rule is their *collective hit rate*.

A team of N interviewers can be either *homogeneous* or *heterogeneous*. Consider the case of two interviewers. They form a homogeneous (nested) set if and only if the second interviewer's correct identifications form a subset of those chosen by the first interviewer. A team of homogeneous interviewers is likely if everyone has been trained to use similar cues to identify top candidates. Two interviewers form a heterogeneous team if their correct identifications are not nested. If two interviewers are heterogeneous, they are likely to rely on different cues to identify the best candidates. A heterogeneous team could be formed with the purpose of covering a broad range of interviewer experience using a large range of cues. Such interviewers will complement each other, focusing on identification of cues that fall outside the other's domain of expertise.

We now turn to the main question. Compared to what the best interviewer can achieve alone, does adding more interviewers lead to better results? Let's begin with the simplest case of two interviewers and then proceed to N > 2 interviewers. In closing, the authors of the current research will consider the situation when the best interviewer is not known and the influence of free riding.

Are two interviewers better than one?

Consider the task of selecting the top 10 out of 100 candidates. For a baseline, randomly picking 10 out of 100 would lead to an expectation of one correct vote, that is, a random hit rate of $h_r = .1$ or, in general terms, m/M. Interviewers that are better than chance have hit rates of $h_r < h_i \leq 1$. If the best interviewer had a hit rate of 1, it is easy to see that adding additional, less experienced interviewers would lead to a lower collective hit rate. Such an omniscient interviewer, however, is unlikely in the real world. More realistically, assume that the best interviewer gets eight out of the top 10 correct ($h_{best} = .8$), that is, misses two by voting for two less suitable candidates. Now let's add a second interviewer with $h_2 = .6$, who gets six correct, which is still much better than by chance. Does this additional interviewer improve the decision?

To answer the question, let's first consider the general case in which two interviewers are randomly selected from a pool with varying hit rates. For any pair of hit rates, the following result holds.

(1) The expected collective hit rate of a team of two interviewers, randomly sampled from a pool of independently and identically distributed hit rates, is $(h_1 + h_2) / 2$.

This perspective implies that adding a second interviewer to one with a higher hit rate will, on average, not increase the collective hit rate, but in fact decrease it. The result is derived in Appendix A, and can be generalized to any distribution of hit rates, whether the best interviewer is known or not. For example, if a firm randomly samples pairs of interviewers with $h_{best} = .8$ and $h_{least} = .6$, the expected collective hit rate is .7.

Result 1 concerns the expected collective hit rate but is mute about the variability. Can adding a second interviewer to the one known to have the best hit rate result in a collective hit rate higher than that of the best interviewer? To answer the question a candidate selection process is simulated using two interviewers with hit rates $h_1 = .8$ and $h_2 = .6$ and repeated 100,000 times. The simulation confirmed that the collective hit rate converges to .7, but also showed substantial variability. The expected collective hit rates of (.1, .2, .3, .4, .5, .6, .7, .8, .9, 1.0) were observed in approximately the following percent of simulations: (0, 0, 0.1, 1, 6.7, 22.5, 37.1, 26, 6.4, 0.3). Note that the obtained distribution of the collective hit rates is not symmetric. The reason is that the range of probabilities is bounded by 0 and 1. A symmetric distribution results if both hit rates are in the middle of the range, for instance, if both interviewers have a .5 hit rate. Hence, the expected collective hit rate of .7 was observed in 37% of the cases. The important result is that collective hit rates of .9 or 1.0 were obtained in about 6.7% of the cases altogether. That is, whereas adding a second interviewer decreases the hit rate on average from .8 to .7, the hit rate is increased in about 1 out of 15 pairs of reviewers. Equally often, however, adding a second reviewer decreases the number of correct identifications below that of the second-best interviewer: In 7.7% of the cases, the collective hit rate was .5 or below.

That is, the range of the collective hit rates of a team of two interviewers extends above the hit rate h_1 of the best interviewer and below the hit rate h_2 of the second interviewer. Can we identify features of individual interviewers that explain when two interviewers perform better than the best one, as in the 6.7% of cases, and when they do worse?

Homogeneous interviewers

First consider the case of two homogenous interviewers, that is, two interviewers whose hits are nested. Can adding a homogeneous interviewer improve the hit rate achieved by the best interviewer? The answer is no. This can be illustrated by adding a second interviewer with $h_2 = .6$ to one with $h_1 = .8$. Eight of the top 10 candidates get one vote from the first interviewer (h_1), and six of these get one vote from the second interviewer (h_2). The remaining 6 $(2 + 4)$ votes are false positives and distributed over 6 non-targets (assuming for simplicity and without loss of generality that none of the non-targets got 2 votes). By majority vote, the six targets with 2 votes each are selected, while from the 8 candidates with one vote each (two targets and six non-targets), 4 candidates are randomly drawn. The resulting expected collective hit rate is again .7. Yet its range is limited to between .6 and .8. Thus, unlike in the general case, adding a homogeneous interviewer can *never* improve on the hit rate of the best interviewer. This leads to the question: Can we identify the features of two homogeneous interviewers who at least do not make things worse?

Here is the general result.

(2) Adding a homogeneous second interviewer can never lead to a collective hit rate higher than that of the best interviewer, neither for the expected hit rate nor for its range. Only if the hit rate of the second interviewer equals that of the first will the collective hit rate be the same as the individual hit rates.

One interesting feature of Result 2 is that the collective hit rate is determined entirely by the nested choices of targets and independent of the distribution of the false positives. To illustrate, assume two interviewers with hit rates of .6 each. Because these interviewers are homogeneous, six targets receive two votes each. Consider now the two extreme cases for the distribution of the remaining votes. First, if both interviewers give their remaining 4 votes to the same non-targets, these four false-positives would be selected by the majority rule, without affecting the overall .6 collective accuracy. Second, if both interviewers give their remaining 4 votes to different non-targets, 8 false positives get one vote each, four of which are randomly selected. Again, the expected collective hit rate of .6 is unaffected by the distribution of votes over non-targets.

Heterogeneous interviewers

Consider now the case of heterogeneous interviewers, that is, the set of interviewers who are not homogeneous. Can adding a heterogeneous interviewer improve the hit rate achieved by the best interviewer? Consider again two cases, the extreme case of adding a “mirror interviewer” who identifies all candidates that the best interviewer misses, and the general case of two interviewers whose targets partially overlap.

Adding a mirror interviewer

First, let's add an interviewer who votes for exactly all the targets that the best interviewer missed and no other targets. That is, $h_1 + h_2 = 1$, and there is no overlap between hits. This interviewer is ideal in the sense of being able to identify cues that are outside the expertise of the best interviewer. In fact, teams are sometimes put together to represent heterogeneous competencies. Will the addition of such a mirror interviewer improve the hit rate?

For mirror interviewers, and heterogeneous interviewers in general, the votes for non-targets matter in answering this question. Thus, let's first consider the extreme sub-case where the interviewers vote for mutually exclusive sets of non-targets. Psychologically, this means that the two interviewers are complementary in identifying different targets but also in being impressed by different sets of non-targets.

Here is a surprising result.

(3) Adding a mirror interviewer who identifies exactly all targets missed by the best interviewer and is impressed by *different* non-targets leads to an expected hit rate of .5. To illustrate, assume that the first reviewer votes for 6 targets and the second the remaining four. Thus, each target has one vote, but 10 non-targets also have one vote each. All selected cases are therefore ties, and the decision has to be based on a random selection from the 20 with one vote each, which leads to an expected overall hit rate of .5. Finding an interviewer who

is complementary to the best interviewer in the sense described here is counterproductive.

The second extreme sub-case is one in which the votes for non-targets are nested. That is, the non-targets selected by the best interviewer form a subset of that of the second-best. Psychologically, this can be because both are impressed by the same misleading cues of the same candidates. Does the addition of a mirror interviewer help in this condition?

(4) Adding a mirror interviewer who identifies exactly all targets missed by the best interviewer and is impressed by the *same* non-targets leads to an expected hit rate lower than .5.

Adding a mirror interviewer hence leads to an even worse collective performance if both interviewers fall prey to the same set of non-targets as opposed to a different set. This is because $m(1 - h_1)$ non-targets will get two votes, thereby decreasing the probability of targets being selected.

All in all, adding a mirror interviewer who votes for exactly all the targets that the best interviewer missed will result in an expected collective hit rate of .5 or lower. Note that this result is independent of whether one knows the identity of the best interviewer. If the second interviewer has a higher hit rate, the same results will be obtained, because the hit rates of both interviewers add up to 1 in the case where all targets get exactly one vote.

Heterogeneous interviewers: The general case

Mirror interviewers form a subset of all heterogeneous interviewers. Fig. 1 illustrates a team of two interviewers who are heterogeneous and partially overlap in their identification of targets. Does it pay to add such an interviewer? The right side of Fig. 1 shows the result of the majority rule: 4 targets receive two votes, while 6 targets and 6 non-targets get one vote each. By choosing randomly 6 out of these 12 with one vote, one expects three more hits, which results in an expected collective hit rate of .7. The range is between 4 (all targets with two votes only) and 10 hits (lucky vote). The general result is as follows.

(5) Adding a heterogeneous second interviewer (whether hits overlap with those of the best interviewer or not) leads to an expected collective hit rate that is lower than that of the best interviewer. The range of the collective hit rates of two heterogeneous interviewers extends above the hit rate h_1 of the best interviewer and below the hit rate h_2 of the second interviewer. That is, the collective hit rate can be higher than that of the best interviewer.

The general consequence of this analysis is that in order to improve the expected hit rate, one should never add a second interviewer. This holds independent of whether the second interviewer can identify all targets the first interviewer missed or not, and of the preferences for non-targets. Instead, one should find means to improve the hit rate of the best interviewer. Do these conclusions also hold for $N \geq 2$ interviewers?

*Are **N** interviewers better than one?*

What about six interviewers, as used by the consulting company mentioned in the introduction? Although this company never performed the analysis offered here, the company might have intuitively chosen the right number. Consider first the case of $N \geq 2$ who form a homogeneous set, that is, every interviewer's correct identifications form a subset of those chosen by the best interviewer. It is easy to see that Result 2 generalizes to this situation: N interviewers can never lead to an expected collective hit rate higher than that of the best interviewer. The second sentence of Result 2 also generalizes if one replaces “the hit rate of the second interviewer” with “the hit rates of all interviewers.” If a set of N homogeneous interviewers cannot lead to a higher expected hit rate than that of the best interviewer, can a set of N heterogeneous interviewers do so?

Similar to the case of $N = 2$, with heterogeneous interviewers, the collective hit rate depends on the exact distribution of the votes within the target group, the distribution of the votes within the non-target group, and the individual hit rates of the N interviewers. To capture these complex relations in a representative way, the authors of the current study simulated selection committees of size $1 \leq N \leq 20$. The question this simulation answers is: How many additional interviewers need to be added to the best interviewer so that the collective hit rate is higher than that of the best interviewer alone?

The hit rates of the best interviewer were varied at $h_{best} = .9, .8, .7$, and .6. The additional $N - 1$ interviewers were sampled from a population of interviewers with $h_{best} \geq h_i \geq 1/2\ h_{best}$, or $h_{best} \geq h_i \geq 3/4\ h_{best}$ (for details, see Appendix B). This means that additional interviewers had strong performances. Some of them were as accurate as the best interviewer, and their hit rates were always in a high range and far above chance. As described previously, each interviewer voted for 10 out of 100 candidates, and the candidates who won the most votes were chosen.

Consider first the question whether one should add one additional interviewer to the best interviewer. Fig. 2 shows that this addition sharply decreases the expected collective hit rate. Thus, under the conditions defined in this analysis, two interviewers can never be expected to be better than one. Consider now adding two interviewers. When the variability between interviewers is high (Fig. 2, top), adding two interviewers leads to *lower* expected hit rates. A team of three interviewers always does better than a team of two interviewers, but not better than the best. Only teams of four interviewers begin to perform better than the best interviewer, and only if the best interviewer has a rather low hit rate ($h_{best} = .6$) and the added interviewers have hit rates

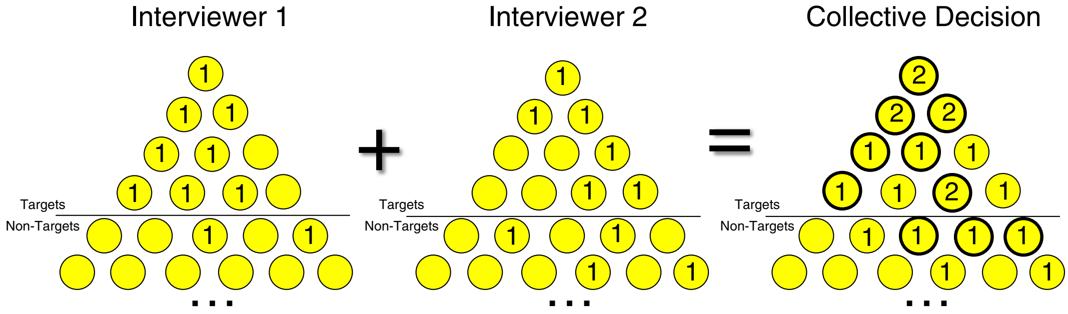

Fig. 1. Does it help to add Interviewer 2, who is able to identify all targets that Interviewer 1 missed, as well as some common targets? Note. Each circle denotes a candidate and indicates the number of votes for the candidate. The 10 target candidates are placed at the pyramid's top. Interviewer 1 has a hit rate of .8, that is, votes for 8 out of the top 10 candidates correctly. Interviewer 2 has a lower hit rate of .6 but picks all targets that Interviewer 1 missed. Using the majority rule, the collective decision is made (right side), and the selected candidates are marked by a bold circle. Surprisingly, the collective decision correctly identifies only 7 targets, that is, fewer than the best interviewer alone.

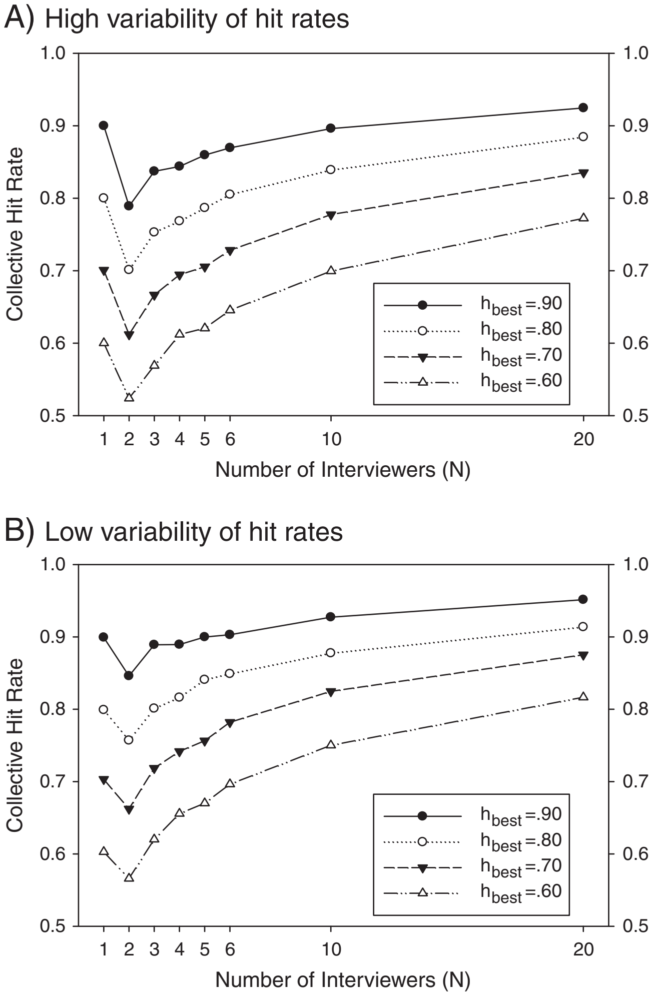

Fig. 2. The expected collective hit rate as a function of the number N of interviewers. Note. The best interviewer is always a member of the committee. Each line represents the changes to the collective hit rate with increasing N, for best interviewers with hit rates from 0.6 to 0.9. The top panel shows committees with members whose individual hit rates vary highly ($h_{best} \geq h_i \geq 1/2\ h_{best}$), and the bottom panel shows committees whose individual hit rates vary little ($h_{best} \geq h_i \geq 3/4\ h_{best}$). Results are based on 10,000 committees for each combination of hbest and N.

ranging between $.6 \geq h_i \geq$ ½ .6. This is shown in the lowest line in Fig. 2, top. If the best interviewer has a high hit rate ($h_{best} = .8$) and the added interviewers have hit rates ranging between $.8 \geq h_i \geq$ ½ .8, one needs a team of ten interviewers to obtain a noticeable improvement over the best interviewer. This is shown in the second line from the top in Fig. 2, top. When the interviewers are very good ($h_{best} = .9$), with added interviewers' hit rates ranging between $.9 \geq h_i \geq$ ½ .9, ten interviewers are no longer enough and one would need a team of 20 interviewers to outperform the best interviewer. This finding appears in the top line in Fig. 2, top.

In contrast, when the variability between interviewers is low (Fig. 2, bottom), the beneficial effect of adding interviewers increases faster than when the variability is high. Two interviewers are still always worse than one interviewer. Yet if the added interviewers have low hit rates, that is, $.6 \geq h_i \geq$ ¾ .6, and $.7 \geq h_i \geq$ ¾ .7, respectively, adding two interviewers leads to higher expected hit rates than that of the top interviewer alone. This is shown in the bottom lines in Fig. 2, bottom. When the added interviewers are very good, with hit rates ranging between $.9 \geq h_i \geq$ ¾ .9, one needs a team of 10 interviewers to outperform the best interviewer. This outcome appears in the top line of Fig. 2, bottom.

These observations lead to the following general results for $N \geq 3$.

(6) For all $N \geq 2$ (but not for $N \geq 1$), the collective hit rate increases with N, with diminishing returns. Beneficial effects of additional interviewers increase faster when the variability of interviewers' hit rates is lower.

(7) The likelihood that the best interviewer outperforms teams of $N \geq 3$ heterogeneous interviewers increases with higher values of h_{best} and higher variability of hit rates within the team of interviewers.

When the best interviewer is not known

Up to this point, the current analysis often depended on knowing who the best interviewer was. Such knowledge may not be always available, for instance, when the interviewers have been part of the company only briefly. Let's consider now the extreme alternative where companies know nothing about the order of quality among their interviewers and pick interviewers randomly. Fig. 3 shows the collective hit rate as a function of the number of interviewers when the best interviewer is not known, and when the interviewers are randomly sampled from a pool of employees. The current analysis shows a strictly monotonic increase in the collective hit rate with N. Thus, when nothing is known about the relative quality of the interviewers, having more interviewers in the committee always pays off, with one exception, $N = 2$. Adding a second interviewer to a randomly picked first interviewer does not lead to any improvement. The explanation is that in the long run, the expected collective hit rate for $N = 2$ is equal to the expected hit rate for $N = 1$. This leads to the following general result:

(8) The collective hit rate of N interviewers whose individual hit rates are not known increases monotonically with N, independent of the variability between interviewers. The only exception is a team of two interviewers, whose expected hit rate is never better than that of one interviewer.

Free riders

Adding more interviewers to a team may increase the quality of collective decision making, but may also lead to free riding. Free riding can be defined as a decrease in individual contribution to the quality of group decision. The psychological mechanism of free riding appears to be the following: As the size of a team increases, individual interviewers pay less attention to acquisition and processing of information about the candidates (Albanese & van Fleet, 1985; Kameda et al., 2011). The reason might be that interviewers feel that their influence on the final

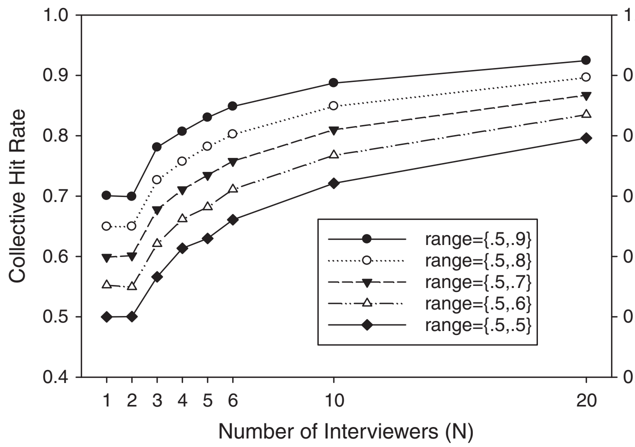

Fig. 3. Collective hit rate as a function of the number of interviewers, when the best interviewer is not known. Note. The interviewers are sampled from uniform distributions with specified ranges of hit rates.

decision is dwindling, or the spontaneous diffusion of responsibility in the presence of others (see Kerr & Tindale, 2004). As a consequence of free riding, the collective hit rate may decrease.

To quantify the free rider effect, the currents study adopts Mukhopadhaya's (2003) general model to the problem of selecting the number of interviewers. The model links the size of a group to the amount of free riding. The effect of free riding is defined by the product between an individual interviewer's hit rate and a free riding factor σ(N, c) ranging between 0 and 1, where N is the size of the team and c are the costs of searching and processing relevant information:

$$\sigma(N,c) = \begin{cases} 1-c^{\frac{1}{N-1}}, N>1 \\ 1, N=1 \end{cases}$$

The modified hit rate of an interviewer i is the hit rate times the free riding factor: $h_i \times \sigma_i(N, c)$.

For a team size N = 1, no free riding is assumed and the value of the free riding factor is σ(1, c) = 1. For N > 1, in a simulation study the costs c were varied between 0 and .001. When c = 0, the individual interviewer devotes full attention and no free riding. When c > 0, free riding occurs. For instance, with c = .001 and N = 10 group members, the individual hit rate is discounted by a factor of σ(10,.001) = 0.536. For c= .00001 and N = 10, the factor is σ(10,.00001) = 0.722.

Fig. 4 shows the expected collective hit rate as a function of the number of interviewers, for different levels of free riding. When the best interviewer is not known (left), this function is non-monotonic. For smaller teams (N = 1,2), the expected collective hit rate is the average of the individual hit rates (as in Fig. 3). For a group size of three interviewers, the expected collective hit rate increases. Unlike what is shown in Fig. 3, the expected collective hit rate for N > 3 may decrease or increase, depending on the amount of free riding.

When the best interviewer is known (Fig. 4, right), adding more interviewers does not help much. The best interviewer matches the expected collective hit rate of the team of size five, and outperforms every team in which free riders exist.

What size of an interviewing team is best, and how large can a team be before free riding is likely to occur?

(9) In the case of expected influence of free riding, a single best interviewer should be chosen instead of a team of interviewers. If the best interviewer is not known, then the team should comprise exactly three members (N = 3).

The N = 3 team does not appear to be affected much by the magnitude of free riding (Fig. 4). The free riding dominates other team sizes within the same free riding curve when the best interviewer is both known or not.

Discussion

One of the authors of the current study was approached by a consulting company with the following question. When forming an interviewing team, how many interviewers should be included to achieve the best results? A common theme in the literature is that the quality of collective decision is improved by adding more members to a group, as long as their individual decisions are better than chance (e.g., Berend & Paroush, 1998; Condorcet, 1785). The *more the better* approach suggests that the size of an interviewing team should be as large as possible. But large teams expend more financial resources, spend more time, deplete firm's resources, and slow down decision-making processes due to aggregation of individual decisions.

How then to determine the size of the interviewing team? The first approach is to employ an optimization procedure (e.g. Grofman, Owen, & Feld, 1983; Karotkin & Paroush, 2003). To find the optimal team size within the constraints of time, resources, and search costs, the optimization model would need the exact prior probabilities, likelihoods, and costs associated with each new interviewer. Unfortunately, these data typically do not exist in practice (but see Dipboye et al., 2001; Heneman, 1975; Pulakos et al., 1996; Zedeck et al., 1983). Consulting firms operating in the global market live in a so-called "large world" (Binmore, 2009) in which it is not possible to capture the statistical regularities with any certainty because the relevant information is only partially known. Whereas optimization models have to make assumptions that are likely to be too strong, the second approach tries to capture this uncertainty with "satisficing rules" to find a good-enough trade-off between accuracy and effort (Payne, Bettman, & Johnson, 1993). In both cases, however, the common assumption is that more interviewers are better, and the only problem is to find a trade-off between increasing accuracy and the increasing costs of more interviewers.

Our argument is different. The problem is not to find the optimal trade-off between accuracy and costs. As shown, independent of the costs, adding more interviewers is not always better—there is not always an accuracy-effort trade-off in the first place. Why would Condorcet's Jury Theorem not hold up in the world specified in the

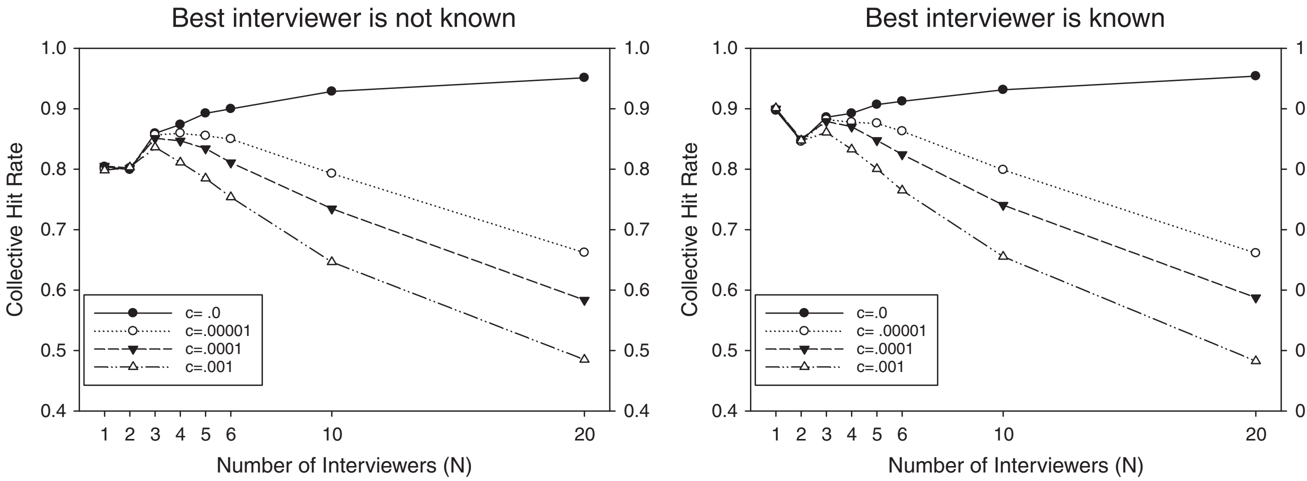

Fig. 4. The influence of free riding on the expected collective hit rate as a function of the number of interviewers. Note. Degrees of free riding are introduced by varying the costs of information c. In both panels, the interviewers were sampled from the uniform distribution of hit rates ranging between .7 and .9. In the right panel, .9 is the hit rate of the best interviewer.

current analysis? The authors of the current study identified two general conditions under which the opposite rule, *less is more*, holds: (1) If the best interviewer is known, the function between N and collective hit rate is no longer generally monotonic. That is, there are regions where adding more interviewers does not help. (2) In a case of free riding, the function is also not generally monotonic. Moreover, the effects of both conditions combine.

Policy implications

Note that all policy implications are relative to the assumptions made in this study: that the best interviewer is known, interviewers vote independently, and the majority rule is followed to arrive at the collective decision.

The first policy implication is that one interviewer is better than two. Adding a second interviewer is never a good idea because it decreases the expected hit rate. Even when the best interviewer is not known a second interviewer does not contribute to expected gain. Second, if one wants to add several (two or more) interviewers, which characteristics should they have? The policy implication is to not ask the best interviewer to teach others how to identify candidates. Doing so would likely result in a set of homogenous interviewers who use the same cues or strategies, and a homogeneous set is never better than the best interviewer alone (Result 2, which generalizes to any N). Nor will there be any improvement by finding a mirror interviewer who can identify exactly the targets missed by the best interviewer (Results 3 and 4). Rather, the solution is to add a larger set of heterogeneous interviewers. The size of this set depends on the range of individual hit rates: The closer the hit rates are to that of the best interviewer and the smaller their variability, the fewer additional interviewers are needed for a given improvement.

Third, if the general situation encourages free riding, one might abstain from trying to find a sufficiently large group of heterogeneous interviewers and instead stick with the best interviewer. The reason is that the policy to increase collective accuracy by increasing N is counteracted by the negative effect of free riding, which also increases with larger N. If the best interviewer is not known and one has no knowledge about the hit rates of the various interviewers, a set of three interviewers appears to be a robust solution.

All in all, the present analysis advocates the general policy to invest in intensive training of the best interviewer rather than distribute resources among many interviewers. A good example of such practice is found in studies reporting that using only one interviewer (Dougherty et al., 1986; Ghiselli, 1966; Yonge, 1956) can result in the best selections.

When assumptions are violated

Multiple rounds

In the current analysis, the authors assumed that all interviews are conducted in one round. Yet many firms divide the interview process into a series of rounds. In the case study that motivated this paper, the interview consisted of two rounds with three interviews each. From the current analysis, using three interviewers would be a good general policy if free riding cannot be ruled out (Fig. 4) or the best interviewer is not known. However, if the best interviewer is known, using three interviewers is likely to decrease the hit rate, both in the presence of free riding or not, and specifically with high variability of hit rates (Fig. 2). Thus, the present analysis can be extended to personnel selection with multiple rounds by applying the analysis to each round.

Votes are not independent

In the current study it is assumed that individuals' votes are not influenced by the votes of the other group members. If team members communicate after voting individually but stand by their decisions, the analysis would still be valid. However, communication between team members can affect collective decisions (Austen-Smith & Banks, 1996; see also Gerling, Grüner, Kiel, & Schulte, 2005). In practice, votes become subject to others' influence in panel sessions in which all interviewers can exchange opinions about the candidates and reach a group consensus about hiring and rejecting decisions. A meta-analysis did not show the advantage of predictive validity of a panel over the independent individual decisions (Wiesner & Conshaw, 1988; but see Conway et al., 1995). The most likely influence appears to be that interviewers assimilate and revise their individual votes, resulting in a more homogeneous set of interviewers. In that case, Result 2 suggests that the more homogenous the group becomes, the less the improvement over the best interviewer will be. A more detailed analysis of the effect of sharing information is beyond the scope of this paper and can be based on work on correlated votes and hidden profiles (Berg, 1993; Grofman et al., 1983; Hogarth, 1978; Lightle, Kagel, & Arkes, 2009; Lombardelli, Proudman, & Talbot, 2005; Reimer & Hoffrage, 2005; Stasser & Titus, 2003; Winkler & Clemen, 2004).

When the best interviewer is exhausted

The general policy conclusion to train and engage the best interviewer only might lead to physical and mental exhaustion. This objection is only partially correct. In a situation where the same set of $N \geq 2$ interviewers sequentially interrogates all candidates, the time commitment of the best interviewer is actually the same as when that interviewer works alone. Yet if the number of candidates is very large, and a company uses rotating teams of interviewers, then using the best interviewer alone can result in an exhausting time commitment that may lead to fatigue and decreased performance. In this case, the policy implication would be to train two or more best interviewers and use each of them alone in a rotating scheme.

Conclusions

Recruitment is key to success in companies, academic departments, and beyond. To ensure the best possible recruits, much time and money is invested, and interviews are almost always part of the process. The question of whether more interviewers are always better appears to have not been studied before. The present study specifies conditions under which higher accuracy, not only higher efficiency, is obtained by reducing the size of the interviewing team. In fact, a single best interviewer can outperform larger committees across a variety of conditions, showing the highest hit rate. Using single interviewers also reduces coordination costs, eliminates the costs of sharing information and aggregation, and reduces free riding to a minimum. In general, better recruitments can be achieved by training and grooming a single top interviewer rather than investing equally in a team of interviewers.

Appendix A

The assumption, here, is that interviewers are independent, have above-chance hit rates, and use the majority rule to determine that collective decision. In case of a tie, the decision is made randomly.

Proof of Result 1: Assume two interviewers with hit rates h_1 and h_2. For each of m targets, each interviewer can vote "+" (identify a target) or "−" (miss a target), resulting in four combinations. The total probability P of all these combinations is $P(+,+) + P(+,-) + P(-,+) + P(-,-) = 1$. The probability that two interviewers both have a hit is $P(+,+) = h_1 \cdot h_2$, and that both miss a target is $P = (-,-) = (1 - h_1)(1 - h_2)$. The probability that only one interviewer has a hit is $P(-,+) + P(+,-) = 1 - P(+,+) - P(-,-) = h_1 + h_2 \ - 2h_1h_2$, which is derived from the total probability of all possible events.

Applying the majority rule:

$$\textit{Collective hit rate} = P(+,+) + 1/2(P(-,+) + P(+,-))$$
$$= h_1 \cdot h_2 + 1/2(h_1 + h_2 - 2h_1h_2) = (h_1 + h_2)/2$$

Appendix B *Details of simulations*

Candidates

Assume that each candidate is characterized by a competency value X. For a group of candidates of a size M, this value is represented as a partially ordered set with a finite number of real values $X = \{X_1, X_2, X_3, \ldots X_j, \ldots X_M\}$, in which for each successive pair the following holds: $X_j < X_{j+1}$, for all j ε {1,2,3,..,M}. For the simulation, X = {1,2,3, ….,100} and m = 10.

Interviewers

The property X cannot be observed directly. An interviewer can estimate X by X′, which is presented as follows:

$$X' = X + e,$$

where e represents the value of internal decision noise error that is normally distributed (N[0,var]).

For a group of candidates of a size M, j ε {1,2,3,..,M}, this leads to:

$$X_j^{\,'} = X_j + e$$

The individual interviewer's hit rate is defined as the proportion of the m = 10 targets that are among the 10 votes. When the variance of e equals zero, then the X′ is equivalent to X, and the interviewer's hit rate equals one.

Group of interviewers

Assume that each interviewer i is characterized by a different value of *e* such that

$$X_{i,j}^{\,'} = X_{i,j} + e_i.$$

The matrix $X_{i,j}'$ represents a matrix with the dimensions of i × (m = 10), that is, the number of interviewers times the number of candidates selected by each interviewer. In the current simulation, the individual hit rates had a uniform distribution, $h_i = \text{Uniform}[h_{\text{least}}, h_{\text{best}}]$, for $1 \geq h_{\text{best}} \geq h_{\text{least}} \geq 0$. When a group of i interviewers were sampled, error values that produce the desired hit rates were calculated.

Majority rule

Each interviewer votes for the 10 candidates (m = 10) with the highest X′ values. From these values, the 10 candidates with the highest number of votes are selected. In the case of a tie, the candidates are randomly selected. The collective hit rate is defined as the proportion of the targets in the 10 selected candidates. The candidate selection process described above was repeated 10,000 times for each number of interviewers (N = 1, 2, 3, 4, 5, 6, 10, 20). The average across these repetitions is reported in the figures as the collective hit rate.

References

Albanese, R., & van Fleet, D.D. (1985). Rational behavior in groups: The free-riding tendency. *Academy of Management Review*, *10*(2), 244–255.

Ariely, D., & Levav, J. (2000). Sequential choice in group settings: Taking the road less traveled and less enjoyed. *Journal of Consumer Research*, *27*(3), 279–290.

Arkes, H. R. (2003). The nonuse of psychological research at two federal agencies. *Psychological Science*, *14*, 1–6.

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners: Norwell.* MA: Kluwer Academic Publishers.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the Condorcet Jury Theorem. *American Political Science Review*, *90*(1), 34–45.

Berend, D., & Paroush, J. (1998). When is Condorcet's Jury Theorem valid? *Social Choice and Welfare*, *15*(4), 481–488.

Berg, S. (1993). Condorcet's Jury Theorem, dependency among jurors. *Social Choice and Welfare*, *10*, 87–96.

Binmore, K. (2009). *Rational decisions.* Princeton, NJ: Princeton University Press.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clemen, R. T., & Winkler, R. L. (1987). Calibrating and combining precipitation probability forecasts. In R. Viertl (Ed.), *Probability and Bayesian Statistics* (pp. 97–110). New York: Plenum.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix [Essay on the application of analysis to the probability of majority decisions].* Paris: Imprimerie Royale.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565–579.

Dipboye, R. L., Gaugler, B. B., Hayes, T. L., & Parker, D. (2001). The validity of unstructured panel interviews: More than meets the eye? *Journal of Business and Psychology*, *16*(1), 35–49.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology*, *71*(1), 9–15.

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. *Personnel Psychology*, *41*(2), 315–327.

Dunnette, M.D. (1972). *Validity study results for jobs relevant to the petroleum refining industry.* Washington, DC: American Petroleum Institute.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158–172.

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

Gerling, K., Grüner, H. P., Kiel, A., & Schulte, E. (2005). Information acquisition and decision making in committees: A survey. *European Journal of Political Economy*, *21*(3), 563–597.

Ghiselli, E. E. (1966). The validity of a personnel interview. *Personnel Psychology*, *19*(4), 389–394.

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, *7*, 398–400.

Graham, J. R. (1996). Is a group of economists better than one? Than none? *Journal of Business*, *69*(2), 193–232.

Grofman, B., Owen, G., & Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*(3), 261–278.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508.

Heneman, H. G. (1975). *The impact of interviewer training and interview structure on the reliability and validity of the selection interview.* Proceedings of Academy of Management, 231–233.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40–46.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, *13*(2), 324–340.

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184–190.

Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior*, *20*(4), 549–560.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72–98.

Johnson, T. R., Budescu, D.V., & Wallsten, T. S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making*, *14*(2), 123–140.

Kameda, T., Tsukasaki, T., Hastie, R., & Berg, N. (2011). Democracy under uncertainty: The wisdom of crowds and the free-rider problem in group decision making. *Psychological Review*, *118*(1), 76–96.

Karotkin, D., & Paroush, J. (2003). Optimum committee size: Quality-versus-quantity dilemma. *Social Choice and Welfare*, *20*(3), 429–441.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655.

Krause, J., & Ruxton, G. D. (2002). *Living in groups.* Oxford, UK: Oxford University Press.

Lightle, J. P., Kagel, J. H., & Arkes, H. R. (2009). Information exchange in group decision making: The hidden profile problem reconsidered. *Management Science*, *55*(4), 568–581.

Lombardelli, C., Proudman, J., & Talbot, J. (2005). Committees versus individuals: An experimental analysis of monetary policy decision making. *International Journal of Central Banking*, *1*(1), 181–205.

McDaniel, M.A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599–616.

Mukhopadhaya, K. (2003). Jury size and the free rider problem. *Journal of Law, Economics, and Organization*, *19*, 24–44.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge, UK: Cambridge University Press.

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology*, *49*(1), 85–102.

Reimer, T., & Hoffrage, U. (2005). Can simple group heuristics detect hidden profiles in randomly generated environments? *Swiss Journal of Psychology*, *64*(1).

Reimer, T., & Katsikopoulos, K. V. (2004). The use of recognition in group decision-making. *Cognitive Science*, *28*(6), 1009–1029.

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology*, *89*, 553–561.

Sorkin, R., West, R., & Robinson, D. (1998). Group performance depends on the majority rule. *Psychological Science*, *9*, 456–463.

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, *3–4*, 302–311.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* New York, NY, US: Doubleday & Co.

Wallsten, T. S., Budescu, D.V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268.

Wiesner, W. H., & Conshaw, S. F. (1988). The moderating impact of interview format and degree of structure on interview validity. *Journal of Occupational Psychology*, *61*, 275–290.

Winkler, R. L., & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, *1*(3), 167.

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

Yonge, K. A. (1956). The value of the interview: An orientation and a pilot study. *Journal of Applied Psychology*, *40*(1), 25–31.

Zedeck, S., Tziner, A., & Middlestadt, S. E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology*, *36*(2), 355–370.