OPERATIONS RESEARCH
Vol. 62, No. 3, May–June 2014, pp. 483–511
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2014.1269




CROSSCUTTING AREAS

# Judge: *Don't Vote*!

**Michel Balinski**
Centre national de la recherché scientifique, Laboratoire d'Econométrie, Ecole Polytechnique, 91128 Palaiseau Cedex, France,
michel.balinski@polytechnique.edu

**Rida Laraki**
Centre national de la recherché scientifique, Laboratoire d'Analyse et modelisation de Systemes pour l'Aide a la DEcision,
Université Paris-Dauphine, 75775 Paris Cedex 16, France; and Département d'Economie, Ecole Polytechnique, 91128 Palaiseau Cedex, France,
rida.laraki@polytechnique.edu

This article argues that the traditional model of the theory of social choice is not a good model and does not lead to acceptable methods of ranking and electing. It presents a more meaningful and realistic model that leads naturally to a method of ranking and electing—*majority judgment*—that better meets the traditional criteria of what constitutes a good method. It gives descriptions of its successful use in several different practical situations and compares it with other methods including Condorcet's, Borda's, first-past-the-post, and approval voting.

*Subject classifications*: methods of electing and ranking; Condorcet and Arrow paradoxes; strategic manipulation; faithful representation; meaningful measurement; figure skating; presidential elections; jury decision.
*Area of review*: OR Practice.
*History*: Received February 2012; revisions received February 2013, August 2013, January 2014; accepted January 2014. Published online in *Articles in Advance* April 28, 2014.

*"The final test of a theory is its capacity to solve the problems which originated it."*
George B. Dantzig (Dantzig 1963, p. vii)

## 1. Why?

George Dantzig's limpid, opening phrase of the preface of his classic work on linear programming and extensions (Dantzig 1963, p. vii) is worth repeating over and over again, for it is far too often forgotten. By his final test, the theory of voting has failed. Despite insightful concepts, fascinating analyses, and surprising theorems, its most famous results are for the most part negative: paradoxes leading to impossibility and incompatibility theorems. We argue that the theory has yielded no really decent methods for practical use and that this is due, in essence, to how voting has been viewed.

Since 1299 (and perhaps before) voting has been modeled in terms of *comparing the relative merits* of candidates. In this conception voters are assumed to rank order the candidates (the inputs), and the problem is to amalgamate these so-called preferences into the rank order of society (the output).

If, instead, voters *evaluate the merit of each candidate* in a well-defined ordinal scale (the inputs) and majorities determine society's evaluation of each candidate and thereby its rank ordering of all (the outputs), then, we claim, the most important paradoxes of the traditional theory of voting are overcome.

Viewed through one lens this change of paradigm is small: a vote on the candidates themselves is replaced by votes on the final grade to be given to each candidate. Viewed through another lens the change looms large: the basic meaning of "majority" is interpreted and practiced differently bringing with it very important theoretical and practical consequences. Significantly, by asking more of voters—permitting much more accurate expressions of their opinions—it places greater confidence in them.

### 1.1. Why *Don't Vote*! in Theory

Rank-order inputs lead to two insurmountable paradoxes that plague practice, and therefore theory:

1. *Condorcet's paradox.* In the presence of at least three candidates, $A$, $B$, and $C$, it is entirely possible that in head-to-head encounters, $A$ defeats $B$, $B$ defeats $C$, and $C$ defeats $A$, so transitivity fails and a *Condorcet cycle* is produced, $A \succ_S B \succ_S C \succ_S A$, where $X \succ_S Y$ means society prefers $X$ to $Y$.
2. *Arrow's paradox.* In the presence of at least three candidates, it is possible for $A$ to win, yet with the same voting opinions $B$ defeats $A$ when $C$ withdraws.

These paradoxes are real. They occur in practice. Condorcet's paradox was observed in a Danish election (Kurrild-Klitgaard 1999). It has occurred in skating (see below). It also occurred in the famous 1976 "Judgment of Paris" where eleven voters—well-known wine experts—evaluated six cabernet sauvignons of California and four of Bordeaux, and the "unthinkable" is supposed to have occurred: in the phrase of *Time* magazine (June 7, 1976) "California defeated all Gaul." In fact, by Condorcet's majority principle, five wines—including three of the four French wines—all

preferred to the other five wines by a majority, were in a Condorcet cycle, $A \approx_S B \succ_S C \approx_S D \succ_S E \succ_S A$, where $X \approx_S Y$ means society or the jury considers $X$ and $Y$ to be tied (Balinski and Laraki 2010, §7.8; Balinski and Laraki 2013a). Moreover, after having seen it happen in practice Charles Dodgson observed in 1876 that voting strategically rather than honestly to optimize the outcome is likely to provoke Condorcet cycles (Dodgson 1876) (confirmed by experiments, see Balinski and Laraki 2010, §19.2).

Arrow's paradox is seen frequently. Had Ralph Nader not been a candidate for the presidency in the 2000 election in Florida, it seems clear that most of his 97,488 votes would have gone to Albert Gore who had 537 votes less than George W. Bush, thus making Gore the winner in Florida and so the national winner with 291 electoral college votes to Bush's 246. Bill Clinton was the winner with 43% of the popular vote in 1992, George Bush and Ross Perot together polling 56%: the evidence suggests Bush would have won pitted against Clinton alone. And the same may be argued for the election of 1912: Woodrow Wilson would most likely have lost against either Theodore Roosevelt or Williams Taft alone (who together had over 50% of the votes).

Arrow's paradox is also seen in judging. According to the rules that were used for years in amalgamating judges' opinions of figure skating performances—where their inputs were rank orders of skaters—it often happened that the relative position of two skaters could invert, or "flip-flop," solely because of another skater's performance (see below for concrete evidence). And the same has occurred in ranking wines: the "winner" among the set of all 10 wines of a competition is not the winner among subsets of them (Balinski and Laraki 2013a).

Behind these paradoxes lurk a host of impossibilities inherent to the traditional model. A brief account is given of several of them. The model is this. Each voter's input is a rank order of the candidates. Their collective input is society's *preference profile* $\Phi$. The output, society's rank order of the candidates, is determined by a rule of voting $F$ that depends on $\Phi$. It must satisfy certain basic demands. (1) *Unlimited domain.* Voters may input whatever rank orders they wish. (2) *Unanimity.* When every voter inputs the same rank order, then society's rank order must be that rank order. (3) *Independence of irrelevant alternatives (IIA).*[1] Suppose that society's rank order over all candidates $\mathscr{C}$ is $F(\Phi^{\mathscr{C}})$ and that over a subset of the candidates, $\mathscr{C}' \subset \mathscr{C}$, it is $F(\Phi^{\mathscr{C}'})$. Then the rank order obtained from $F(\Phi^{\mathscr{C}})$ by dropping all candidates not in $\mathscr{C}'$ must be $F(\Phi^{\mathscr{C}'})$. (4) *Nondictatorial.* No one voter's input can always determine society's rank order whatever the rank orders of the others.

Theorem 1 (Impossibility (Arrow 1951)). *There is no rule of voting that satisfies the properties* (1) *to* (4) (*when there are at least three candidates*).

Arrow's theorem explicitly ignores the possibility that voters have strategies. It assumes voters' "true" opinions may be expressed as rank orders and that they are their inputs, not some other inputs chosen strategically to maximize the outcome they wish. A rule of voting is *strategy proof* or incentive compatible when every voter's best strategy is to announce his true preference order; otherwise, the rule is *manipulable*. Strategy-proof or incentive compatible rules are desirable, for then the true preferences of the voters are amalgamated into a decision of society rather than some other set of strategically chosen preferences. Regrettably they do not exist.

However, the very formulation of the theorem that proves they do not exist underlines a defect in the traditional model. In general, the output of a rule of voting is society's rank order. Voters usually "prefer" one rank order to another, viz., the rank order of the candidates is important to a voter, the rank order of figure skaters in Olympic competitions is important to skaters, judges, and the public at large. But voters and judges have no way of expressing their preferences over rank orders. In the spirit of the traditional approach they should be asked for their rank orders of the rank orders (for a more detailed discussion of this point see Balinski and Laraki 2010, §§4.6 and 9.4). Be that as it may, when strategic choices are introduced in the context of the traditional approach something must be assumed about the preferences of the voters to be able to analyze their behavior. It is standard to assume that voters only care about who wins, i.e., voters' utility functions depend only on who is elected. This is certainly not true for judges of competitions. This is also false for many voters:[2] why, otherwise, did so many U.S. voters opt for Ralph Nader in the presidential election of 2000 in the knowledge that he could never be the winner, or why do so many voters opt for minor candidates in all of France's presidential elections knowing they can never win? Voters vote because they wish to send messages that express their opinions.

Each voter's input is now a rank order that is chosen strategically, so it may or may not correspond to her true "preferences." A rule of voting is assumed to produce a winner only, and unanimous means that when all the voters place a candidate first on their lists then so does the rule.

Theorem 2 (Impossibility (Gibbard 1973, Satterthwaite 1975)). *There is no rule of voting that is unanimous, nondictatorial, and strategy proof for all possible preference profiles* (*when there are at least three candidates*).

In carefully analyzing a proposal of Condorcet, Young noticed that there was a conflict between, on one hand, a winner, and on the other hand, the first in an order-of-finish (Young 1988). A third result shows that this conflict is inescapable in the context of the traditional approach. To explain, it an additional concept must be invoked. When there are $n$ candidates $A_i$ $(i = 1, \ldots, n)$, a set of $kn$ voters with the preference profile

$$
\begin{array}{lccccccccc}
k\colon & A_1 & \succ & A_2 & \succ & \cdots & \succ & A_{n-1} & \succ & A_n \\
k\colon & A_2 & \succ & A_3 & \succ & \cdots & \succ & A_n & \succ & A_1 \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
k\colon & A_n & \succ & A_1 & \succ & \cdots & \succ & A_{n-2} & \succ & A_{n-1}
\end{array}
$$

(the first line meaning, for example, that $k$ voters have the preference $A_1 \succ A_2 \succ \cdots \succ A_{n-1} \succ A_n$) is called a *Condorcet component*. Each candidate appears in each place of the order $k$ times. Given a preference profile that is a Condorcet component, every candidate has the same claim to the first, the last, or any other place in the order-of-finish: there is a vast tie among all candidates for every place.

The model is now this. Voters input rank orders, a rule amalgamates them into society's rank order. The first-place candidate is the winner, the last-place candidate is the loser. The rule must enjoy three properties: (1) *Winner-loser unanimous.* Whenever all voters rank a candidate first (respectively, last) he must be the winner (the loser). (2) *Choice compatible.* Whenever all voters rank a candidate first (respectively, last) and a Condorcet component is added to the profile, that candidate must be the winner (the loser). (3) *Rank compatible.* Whenever a loser is removed from the set of candidates, the new ranking of the remaining candidates must be the same as their original ranking (a weak IIA).

THEOREM 3 (WINNER/RANKING INCOMPATIBILITY (BALINSKI AND LARAKI 2007, 2010)). *There is no rule of voting that is winner-loser unanimous, choice and rank compatible* (*when there are at least three candidates*).

Theorem 3 shows that there is an inherent incompatibility between winners or losers and orders-of-finish. Imagine the following situation: All but one figure skater, Miss $LS$, have performed, and Miss $FS$ is in first-place among them. Then Miss $LS$ performs. The result is that she finishes last but Miss $FS$ is no longer in first place. Rank compatibility is violated, but a method that guarantees it is satisfied implies one of the other two properties may not be met, which is unthinkable.

There is still another fundamental difficulty with the traditional model. Clearly, if a voter has a change of opinion and decides to move some candidate up in her ranking that candidate should not as a consequence end up lower in the final ranking, that is, the method of voting should be "choice monotone." Monotonicity is essential to any practically acceptable method: how can one accept the idea that when a candidate rises in the inputs he falls in the output? But there are various ways of formulating the underlying idea. Another is "rank monotone": if one or several voters move the winner up in their inputs, not only should he remain the winner but the final ranking among the others should not change.

THEOREM 4 (MONOTONIC INCOMPATIBILITY (BALINSKI ET AL. 2009)). *There is no unanimous, impartial rule of voting that is both choice monotone and rank monotone.*[3]

Moreover, when some nonwinner falls in the inputs of one or more voters, no method of the traditional model can guarantee that the winner remains the winner (none is "strongly monotone" (Muller and Satterthwaite 1977)). Why all of this happens is simple: moving some candidate up necessarily moves some candidate(s) down, though there may be no change of opinion regarding *them*.

In short, these four theorems show, we believe, that there can be no good method of voting. But operations research is not *only* theorems and algorithms, it is *also* formulating adequate models. To begin, a problem must be understood as best as can be. Next, a model must be formulated that attempts to capture the essentials of the real situation. It must then be challenged by the gritty details of the real problem. Only then is it worthwhile to develop and explore the mathematical properties of the model. But this, in turn, can—invariably, will—lead to new understandings of the problem, to refinements and reformulations of the model, and so eventually to new probing conclusions. Indeed, operations research that seeks to solve real problems consists of a sequence of repetitions of this process.

What is amazing about the theory of social choice is that the basic model has not changed over seven centuries. Comparing candidates has steadfastly remained the paradigm of voting. And yet, both common sense and practice show that voters and judges do not formulate their opinions as rank orders. Rank orders are grossly insufficient expressions of opinion, because a candidate who is second (or in any other place) of an input may be held in high esteem by one voter but in very low esteem by another.

Moreover, rank ordering competitors is difficult to do. There is ample evidence for this. With the old rules for judging figure skaters, the inputs of judges were rank orders of the performers, but the judges were not asked to submit rank orders, for that is much too difficult. Instead, they were asked to give number grades, and their number grades were used to deduce their rank orders. Indeed, this is the routine in schools and universities where students' grades are used to determine their standings. In the last three presidential elections held in France, there were, respectively, 16, 12, and 10 candidates. Voters certainly did not rank order the candidates. Instead, they rejected most and chose one among several whom they held in some degree of esteem (possibly high, often rather low, though it was impossible for them to express such sentiments). A voting experiment carried out in parallel with the 2007 presidential election showed that fully one-third of the voters did not have a single preferred candidate and that the merits of candidates ranked highest in a voter's input, or ranked second highest in his input, etc., were seen to be quite different (Balinski and Laraki 2010, 2011). Is it at all reasonable, then, to count the highest ranked (or the second highest ranked, etc.) candidate of two voters in the same way?

Thus the traditional approach to voting fails for two separate reasons:

- The model's inputs are inadequate.
- The model's implications exclude a satisfactory procedure.

The goal of this paper is to give a brief account of a new paradigm and model for a theory of social choice that (1) enables judges and voters to express their opinions

**Table 1.** Scores of competitors given by nine judges (performance plus technical marks).

| Name | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | $J_6$ | $J_7$ | $J_8$ | $J_9$ | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| T. Eldredge | 11.3 | 11.6 | 11.3 | 11.4 | 11.4 | 11.7 | 11.4 | 11.2 | 11.5 | 11.42 |
| C. Li | 10.8 | $11.2^+$ | 11.0 | 10.9 | 10.6 | 11.0 | 10.8 | 10.9 | 11.2 | 10.93 |
| M. Savoie | 11.1 | $10.8^+$ | 11.1 | $10.8^+$ | 10.5 | 10.8 | 10.6 | 10.5 | 11.1 | 10.81 |
| T. Honda | 10.3 | 11.2 | 10.9 | 11.0 | 10.8 | $10.9^+$ | 10.4 | 10.3 | 10.7 | 10.72 |
| M. Weiss | 10.6 | 11.1 | 10.6 | 10.8 | 10.4 | 10.9 | 10.9 | 10.4 | 10.9 | 10.73 |
| Y. Tamura | 09.8 | 10.8 | 10.1 | 10.4 | 11.0 | 11.6 | 10.7 | 10.6 | 10.8 | 10.64 |

*Note.* $x^+$ is ranked above $x$.

naturally and much more accurately than rank orders; and (2) escapes the traditional impossibilities just discussed. For a complete presentation of the theory, a detailed justification of its basic paradigm, and descriptions of its uses to date and of experiments that have been conducted to test it, see Balinski and Laraki (2010).

### 1.2. Why *Don't Vote*! in Practice

Everything is ranked all of the time: architectural projects, beauty queens, cities, dogs, economists, figure skaters, graduates, hotels, investments, journals, kung fu fighters, light heavyweight boxers, musicians, novelists, operations research analysts, ..., and zoologists, not only candidates for offices. How? Usually by evaluating them in a common language of grades. That it is natural to do so is evident since it is so often done—and shows the reason why a theory is needed to determine how the grades should be amalgamated. In most *real* competitions (other than elections) the order-of-finish of competitors is a function of number grades attributed by judges. Usually the functions used to amalgamate judges' grades are their sums, or equivalently, their averages. But this is not nor was always so—and it need not be so. The recent changes in the rules used in figure skating offer a particularly interesting case study.

**1.2.1. Condorcet's and Arrow's Paradoxes.** Although there already had been occurrences of Arrow's paradox in the past, including the 1995 woman's world championship, what happened in the 1997 men's figure skating European championships was the extra drop that caused a flood. Before A. Vlascenko's performance, the rule's top finishers were A. Urmanov first, V. Zagorodniuk second, and P. Candeloro third. Then Vlascenko performed. The final order-of-finish placed him sixth, confirmed Urmanov's first, but put Candeloro in second place and Zagorodniuk in third. The outcry over this flip-flop was so strident that the president of the International Skating Union (ISU) finally admitted something must be wrong with the rule in use and promised it would be fixed. Accordingly, the rules were changed. The ISU adopted the OBO rule ("one-by-one") in 1998. It is explained via a *real problem* that dramatically shows the many difficulties that may be encountered with the traditional approach (for us this example is as important as a theorem).

The Four Continents Figure Skating Championships are annual competitions with skaters from all the continents save Europe (whence the "four"). In 2001 they were held in Salt Lake City, Utah. The example discussed comes from the men's "short program." There were 22 competitors and nine judges. The analysis is confined to the six leading finishers. It happens that doing so gives exactly the same order-of-finish among the six as is obtained with all competitors (it ain't necessarily so!). Every judge assigns to every competitor two grades, each ranging between 0 and 6, one "presentation mark" and one "technical mark." Their sums determine each judge's input. The data concerning the six skaters is given in Table 1.

Contrary to public belief the sum or the average of the scores given a skater did not determine a skater's standing. They were only used as a device to determine each judge's rank order of the competitors.

When two sums are the same but the presentation mark of one competitor is higher than the other's, then that competitor is taken to lead the other in the judge's input. This ISU rule breaks all ties in the example; when a tie occurs a "+" is adjoined next to the number (in Table 1) that indicates a higher presentation mark, so indicates higher in the ranking. The judges' rank orders of the competitors—their inputs to the OBO rule—are given in Table 2. Thus, for example, judge $J_1$ ranked Eldredge first, Savoie second, ..., and Tamura last.

Up to here, the new rule is identical to the old one (for details see Balinski and Laraki 2010). The innovation was in how the judges' inputs are amalgamated into a decision. The OBO system combines two of the oldest and best known voting rules, Llull's—a generalization of Condorcet's known by some as Copeland's (Copeland 1951)—and Cusanus's—best known as Borda's method. To use what we will call Llull's and Borda's rules, Table 3 gives the numbers of judges that prefer one competitor to another for all pairs of

**Table 2.** Judges' inputs (indicating rank orders of the six competitors).

| Name | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | $J_6$ | $J_7$ | $J_8$ | $J_9$ |
|---|---|---|---|---|---|---|---|---|---|
| T. Eldredge | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C. Li | 3 | 2 | 3 | 3 | 4 | 3 | 3 | 2 | 2 |
| M. Savoie | 2 | 5 | 2 | 4 | 5 | 6 | 5 | 4 | 3 |
| T. Honda | 5 | 3 | 4 | 2 | 3 | 4 | 6 | 6 | 6 |
| M. Weiss | 4 | 4 | 5 | 5 | 6 | 5 | 2 | 5 | 4 |
| Y. Tamura | 6 | 6 | 6 | 6 | 2 | 2 | 4 | 3 | 5 |

**Table 3.** Judges' majority votes in all head-to-head comparisons.

| | T. Eldredge | C. Li | M. Savoie | T. Honda | M. Weiss | Y. Tamura | Number of wins | Borda score |
|---|---|---|---|---|---|---|---|---|
| T. Eldredge | — | 9 | 9 | 9 | 9 | 9 | 5 | 45 |
| C. Li | 0 | — | 7 | 7 | 8 | 7 | 4 | 29 |
| M. Savoie | 0 | 2 | — | 5 | 6 | 5 | 3 | 18 |
| T. Honda | 0 | 2 | 4 | — | 5 | 4 | 1 | 15 |
| M. Weiss | 0 | 1 | 3 | 4 | — | 6 | 1 | 14 |
| Y. Tamura | 0 | 2 | 4 | 5 | 3 | — | 1 | 14 |

competitors. Thus, for example, Savoie is ranked higher than Weiss by six judges, so ranked lower by three.

Condorcet was for declaring one competitor ahead of another if a majority of judges preferred him to the other. But, of course, his paradox may arise. It does in this example,

$$\text{Honda} \succ_S \text{Weiss} \succ_S \text{Tamura} \succ_S \text{Honda}.$$

A more general rule than Condorcet's was proposed in 1299 by Ramon Llull (Hägele and Pukelsheim 2001): *Llull's method*. Rank the competitors according to their numbers of wins plus ties.[4] It is a more general rule because a Condorcet winner is necessarily a Llull winner. Eldredge is the Condorcet winner and Llull winner, and Llull's rule yields the ranking

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Honda} \approx_S \text{Weiss} \approx_S \text{Tamura}.$$

The first three places are clear, but there is a tie for the next three places. Eldredge is the *Condorcet winner* because he is ranked higher by a majority of judges in all pair-by-pair comparisons. There is no *Condorcet loser* because no skater is ranked lower by a majority in all pair-by-pair comparisons.

Cusanus in 1433 (Hägele and Pukelsheim 2008) and later Borda in 1770 (Borda 1784) had an entirely different idea. In *Borda's method* (it is so well known under this name that we use it too) a competitor $C$ receives $k$ *Borda points* if $k$ competitors are below $C$ in a judge's rank order; $C$'s *Borda score* is the sum of his Borda points over all judges; and the Borda ranking is determined by the competitors' Borda scores. Alternatively, a competitor's Borda score is the sum of the votes he receives in all pair by pair votes. Thus the Borda scores in Table 3 are simply the sums of votes in the rows, and the Borda ranking of the six candidates is

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Honda} \succ_S \text{Weiss} \approx_S \text{Tamura}.$$

Borda's method, however, often denies first place to a Condorcet winner or last place to a Condorcet loser, and that has caused many to be bewitched, bothered, and bewildered (though Borda's method suffers from much worse defects as will soon become apparent).

There is an essential difference in the two approaches. Whereas Llull and Condorcet rely on each candidate's total number of wins against all other candidates in head-to-head confrontations, Cusanus and Borda rely on each candidate's total number of votes against all other candidates in head-to-head confrontations.

The *OBO rule* used in skating is the following:

1. Rank the competitors by their number of wins (thereby giving precedence to the Llull and Condorcet idea).
2. Break any ties by using Borda's rule.

In this case Borda's rule yields a refinement of Llull's, so the OBO rule ranks the six skaters as does Borda,

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Honda} \succ_S \text{Weiss} \approx_S \text{Tamura}.$$

This was the official order-of-finish. The OBO rule is also known as *Dasgupta-Maskin's method* (Dasgupta and Maskin 2004, 2008). They proposed it with elaborate theoretical arguments, calling it "the fairest vote of all," though it had been tried and discarded in skating.

The OBO rule produces a linear order, so is not subject to Condorcet's paradox, but it is (unavoidably) subject to Arrow's paradox, viciously so in this example. For suppose that the order of the performances had been first Honda, then Weiss, Tamura, Savoie, Li, and Eldredge. After each performance, the results are announced. Among the first three, the judges' inputs are the ones shown in Table 4. This yields the majority votes, numbers of wins, and Borda-scores in Table 5, so the result is

$$\text{Weiss} \succ_S \text{Honda} \succ_S \text{Tamura}$$

(note that majority voting yields a Condorcet cycle, Honda $\succ_S$ Weiss $\succ_S$ Tamura $\succ_S$ Honda).

**Table 4.** Judges' inputs, three competitors.

| Name | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | $J_6$ | $J_7$ | $J_8$ | $J_9$ |
|---|---|---|---|---|---|---|---|---|---|
| T. Honda | 2 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 |
| M. Weiss | 1 | 2 | 2 | 2 | 3 | 3 | 1 | 2 | 1 |
| Y. Tamura | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 1 | 2 |

**Table 5.** Majority votes in head-to-head comparisons, three competitors.

| | T. Honda | M. Weiss | Y. Tamura | Number of wins | Borda score |
|---|---|---|---|---|---|
| T. Honda | — | 5 | 4 | 1 | 9 |
| M. Weiss | 4 | — | 6 | 1 | 10 |
| Y. Tamura | 5 | 3 | — | 1 | 8 |

**Table 6.** Judges' inputs, four competitors.

| Name | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | $J_6$ | $J_7$ | $J_8$ | $J_9$ |
|---|---|---|---|---|---|---|---|---|---|
| M. Savoie | 1 | 3 | 1 | 2 | 3 | 4 | 3 | 2 | 1 |
| T. Honda | 3 | 1 | 2 | 1 | 2 | 2 | 4 | 4 | 4 |
| M. Weiss | 2 | 2 | 3 | 3 | 4 | 3 | 1 | 3 | 2 |
| Y. Tamura | 4 | 4 | 4 | 4 | 1 | 1 | 2 | 1 | 3 |

**Table 7.** Majority votes in head-to-head comparisons, four competitors.

| | M. Savoie | T. Honda | M. Weiss | Y. Tamura | Number of wins | Borda score |
|---|---|---|---|---|---|---|
| M. Savoie | — | 5 | 6 | 5 | 3 | 16 |
| T. Honda | 4 | — | 5 | 4 | 1 | 13 |
| M. Weiss | 3 | 4 | — | 6 | 1 | 13 |
| Y. Tamura | 4 | 5 | 3 | — | 1 | 12 |

For the first four skaters the judges' inputs are shown in Table 6, yielding the majority votes, numbers of wins and Borda scores in Table 7, so the result is

$$\text{Savoie} \succ_S \text{Weiss} \approx_S \text{Honda} \succ_S \text{Tamura}.$$

Before Savoie's performance Weiss led Honda; afterward they were tied.

Compare this with the final standings among all six skaters after the performances of Eldredge and Li (already computed):

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Honda} \succ_S \text{Weiss} \approx_S \text{Tamura}.$$

The last three did not perform, and yet Honda—who had once been tied with Weiss and once behind him—is now ahead of him, and Weiss—who had been ahead of Tamura—is now tied with him.

The ISU had discarded its old *ordinal rule*—used for many years—in 1998. It prescribed a competitor's median place in the standings as his final place in the standings (an idea first advanced by Galton (Galton 1907)), giving the result (where the median place is in parentheses following the names of each skater)

$$\text{Eldredge}(1) \succ_S \text{Li}(3) \succ_S \text{Savoie}(4) \approx_S \text{Honda}(4) \succ_S \text{Weiss}(5) \approx_S \text{Tamura}(5).$$

Recent social choice literature first proposed the median as a rule for the traditional model only after it had been discarded by the ISU (Bassett Jr and Persky 1999) (without, it seems, realizing that Galton had done so earlier). They advanced the median because of its statistical robustness. However, they made no provisions for ties, and no rule when the number of judges is even. The ISU resolved ties by the size of the majority in favor of at least the competitor's final place, which in this case puts Weiss (with seven) ahead of Tamura (with five) but leaves Savoie and Honda tied (at five). The ISU resolved further ties by summing up the numbers that corresponded to the candidates' final place or better,[5] which in this case puts Savoie (with 15) ahead of Weiss (with 16), and gives the result

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Honda} \succ_S \text{Weiss} \succ_S \text{Tamura}.$$

Note however that—as with the OBO rule—flip-flops can occur, and do: the ordinal rule gives the order Weiss $\succ_S$ Honda $\succ_S$ Tamura among the three alone. Indeed, in the women's world championships of 1995 the fourth place finisher performed after the three who finished ahead of her, but her performance changed the silver and bronze medals.

This chaotic behavior of repeated flip-flops is completely unacceptable to spectators, competitors, and of course common sense. It is no isolated phenomenon. Similar chaotic behavior occurs in the famous 1976 Paris wine tasting (Balinski and Laraki 2013a). It is inherent to the old ordinal rule, the OBO, Borda, and other methods as well.

**1.2.2. Strategic Manipulation.** The OBO rule was abandoned by the ISU following the big scandal of the 2002 Winter Olympics (also held in Salt Lake City). In the pairs figure skating competition the gold medal went to a Russian pair, the silver to a Canadian pair. The vast majority of the public, and many experts as well, were convinced that the gold should have gone to the Canadians, the silver to the Russians. A French judge confessed having favored the Russian over the Canadian pair, saying she had yielded to pressure from her hierarchy, only to deny it later. That judges manipulate their inputs—reporting grades not in keeping with their professional opinions—is known. A recent statistical analysis concluded "[Judges] . . . appear to engage in bloc judging or vote trading. A skater whose country is not represented on the judging panel is at a serious disadvantage. The data suggests that countries are divided into two blocs, with the United States, Canada, Germany, and Italy on one side and Russia, the Ukraine, France and Poland on the other" (Zitzewitz 2006). Once again the skating world entered into fierce fights over how to express and how to amalgamate the opinions of judges. Finally—thankfully—the idea that judges' inputs should be rank orders was abandoned. In so doing, the ISU joined the growing number of organizations whose rules direct judges to assign number grades to candidates, and the candidates' average grades determine the orders-of-finish (including diving, wine tasting, gymnastics, pianists, restaurants, and many others).

Such rules are usually known as *point-summing methods*; in the context of elections some call it *range voting*. The judges' scores in the 2001 Four Continents Figure Skating Championships provide an immediate example. Take the judges' inputs to be the scores themselves. They range from a low of 0 to a high of 12. The candidates' average scores are given in Table 1 and yield an order-of-finish that differs from that of the Borda and OBO rules:

$$\text{Eldredge} \succ_S \text{Li} \succ_S \text{Savoie} \succ_S \text{Weiss} \succ_S \text{Honda} \succ_S \text{Tamura}.$$

**Table 8.** Judge $J_2$'s manipulations that change the order-of-finish to what she wishes (given in the first row).

| | T. Eldredge | C. Li | M. Savoie | T. Honda | M. Weiss | Y. Tamura |
|---|---|---|---|---|---|---|
| $J_2$: | 1st | 2nd | 5th | 3rd | 4th | 6th |
| | 11.6 | $11.2^+$ | $10.8^+$ | 11.2 | 11.1 | 10.8 |
| | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| | 12.0 | 11.9 | $10.2^+$ | 11.8 | 11.4 | 10.2 |
| Averages: | 11.42 | 10.93 | 10.81 | 10.72 | 10.73 | 10.64 |
| | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| | 11.47 | 11.01 | 10.74 | 10.79 | 10.77 | 10.58 |

*Note.* Note that her new grades define the same order.

It is at once evident that judges can easily manipulate the outcome by assigning their grades strategically. Every judge can both increase and decrease the final score of every competitor by increasing or decreasing the score given to that competitor.

In this case it is particularly tempting for judges to assign scores strategically. Suppose they reported the grades they believed were merited. Take, for example, judge $J_2$. She can change her scores (as indicated in the top part of Table 8, e.g., increasing that of Eldredge from 11.6 to 12.0 so that his average goes from 11.42 to 11.47) so that the final order-of-finish is exactly the one she believes is merited. Moreover, the new scores she gives agree with the order of merit she believes is correct. But judge $J_2$ is not unique in being able to do this: every single judge can alone manipulate to achieve precisely the order-of-finish he prefers by changing his scores. And each can do it while maintaining the order in which they placed them initially (given in Table 2). Results are announced following every performance, so judges accumulate information as the competition progresses and may obtain insights as how to best manipulate.

This analysis shows how extremely sensitive point-summing methods are to strategic manipulation; in fact, they are more open to manipulation than any other method of voting. This is important because the reason for voting is to arrive at the true collective decision of a society or jury.

**1.2.3. Faithful Representation and Meaningfulness.** How to construct a scale is a science—measurement theory—that raises two key problems (Krantz et al. 1971). First, *the faithful representation problem*: What scale? "When measuring some attribute of a class of objects or events, we associate numbers . . . with the objects in such a way that the properties of the attribute are faithfully represented as numerical properties" (Krantz et al. 1971, p. 1). For example, if the scale is a finite set of numbers from 0 to 20, should they be spaced evenly or otherwise? Second, *the meaningfulness problem*: Given a faithful representation, what analyses of sets of measurements are valid? For example, if the scale consists of the integers $0, 1, \ldots, 20$ when is it justified to sum and take averages of measurements?

Pain, for example, is measured on an 11-point *ordinal* scale going from 0 to 10, each number endowed with a careful verbal description: it is not meaningful to sum or average such measures since an increase from (say) 2 to 3 cannot be equated with an increase from 8 to 9. Temperature, Celsius or Fahrenheit, is an *interval* scale because equal intervals have the same significance: sums and averages are meaningful but multiplication is not, for there is no absolute 0. Ounces, inches, and the Kelvin temperature scale are *ratio* scales: they are interval scales where 0 has an absolute sense and multiplication is meaningful as well.

To appreciate the significance of what it means to add scores in competitions—that is, to construct an interval measure—consider two practical examples. The decathlon is an athletic competition consisting of 10 track and field events. For each event a competitor receives a number of points depending on his performance. The sum of the points across all events is the competitor's final score. How should the points be related to the performance? This is a nontrivial problem. In practice the formula for the 100-meter dash gives 651 points for 12 seconds, 861 for 11 seconds, 1,096 for 10 seconds, and 1,357 for nine seconds. Going from 12 seconds to 11 adds 210 additional points; from 10 seconds to nine garners an additional 285, although no human being has ever run that distance in nine seconds. The merit of reducing the time by one second should not be measured linearly: it should be related to the difficulty of the improvement if the points are to constitute a valid interval measure. That difficulty may be assessed by the frequency with which it is realized: the distribution of the performances across "all" competitors determines how the points are assigned. So, given a distribution for the 100-meter dash, *ideally* each time should be mapped into points so that the same percentage of performances belong to any two intervals of points $[x, x+\epsilon]$ and $[y, y+\epsilon]$. This gives to each interval of the same length the same meaning, and so transforms the performances into points that belong to an interval measure. Similarly, any distribution of performances may be mapped into a uniform distribution in an interval scale of points.

A second practical example confirms this interpretation—Denmark's new seven-grade number language adopted for the academic year 2006–2007. It has seven numerical grades: 12, 10, 7, 4, 2, 0, or −3. For sums and averages to make any sense at all, this scale must be an interval measure. The language of grades is described as follows:

> 12 (*A*)—*outstanding*, no or few inconsiderable flaws, 10% of passing students,
>
> 10 (*B*)—*excellent*, few considerable flaws, 25% of passing students,

7 (*C*)—*good*, numerous flaws, 30% of passing students,
4 (*D*)—*fair*, numerous considerable flaws, 25% of passing students,
2 (*E*)—*adequate*, the minimum acceptable, 10% of passing students,
0 (*Fx*)—*inadequate*,
−3 (*F*)—*entirely inadequate*.

Is there any relation between these seemingly peculiar scores and the prescribed distributions? Imagine that all the real numbers from two up to 12 are possible passing grades in an examination. Underlying the idea of an interval measure is that over the grades of many students in the closed interval $[2, 12]$, the percentages of students who obtain grades in intervals of the same length are the same. Which of the five passing grades should be assigned to a 5.7? The grade whose number is closest to 5.7, namely, 7 or *good*; or, more generally, any number from the interval $[5.5, 8.5]$ should be mapped into a *good*. By the same token any grade from the interval $[2, 3]$ is mapped into an *adequate*, from $[3, 5.5]$ into a *fair*, from $[8.5, 11]$ into an *excellent*, and from $[11, 12]$ into an *outstanding*. The five numbers (2, 4, 7, 10, 12) seem to have been chosen so that the intervals occupy, respectively, the percentages of the whole equal to the percentages of passing grades specified in the definition: $[2, 3]$ occupies 10% of the interval, $[3, 5.5]$ occupies 25%, $[5.5, 8.5]$ occupies 30%, $[8.5, 11]$ occupies 25%, and $[11, 12]$ occupies 10%. Thus equal intervals do have the same significance: on average, the same percentage of passing students belong to each interval and on average, 10% are *outstanding*, 25% are *excellent*, and so on down to 10% are *adequate*. Thus the Danish system attempts to construct an interval measure so that it is meaningful to add and compute averages of the numbers it assigns students.

More formally, suppose $k$ number grades, $x_1 < x_2 < \cdots < x_k$, are to be given, and their percentages are to be $(p_1, p_2, \ldots, p_k)$, so $\sum p_j = 100$. The grades constitute an interval measure when for all $i$, $x_i$ is in the interval $[p_1 + \cdots + p_{i-1}, p_1 + \cdots + p_i]$ and $\sum_{j=1}^{i} p_j$ is the midpoint of the interval $[x_i, x_{i+1}]$.

Let $q_i = \sum_1^i (-1)^{j+1} p_j$ for $i = 1, \ldots, k$.

Theorem 5 (Balinski and Laraki 2010, p. 172). *There exist number grades $x = (x_1, \ldots, x_k)$ that constitute an interval measure for the percentage distribution $(p_1, \ldots, p_k)$ if and only if there exists a $\delta \geqslant 0$ that satisfies*

$$\max_i q_{2i} \leqslant \delta \leqslant \min_j q_{2j+1}.$$

*When such $\delta$ exist, $x$ satisfying*

$$x_{2i} = -\delta + 2\sum_j^i p_{2j-1} \quad \text{and} \quad x_{2i+1} = \delta + 2\sum_j^i p_{2j}$$

*defines a set of interval measure grades for each possible value of $\delta$.*

The theorem is proven by taking $x_1 = \delta$ and doing a bit of algebraic manipulation.

In the Danish case—namely, $p = (10, 25, 30, 25, 10)$—there is a unique $\delta = 0$ because $q = (10, -15, 15, -10, 0)$ and $\max\{-15, -10\} \leqslant \min\{10, 15, 0\}$. Thus, $\delta = 0$ and $x = (0, 20, 50, 80, 100)$. Rescaling them by dividing by 10, then translating up by 2 yields the equivalent Danish grades. If instead the Danes had observed or stipulated the percentages $p = (10, 19, 42, 19, 10)$, then $q = (10, -9, 33, 14, 24)$ so $\max\{-9, 14\} > \min\{10, 33, 24\}$: there would be no set of interval measure grades.

Sometimes the percentages stipulated or observed admit an interval measure, sometimes not. When several are possible they are not equivalent: one set cannot be obtained from the other by scaling and translating since a change in the value of $\delta$ moves the grades with odd indices in the opposite direction of the grades with even indices. When the value of $\delta$ is unique, the solution is unstable, for some small perturbation in the percentages always renders an interval measure impossible. For example, for an $\epsilon > 0$ perturbation of the Dane's original percentages, $p = (10, 25 + \epsilon, 30 - \epsilon, 25, 10)$ there is no set of interval measure grades. In conclusion, for any given set of percentages either there is no set of interval measure grades, or it is unique but unstable, or there are several sets that are not equivalent: these are troublesome facts that together suggest mechanisms that depend on adding or averaging should be shunned.

Nevertheless, point-summing methods are pervasive (and very old). Since they sum candidates' scores they must—to be meaningful—be drawn from a common interval scale, yet typically they are not. Although in many applications such as figure skating the numbers of the scale have commonly understood meanings, an increase of one base unit invariably becomes more difficult to obtain the higher the score, implying scores do not constitute an interval scale, and suggesting that their sums and averages are not *meaningful* in the sense of measurement theory. Another application to which the same remarks apply is the 1976 Paris wine tasting: a point-summing method was used, it did not rely on an interval scale, and the resulting ranking was highly questionable (Balinski and Laraki 2013a).

Recently point-summing methods have been proposed for political elections by bloggers in France and the United States. *Range voting*[6] uses the scale $[0, 100]$. The scores are not defined, they are given no common meaning, so one voter's 71 may mean something entirely different from another's 71: the scale is not a faithful representation. *Vote de valeur*[7] has five scores, $0, \pm 1, \pm 2$, but here, in response to our criticisms, they have been assigned meanings: $+2$ is *very favorable*, $+1$ *favorable*, 0 *neutral*, $-1$ *hostile*, $-2$ *very hostile*: the scale is a faithful representation. But in either case nothing justifies the choice of the numbers, nor does anything justify summing them: they are not interval scales so sums and averages are not meaningful in the sense of measurement theory.

**Table 9.** Results, Institut BVA polls, March 22, 2007 (a month before the first round of the French presidential election of 2007).

| | Would each of the following be a good President of France? | | | | | |
|---|---|---|---|---|---|---|
| | Yes, certainly (%) | Yes, probably (%) | **Yes** (%) | Not really (%) | Not at all (%) | **No** (%) |
| Ségolène Royal | 21 | 28 | **49** | 22 | 26 | **48** |
| François Bayrou | 18 | 42 | **60** | 22 | 14 | **36** |
| Nicolas Sarkozy | 28 | 31 | **59** | 18 | 20 | **38** |
| Jean-Marie Le Pen | 4 | 8 | **12** | 13 | 71 | **84** |

| | Do you personally wish each of the following to win the presidential election? | | | | Could you personally vote for each of the following in the presidential election? | | | |
|---|---|---|---|---|---|---|---|---|
| | Yes, certainly (%) | Yes, somewhat (%) | **Yes** (%) | **No** (%) | Yes, certainly (%) | Yes, probably (%) | **Yes** (%) | **No** (%) |
| Ségolène Royal | 14 | 22 | **36** | **48** | 27 | 26 | **53** | **42** |
| François Bayrou | 6 | 22 | **28** | **53** | 25 | 44 | **59** | **26** |
| Nicolas Sarkozy | 13 | 17 | **30** | **53** | 28 | 26 | **54** | **41** |
| Jean-Marie Le Pen | — | — | — | — | 8 | 11 | **19** | **76** |

*Notes.* The answers were given for each candidate independently (the difference between 100% and total **yes**'s plus **no**'s in each row is the percentage of no responses, e.g., in the top table 3% gave no response on Royal). Figures for Le Pen were not given in the "personal wish" question.

*Approval voting* (Brams and Fishburn 1983)—a voter assigns a 1 ("approves") or a 0 to each candidate and the candidates are ranked according to their total numbers of 1's—suffers for similar reasons. It has been *practiced* as a point-summing method—e.g., in the words of the Social Choice and Welfare Society's 2007 ballot for electing its president, "You can vote for any number of candidates by ticking the appropriate boxes," the number of ticks determining the candidates' order of finish—though it has been *analyzed* via the traditional model. Both points of view invite comparisons, so strategic voting, and thus Arrow's paradox may occur (e.g., if some voter's favorite candidate withdraws she may change her vote and decide to give a tick to one or more other candidate(s), causing a change in the order-of-finish among the candidates that remain). But its most fundamental problem is that one person's tick may mean something altogether different than another's. So, ticks do not constitute a faithful representation of the quality of candidates and their sum—not meaningful in terms of measurement theory—are at best very rough approximations.

A French national poll proves the point. It posed seemingly close but different questions in several polls preceding the French presidential election of 2007 (see Table 9). Different questions elicit different responses: so, confronted by *no question* voters supply their own, respond accordingly, and the results are not interpretable. Indeed, asked to answer "yes" or "no" the same polls illustrate these can have very different gradations. Thus for voters or judges to express themselves adequately, the scale must contain more than two levels.

*First-past-the-post* or plurality voting—a voter is allowed to give one tick at most and a candidate's total ticks decides his place in the order of finish—is worse. A poll conducted by the BVA Institute on April 10, 2007 (12 days before the election) asked: "When voting in the first round of the [coming] presidential election, which of the following two attitudes correspond most closely to the way in which you will vote?

- I vote for the candidate on my side of the political spectrum who has the greatest chance of making the runoff.
- I vote for the candidate closest to my ideas even if he has little chance of making the runoff."

Thirty-two percent indicated the first attitude, 55% the second (13% indicated neither). Ticking exactly one candidate does not even provide a scale, so their sums have even less meaning. Note, moreover, that this shows some 55% of French voters do not care only about who wins, *their utility functions depend also on factors other than who is elected.*

To summarize, *voting or judging is measuring.* The scale used by approval voting and first-past-the-post is not a faithful representation of voters' opinions; moreover, the semantics are confusing, one tick lumping all kinds of different meanings into one. Taking their sum is tantamount to declaring one mile + one meter + one inch = three and is at best a very imprecise measure. The semantics of rank-order inputs are perfectly clear, but they are far too limited to permit a faithful expression of opinion, deny the existence of any common scale, and lead to unacceptable methods. Point-summing methods exaggerate in the other direction, assuming the existence of a perfect scale of measurement—an interval scale—which is almost impossible to achieve and, in any case, leads to highly manipulable methods. There is, however, a middle ground that asks for more than rank orders but less than an interval scale: an ordinal scale of merit.

### 1.3. A More Realistic Model

Postulate a finite number of *competitors* or *candidates* $\mathscr{C} = \{C_1, \ldots, C_m\}$; a finite number of *judges* or *voters* $\mathscr{J} = \{1, \ldots, n\}$; and a *common language of grades* $\Lambda = \{\alpha, \beta, \gamma, \ldots\}$ that is a totally ordered set.

In practice (e.g., piano competitions, figure skating, gymnastics, diving, wine competitions), common languages of grades are invented to suit the purpose and are carefully defined and explained. Their words are clearly understood, much as the words of an ordinary language, or the measurements of physics. But they almost surely do not constitute interval scales. The grades or words are "absolute" in the sense that every judge uses them to measure the merit of each competitor independently. They are "common" in the sense that judges assign them with respect to a set of benchmarks that constitute a shared scale of evaluation. They are ordinal scales and constitute faithful representations.

What scales are adequate? That depends on the particular application. In wines, a common language of seven words—*excellent, very good, good, passable, inadequate, mediocre, bad*—is used by judges to evaluate each of 14 attributes (concerning aspect, aroma, taste, flavor, ...).[8] In judging diving, 21 numbers—multiples of one-half in the interval [0, 10], carefully defined—are used by judges to evaluate a dive (which has a degree of difficulty).[9] In reaching their decision on the 2009 Louis Lyons Award for Conscience and Integrity in Journalism, the judges at the Nieman Foundation at Harvard University used majority judgment. They chose to use a common language of seven grades—*absolutely outstanding, outstanding, excellent, very strong, strong, commendable, neutral*—to rank five very highly considered nominees. Had each of the judges in these cases ranked the competitors, their inputs would have been merely relative, barring any scale of evaluation and ignoring any sense of shared benchmarks. In general, the more grades the better given that judges can naturally distinguish their meanings. Professional judges are typically able to distinguish more levels than a "general" public. In political elections some six or seven levels seems best (as seen below). There is more meaning in common when voters assign about seven grades than fewer or more (Miller 1956).

A problem is specified by its *inputs*, a *profile*

$$\Phi = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots & \vdots \\ \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{in-1} & \alpha_{in} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kn-1} & \alpha_{kn} \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{pmatrix}$$

where $\alpha_{ij} = \Phi(C_i, j) \in \Lambda$ is the grade assigned by judge $j \in \mathscr{J}$ to competitor $C_i \in \mathscr{C}$. With this formulation of inputs voters specify rank orders determined by the grades (that may be strict if the scale of grades is fine enough), so in this sense the inputs include those of the traditional model. Experience proves they are simple and cognitively natural.

Suppose competitor $C$ is assigned the grades $(\alpha_1, \ldots, \alpha_n)$ and competitor $C'$ the grades $(\beta_1, \ldots, \beta_n)$. A *method of ranking* is a nonsymmetric binary relation $\succeq_S$ that compares any two competitors whose grades belong to some profile. By definition $C \succeq_S C'$ and $C' \succeq_S C$ means $C \approx_S C'$; and $C \succ_S C'$ if $C \succeq_S C'$ and not $C \approx_S C'$. So $\succeq_S$ is a complete binary relation.

What properties should any reasonable method of ranking $\succeq_S$ possess?

1. *Neutrality*. When $C \succeq_S C'$ for the profile $\Phi$, $C \succeq_S C'$ for the profile $\sigma\Phi$ for any permutation $\sigma$ of the competitors (or rows). That is, the competitors' ranks do not depend on where their grades are given in the inputs.

2. *Anonymity*. When $C \succeq_S C'$ for the profile $\Phi$, $C \succeq_S C'$ for the profile $\Phi\sigma$ for any permutation $\sigma$ of the voters (or columns). That is, no judge has more weight than another judge in determining the ranks of competitors. When a rule satisfies these first two properties, it is called *impartial*.

3. *Transitivity*. If $C \succeq_S C'$ and $C' \succeq_S C''$ then $C \succeq_S C''$. That is, Condorcet's paradox cannot occur.

4. *Independence of irrelevant alternatives in ranking* (IIAR). When $C \succeq_S C'$ for the profile $\Phi$, $C \succeq_S C'$ for any profile $\Phi'$ obtained by eliminating or adjoining other competitors (or rows). That is, Arrow's paradox cannot occur.

These four are the rock-bottom necessities in the theory developed here. They are basic to Arrow's theory (Arrow 1951), the recent method of Dasgupta-Maskin (Dasgupta and Maskin 2004, 2008), and are central to all debates on voting. Together they severely restrict the choice of a method of ranking.

Definition 1. A method of ranking *respects grades* if the rank order between them depends only on their sets of grades; in particular, when two competitors $C$ and $C'$ have the same set of grades, they are tied.

With such methods the rank orders induced by the voters' grades must be forgotten, only the sets of grades count, not which voter assigned which grade. Said differently, if two voters switch the grades they give a competitor, this has no effect on the electorate's ranking of the competitors.

Theorem 6 (Balinski and Laraki (2010), p. 182). *A method of ranking is impartial, transitive, and independent of irrelevant alternatives in ranking if and only if it is transitive and respects grades.*

This simple theorem is essential: it says that if Arrow's and Condorcet's paradoxes are to be avoided, then the traditional model and paradigm *must* be abandoned. *Who gave what grade cannot be taken into account*. Not only do rank-order inputs not permit voters to express themselves as they wish, but they are the culprits that lead to all of the impossibilities and incompatibilities.

**Table 10.** Electorate's evaluations.

| | *Good* (%) | *Pass* (%) | *Bad* (%) |
|---|---|---|---|
| $C$: | 40 | 35 | 25 |
| $C'$: | 35 | 30 | 35 |

Definition 2. A *social-ranking function* is a method of ranking that is impartial, transitive, and IIAR.

By Theorem 6 such functions must respect grades and so depend only on the grades of each of the competitors. To see more clearly the implications of the theorem—or of using a social-ranking function—suppose there were three grades—*good*, *pass*, and *bad*—and that an electorate evaluated two candidates $C$ and $C'$ as shown in Table 10.

$C$'s percentages of *good* and *pass* are both above $C'$'s, her percentage of *bad* below $C'$'s, so there is no doubt that in the electorate's evaluation $C$ leads $C'$. But what does a majority vote say? That all depends. If the electorate's preference profile is as shown in Table 11 (consistent with the distributions of grades) then $C$ wins with 65% of the votes (assuming a voter gives her vote to the candidate with the higher grade). On the other hand, if the electorate's profile is as in Table 12 (also consistent with the distributions of grades), then $C'$ wins with 60% of the votes. Thus more precise information about voters' evaluations of candidates shows that majority voting and the traditional model may fail even when comparing only two candidates.

The theorem suggests that what is needed is a function that transforms the grades given any competitor into a final grade, the order among the final grades determining the order-of-finish of the competitors. The usual practice, as was mentioned, is to use the average grade (a point-summing method), though sometimes the top and bottom grades, or top two and bottom two grades, are omitted.

Functions that assign a final grade to a competitor based only on the competitor's set of grades should enjoy at least two other properties. First, if the voters all assign the same grade to a competitor it should be his final grade. Second, in comparing two ordered sets of grades, when each in the first set is at least as high as the corresponding grade in the second set, the final grade given the first should be no lower than that given the second; moreover, when each in the first set is strictly higher than the corresponding grade in the second set, the final grade given the first should be strictly higher than that given the second.

**Table 11.** Electorate's possible preference profile.

| | 30% | 10% | 10% | 25% | 25% |
|---|---|---|---|---|---|
| $C$: | *Good* | *Good* | *Pass* | *Pass* | *Bad* |
| $C'$: | *Pass* | *Bad* | *Good* | *Bad* | *Good* |

**Table 12.** Electorate's possible preference profile.

| | 5% | 35% | 35% | 25% |
|---|---|---|---|---|
| $C$: | *Good* | *Good* | *Pass* | *Bad* |
| $C'$: | *Pass* | *Bad* | *Good* | *Pass* |

Definition 3. A function $f\colon \Lambda^n \to \Lambda$ that transforms grades given a competitor into a final grade is a *social-grading function* if it satisfies three properties:

- *Anonymity*: $f(\ldots, \alpha, \ldots, \beta, \ldots) = f(\ldots, \beta, \ldots, \alpha, \ldots)$
- *Unanimity*: $f(\alpha, \alpha, \ldots, \alpha) = \alpha$
- *Monotonicity*:

$$\alpha_j \preceq \beta_j \quad \text{for all } j \quad \Rightarrow \quad f(\alpha_1, \ldots, \alpha_n) \preceq f(\beta_1, \ldots, \beta_n)$$

and

$$\alpha_j \prec \beta_j \quad \text{for all } j \quad \Rightarrow \quad f(\alpha_1, \ldots, \alpha_n) \prec f(\beta_1, \ldots, \beta_n).$$

Social-grading functions serve two separate though related purposes: (1) they assign a final grade to each competitor, and (2) used as social-ranking functions, they determine the order-of-finish of all competitors. Obvious examples of social-grading functions are the arithmetic mean or average, any other mean such as the geometric or harmonic mean, and the $k$th order function $f^k$ that is the $k$th highest grade (for $k = 1, 2, \ldots, n$).

In practice grades are almost always numbers and, since final grades can be determined by functions such as means, a discrete scale of inputs (including word grades that have been assigned numbers) may well yield a richer set of outputs, and this in turn may naturally lead to defining a richer set of input grades. So it is reasonable—and permits a cleaner and more elegant theory—to assume from the outset that grades belong to an interval of the real line. It turns out not to matter whether this interval is open, half-open, bounded, or not: so the choice has been to take the closed interval $[0, R]$ (in keeping with the often used $[0, 100]$ in the United States and $[0, 20]$ in France).

Small changes in the input grades should naturally imply small changes in the outputs, so it is natural to assume that a social-grading function is continuous. Some theorems require continuity (e.g., Theorem 12); some do not but require a sufficiently large finite set of grades (e.g., Theorem 9); some require neither (e.g., Theorem 7). When a voter or judge has no interest in deviating from a particular grade in a rich set of grades, he has no interest in deviating from that same grade in a subset of them. Although continuity or a sufficiently rich set of grades are necessary for some characterizations the properties of the functions that are characterized hold for finite sets of grades. In any case we believe that the same criteria should be used whatever the size of the language of grades.

The question that presents itself is, *Which social-grading function(s) of the grades of competitors should be used to grade and which to rank*?

## 2. Majority Judgment

In addition to treating voters and candidates impartially, nine desirable properties of a method of voting or judging emerge:

1. Determine (generically) a winner, i.e., a transitive order-of-finish (or avoid Condorcet's paradox).
2. Guarantee that the final order between two candidates does not depend on other candidacies, i.e., satisfy IIAR (or avoid Arrow's paradox).
3. Faithfully represent voters' opinions in practice.
4. Elicit honest voting (make an honest vote be good strategy).
5. Use measures meaningfully (add numbers whose sums make sense).
6. Resist manipulation (minimize the possibility of successful cheating).
7. Heed the majority's will (seek true consensus).
8. Elect the Condorcet winner when he exists (be *Condorcet consistent*).
9. Ensure that a voter always helps his favorite candidate (no "no-show paradox").

Although majority judgment satisfies some desirable properties, it fails to satisfy others: "Nothing is perfect. There are lumps in it." It is the one method that satisfies the first seven properties. It may be characterized mathematically in several ways, among them by 1, 2, and 6 and also by 1, 2, 5, and 7.

There exists no method that meets 1, 2, and 8 (Balinski and Laraki 2010). The question becomes, which lumps are important? We believe that the Condorcet and Arrow lumps (1 and 2) are very important. The Arrow paradox occurs frequently (with potentially dramatic effects, as was seen). The Condorcet paradox has not often been recorded in elections because it is rarely possible to do so—voters' inputs in real elections are rarely rank orders—but it is essential to determine a winner, and when judges' rank orders are known it occurs (e.g., skating and wines (Balinski and Laraki 2013a)).

One of the two critiques of majority judgment is its potential violation of Condorcet consistency (point 8): we argue below that it is of little importance. The second critique is that it violates 9, the no-show paradox (Felsenthal and Machover 2008). First, there exists no method that is Condorcet consistent *and* avoids the no-show paradox (Moulin 1988). Second, the only methods that satisfy 1, 2, and 9 violate 4, 5, 6, 7, and 8 (Balinski and Laraki 2010). Third, in practice majority judgment violates 9 less often than a tie in ordinary majority voting. These points are discussed more fully below.

### 2.1. Majority Judgment: Description

Suppose there are $n$ judges or voters who assign competitors grades.

Definition 4. The $k$th *order function* $f^k$ is the social-grading function whose value is the $k$*th* highest grade. When the set of grades $\mathbf{r}$ of a competitor is ordered from highest to lowest,

$$\mathbf{r} = (r_1 \succeq r_2 \succeq \cdots \succeq r_n) \quad \Rightarrow \quad f^k(\mathbf{r}) = r_k.$$

Definition 5. A competitor's *majority grade* $f^{\text{maj}}$ is the grade that obtains an absolute majority of the voters against any lower grade and an absolute majority or a tie against any higher grade: it is his middlemost or median grade when $n$ is odd, his lower middlemost when $n$ is even:

$$f^{\text{maj}} = \begin{cases} f^{(n+1)/2} & \text{if } n \text{ is odd,} \\ f^{(n+2)/2} & \text{if } n \text{ is even.} \end{cases}$$

A majority grade is not a median because there is no median when $n$ is even. A separate term is needed. The lower middlemost rather than the upper middlemost is taken for two reasons. First, it insists on an absolute majority for a high grade rather than for a low one. Second, it is the logical consequence of "respecting consensus" (developed below), which in essence comes down to this. If two competitors have two grades, $A$ with 10.9 and 9.8, $B$ with 10.5 and 10.2, that candidate whose grades are more consensual should lead, so $B$ should lead $A$.

In the following the judges' scores from Table 1 are interpreted as the grades of a finite common language (going from 0 to 12 in tenths). Ordering each competitor's grades from highest to lowest gives Table 13. The order-of-finish of the competitors is determined by their majority grades. In this case there is a three-way tie for third place. So a finer distinction is needed. If two competitors such as Savoie and Honda have the same majority grade, then the order between them must depend on their sets of grades excluding that one common grade. So it is dropped, and the majority grades of the remaining eight grades are determined. In this case Savoie's is 10.8, Honda's is 10.7: Savoie's is higher, so he leads Honda by majority judgment.

**Table 13.** Competitors' scores ordered from highest to lowest (identities of judges forgotten).

| | $f^1$ | $f^2$ | $f^3$ | $f^4$ | $f^{\text{maj}}$ | $f^6$ | $f^7$ | $f^8$ | $f^9$ |
|---|---|---|---|---|---|---|---|---|---|
| T. Eldredge | 11.7 | 11.6 | 11.5 | 11.4 | *11.4* | 11.4 | 11.3 | 11.3 | 11.2 |
| C. Li | 11.2 | 11.2 | 11.0 | 11.0 | *10.9* | 10.9 | 10.8 | 10.8 | 10.6 |
| M. Savoie | 11.1 | 11.1 | 11.1 | 10.8 | *10.8* | 10.8 | 10.6 | 10.5 | 10.5 |
| T. Honda | 11.2 | 11.0 | 10.9 | 10.9 | *10.8* | 10.7 | 10.4 | 10.3 | 10.3 |
| M. Weiss | 11.1 | 10.9 | 10.9 | 10.9 | *10.8* | 10.6 | 10.6 | 10.4 | 10.4 |
| Y. Tamura | 11.6 | 11.0 | 10.8 | 10.8 | *10.7* | 10.6 | 10.4 | 10.1 | 09.8 |

*Note.* Majority grades are italicized.

In general, suppose a competitor's grades are

$$r_1 \succeq r_2 \succeq \cdots \succeq r_n.$$

Her *majority value* is an ordered sequence of these grades. The first in the sequence is her majority grade; the second is the majority grade of her grades when her (first) majority

**Table 14.** First-past-the-post, French 2012 presidential election, OpinionWay poll.

| Martine Aubry | Marine Le Pen | Nicolas Sarkozy | François Bayrou | Jean-Louis Borloo | Eva Joly | Jean-Luc Mélenchon | Dominique de Villepin | Olivier Besancenot | Jean-Pierre Chevènement | Nicolas Dupoint-Aignan | Nathalie Arthaud |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.7% | 20.6% | 19.1% | 8.5% | 7.8% | 7.4% | 4.2% | 3.7% | 2.9% | 1.9% | 1.4% | 0.8% |

*Note.* Conducted April 6–7, 2011.

grade has been dropped (it is her "second majority grade"); the third is the majority grade of her grades when her first two majority grades have been dropped; and so on. Thus, when there is an odd number of voters $n = 2t - 1$, a competitor's *majority value* is the sequence that begins at the middle, $r_t$, and fans out alternately from the center starting from below:

$$\vec{\mathbf{r}} = (r_t, r_{t+1}, r_{t-1}, r_{t+2}, r_{t-2}, \ldots, r_{2t-1}, r_1).$$

When there is an even number of voters $n = 2t - 2$, the majority-value begins at the lower middle and fans out alternatively from the center starting from above:

$$\vec{\mathbf{r}} = (r_t, r_{t-1}, r_{t+1}, r_{t-2}, r_{t+2}, \ldots, r_{2t-2}, r_1).$$

If the majority values of two competitors $A$ and $B$ are, respectively, $\vec{\mathbf{r}}_A$ and $\vec{\mathbf{r}}_B$, the *majority ranking* $\succ_{\text{maj}}$ is defined by

$$A \succ_{\text{maj}} B \text{ when } \vec{\mathbf{r}}_A \succ_{\text{lexi}} \vec{\mathbf{r}}_B,$$

where $\succ_{\text{lexi}}$ means lexicographically greater, i.e., the first grade where $\vec{\mathbf{r}}_A$ and $\vec{\mathbf{r}}_B$ differ $A$'s is higher. The majority ranking in the skating competition is

$$\text{Eldredge} \succ_{\text{maj}} \text{Li} \succ_{\text{maj}} \text{Savoie} \succ_{\text{maj}} \text{Honda} \succ_{\text{maj}} \text{Weiss} \succ_{\text{maj}} \text{Tamura}.$$

*There can be a tie only if two competitors have precisely the same set of grades.*

A key point should be noted. Consider any judge or set of judges who assigned a competitor a grade higher than his majority grade; e.g., Honda's majority grade is 10.8 and four judges—$J_2$, $J_3$, $J_4$, $J_6$—believed he merited a higher grade: neither one of them nor all of them acting together can raise his majority grade by changing the grades they assigned. Symmetrically, four judges—$J_1$, $J_7$, $J_8$, $J_9$—believed he merited a lower grade: neither one of them nor all of them acting together can do anything to lowering his majority grade. The best strategy of a judge who wishes that a competitor be awarded a particular majority grade is to assign him that grade: honesty is the best policy.

**2.1.1. Majority Judgment with Large Electorates in Use.** With many voters and few grades it is almost certain that a candidate's middlemost grade will be repeated many times. Thus, an absolute majority of voters assign a candidate at least her majority grade, and also an absolute majority of voters assign the candidate at most her majority grade. Moreover, a simplified procedure is almost sure to determine the majority ranking.

Majority judgment has been tested in several political settings (Balinski and Laraki 2010, 2011; Terra Nova 2011; *Rue89* 2011; *Slate.fr* 2011, 2012). Terra Nova (a Paris-based think tank) sponsored a poll conducted by the national polling agency OpinionWay on April 6 and 7, 2011. Entitled "And if the presidential election of 2012 used Majority Judgment," a sample of 1,025 persons 18 years old or above and representative of the French population was questioned. The results concern 991 of them who were registered voters and responded to at least one question. French presidential elections use *two-past-the-post*: a voter names (or ticks) at most one candidate, the candidate most often named (or with the most ticks) wins if she obtains an absolute majority; otherwise, there is a run-off between the two candidates most often named. The central interest of this poll is that the identical set of people voted with the usual method and with majority judgment, permitting comparisons.

*The poll's first question*:[10] "If the first round of the 2012 presidential elections were to be held next Sunday, for which of the following candidates would you most likely vote?" The answers are given in Table 14.

*The poll's second question*: "If the second round of the 2012 presidential elections were to be held next Sunday, for which of the following candidates would you most likely vote for?" The answers in each of three possible run-offs are shown in Table 15.

Voting measures to determine winners and orders-of-finish. The results show majority voting—in one round or two—measures badly. The incumbent president Sarkozy is eliminated, yet he would easily defeat Le Pen. Since the poll admits a 2 to 3% error, any one of the three leading finishers could be eliminated—including the candidate truly wished by the electorate (with first- or two-past-the-post). It seems that the electorate's choice is among the major candidate of the left Aubry, the extreme rightist Le Pen and the major candidate of the right Sarkozy, the remaining candidates being relegated to minor roles. It also seems clear

**Table 15.** Head-to-head votes, French presidential election, OpinionWay poll.

| Runoff 1 | | Runoff 2 | | Runoff 3 | |
|---|---|---|---|---|---|
| Aubry | Le Pen | Aubry | Sarkozy | Sarkozy | Le Pen |
| 63.2% | 36.8% | 56.0% | 44% | 63.3% | 36.7% |

**Table 16.** Majority judgment ballot, French 2012 presidential election, OpinionWay poll.

| Election of the President of France 2012<br>*Having taken into account all relevant considerations, I judge, in conscience, that as President of France each of these candidates would be*: | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Outstanding* | *Excellent* | *Very Good* | *Good* | *Acceptable* | *Poor* | *To reject* |
| Candidate | | | | | | | |

*Note.* You must check one single grade in the line of each candidate. Conducted April 6–7, 2011 (one line for each candidate).

that the strategy of the major candidates of left (Aubry) and right (Sarkozy) is to encourage multiple candidacies in the opposite camp to dilute their total vote in the hopes of a runoff against Le Pen.

*The poll's third question* asked participants to vote with majority judgment using the ballot given in Table 16. The results are given in Table 17. In this poll: (1) three of every four voters accord no *outstanding*; (2) half of the voters accord no *outstanding* and no *excellent*, (3) one of every five voters only assign *acceptable* or below, (4) one of every five voters give their highest grade to two candidates, and (5) one of every five voters give their highest grade to at least three candidates. This behavior is by and large consistent with that observed every time majority judgment is used in political elections. It shows the inadequacy of the traditional inputs, be they rankings or ticks.

A simpler procedure than finding the candidates' majority values determines the majority ranking. Suppose a candidate's majority grade is $\alpha$, and that $p\%$ of his grades are higher than $\alpha$ and $q\%$ are lower. Then his *majority gauge* is $(p, \alpha\pm, q)$, where $p > q$ implies $\alpha$ is endowed with a "+," and otherwise it is endowed with a "−." Thus Aubry's majority gauge is $(38.0\%, \mathit{Good}-, 49.3\%)$.

The majority gauges $(p, \alpha\pm, q)$ determine the *majority ranking* of the candidates (see Table 18). If two candidates have the same majority grade $\alpha$ (ignoring for the moment the signs), four sets of voters disagree. Their respective sizes are measured by the two candidates' $p$'s and $q$'s. The largest set decides if it is a $p$ then that candidate leads the other, if it is a $q$ then that candidate trails the other. For example, Sarkozy and Chèvenement both are *poor*, the largest set is the 46.9% for a higher grade for Sarkozy, so he leads; Mélenchon and Besancenot both are *poor* as well, the largest set is the 44.2% for a lower grade for Besancenot, so he trails. Notice that the rule implies a candidate with $\alpha+$ leads a candidate with $\alpha-$. When the majority gauge distinguishes between two candidates, it necessarily agrees with the majority ranking. With some 20 voters or more and a common language of some six grades, the majority gauge (not the more precise majority value) has sufficed to distinguish candidates in all uses to date.

With simple majority voting Martine Aubry appears to be in the lead by a margin so small that it is statistically insignificant. Majority judgment shows that she leads comfortably in the esteem of the electorate: she has more *outstandings* and *excellents*, fewer *to rejects*. The extreme rightist Marine Le Pen is a very close second according to simple majority voting, but that only takes into account her supporters. Majority judgment shows that a large majority reject her out of hand, so that she has no chance of winning whatsoever (in the climate of opinion of early April 2011). This is, of course, verified by the face-to-face scores of the poll. Candidates Borloo and Villepin of the moderate right wing Union pour un Mouvement Populaire (UMP) (Sarkozy's party) and the centrist Bayrou—low in the majority voting ranking with a third of the votes of Sarkozy or less—are shown by majority judgment to be in fact favored to Sarkozy. Their majority grades are all *acceptable*+, placing them much higher than Sarkozy and Le Pen. Majority judgment takes all of a candidate's grades into account—the good ones

**Table 17.** Results, French 2012 presidential election, OpinionWay poll.

| | *Outstanding* (%) | *Excellent* (%) | *Very Good* (%) | *Good* (%) | *Acceptable* (%) | *Poor* (%) | *To reject* (%) |
|---|---|---|---|---|---|---|---|
| Arthaud | 0.1 | 00.9 | 03.3 | 07.7 | 13.7 | 26.1 | 48.0 |
| Besancenot | 0.8 | 01.7 | 06.9 | 09.9 | 16.1 | 20.4 | 44.2 |
| Mélenchon | 1.3 | 02.7 | 05.0 | 11.2 | 16.5 | 21.4 | 41.8 |
| Joly | 3.2 | 04.7 | 07.4 | 14.5 | 20.3 | 19.0 | 30.9 |
| Aubry | 8.2 | 12.9 | 17.0 | 12.6 | 19.6 | 11.4 | 18.4 |
| Chevènement | 0.5 | 01.1 | 05.8 | 12.9 | 22.8 | 24.7 | 32.2 |
| Bayrou | 1.2 | 04.7 | 12.8 | 19.2 | 26.1 | 16.6 | 19.3 |
| Borloo | 2.2 | 06.2 | 15.3 | 22.3 | 19.6 | 15.9 | 18.5 |
| Villepin | 2.0 | 05.8 | 11.9 | 20.4 | 20.7 | 17.4 | 21.9 |
| Sarkozy | 4.1 | 08.7 | 11.1 | 09.5 | 13.5 | 11.8 | 41.3 |
| Dupont-Aignan | 0.5 | 01.4 | 02.7 | 07.0 | 13.9 | 27.7 | 46.7 |
| Le Pen | 6.8 | 06.5 | 07.0 | 07.2 | 07.8 | 09.3 | 55.6 |

*Note.* Conducted April 6–7, 2011.

**Table 18.** Majority gauges and majority ranking, French 2012 presidential election, OpinionWay poll.

| Majority judgment ranking | Majority gauge | | | Majority vote ranking |
|---|---|---|---|---|
| | $p$ (%) | $\alpha\pm$ | $q$ (%) | |
| 1. Martine Aubry | 38.0 | *Good*− | 49.3 | 1 |
| 2. Jean-Louis Borloo | 46.0 | *Acceptable*+ | 34.4 | 5 |
| 3. Dominique de Villepin | 40.1 | *Acceptable*+ | 39.3 | 8 |
| 4. François Bayrou | 37.9 | *Acceptable*+ | 35.9 | 4 |
| 5. Eva Joly | 29.9 | *Acceptable*− | 49.8 | 6 |
| 6. Nicolas Sarkozy | 46.9 | *Poor*+ | 41.3 | 3 |
| 7. Chèvenement | 43.1 | *Poor*+ | 32.2 | 10 |
| 8. Mélenchon | 36.8 | *Poor*− | 41.8 | 7 |
| 9. Besancenot | 35.4 | *Poor*− | 44.2 | 9 |
| 10. Dupont-Aignan | 25.5 | *Poor*− | 46.7 | 11 |
| 11. Nathalie Arthaud | 25.8 | *Poor*− | 48.0 | 12 |
| 12. Marine Le Pen | 44.4 | *to Reject* | — | 2 |

*Note.* Conducted April 6–7, 2011.

and the bad ones—to determine her place in the ranking. In contrast, simple majority voting takes into account only a mixture of supposedly favorable opinions.

One of the very interesting outcomes of this poll is the consistently low majority grades of all candidates. Seven candidates are judged to be *poor* or worse, including the incumbent president, Sarkozy; and only one candidate, Aubry, is judged *good*. This is an echo of the disregard for politicians regularly reported by opinion polls and also of the fact that the campaign had not yet begun.

Notice that voters who believed Borloo (for example) merited a higher majority grade than *acceptable*—and 46.0% were of that persuasion—could do nothing alone or in concert to raise his majority gauge. Symmetrically, those who believed he merited a lower majority grade—34.4% of them—could do nothing alone or in concert to lower his majority gauge. The best strategy of a voter who wishes that a candidate be awarded a particular majority grade is to assign him that grade.

## 2.2. Majority Judgment: Theory

When a social-grading function is used to amalgamate the grades voters or judges assign competitors, and the grades determine the order-of-finish of the competitors, the Condorcet and Arrow paradoxes cannot occur—transitivity is assured and there can be no flip-flops—as has been proven. Since the grades have only ordinal significance and are neither summed nor averaged, the method is meaningful. Thus three of the seven essential demands are necessarily met.

**2.2.1. Elicits Honesty.** Assigning grades to competitors is a game played by voters or judges. As early as 1907 Sir Francis Galton pointed out that when a jury is to decide on an amount of money—e.g., to allocate to a project, or in assessing damages in an insurance claim—"that conclusion is clearly *not* the *average* of all the estimates, which would give a voting power to 'cranks' in proportion to their crankiness. . . . I wish to point out that the estimate to which least objection can be raised is the *middlemost* estimate, the number of votes that it is too high being exactly balanced by the number of votes that it is too low" (our emphasis), (Galton 1907). He realized that point-summing methods do not elicit honesty (equivalently, that one extreme assignment of points or one extreme money estimate can completely alter the collective outcome). The idea of using the median in voting is Galton's.

The strategy a voter adopts depends on her personal likes and dislikes. Some voters and judges may care most about assigning the grades they believe are truly merited. Some may care most about the final grades assigned each competitor—and are ready to adjust their assignments so as to attain that end. Others may not care at all about the final grades but only about the order-of-finish of the competitors. Still others may think that only the identity of the winner is of importance. Some few may be bought or bribed. Some other few may simply be completely incompetent judges who assign unwarranted grades. The final grade a voter wishes a competitor to be awarded, the final grade he believes the competitor merits, and the grade he gives may all be different. Some juries and electorates almost certainly include judges and voters who honestly wish grades to be assigned according to merit, and in certain cases it is perfectly reasonable to assume that all the players share this intent. Nevertheless, a very complex set of unknown wishes, opinions, expectations, and anticipations—the voters' or judges' *utility functions*—determines the grades they give.

How is a social-grading function to elicit honesty? By making it impossible or difficult for individual voters to change the outcome by using devious strategies.

Definition 6. Suppose that a competitor's final grade is $r^*$. A social-grading function is *strategy-proof-in-grading* if, when a voter's input grade is higher than the final grade, $r^+ > r^*$, any change in his input can only lead to a lower final grade; and if, when a voter's input grade is lower than the final grade, $r^- < r^*$, any change in his input can only lead to a higher final grade.

It is easy to see that the majority judgment is not only strategy proof in grading but also *group strategy proof in grading* in that a group whose inputs are higher (or lower) than the final grade can only lower (raise) the final grade. Thus, one or *all* of those who gave Aubry a grade above her majority grade (*good*) cannot change her majority grade or gauge *except* to lower it (presumably not their intention). Similarly, one or *all* of those who gave her a grade below her majority grade cannot change her majority grade or gauge *except* to raise it (presumably not their intention).

Assume the more a final grade deviates from the grade a voter wishes it to be the less she likes it ("single-peaked preferences over grades"), so that the voter's utility function $u_j(\mathbf{r}^*, \mathbf{r}, f, \mathscr{C}, \Lambda)$ could be a monotonic transformation of $|r_j^* - f(r_1, \ldots, r_n)|$. Then—as is well known in the work on single-peaked preferences (e.g., Galton 1907, Black 1958,

Moulin 1980)—it is a *dominant strategy* for her to assign the grade she believes is merited: i.e., it is at least as good as any other strategy and strictly better in some cases.

Theorem 7 (Moulin 1980; Balinski and Laraki 2010, pp. 191–192). *The unique strategy-proof-in-grading social-grading functions are the order functions* (*for a finite or an infinite number of grades, continuous or discontinuous functions*).[11]

A competitor who receives a higher majority grade than another is naturally ranked higher in the order of the candidates or alternatives than the other: grades imply orders. But when an important component of the voters' utilities are the orders of finish and not merely the final grades of competitors, their strategic behavior may well alter.

Given a profile of grades $(r_j^C)$, $C \in \mathscr{C}$, and $j \in \mathscr{J}$ with $r_j^C \in [0, R]$, let the vector of final grades be $(r^C)$. Suppose the final grades of some two competitors $A, B \in \mathscr{C}$ are $r^A < r^B$, but that some voter $j$ is of the opposite conviction, $r_j^A > r_j^B$. She would like either to increase $A$'s final grade, or decrease $B$'s final grade, or better yet do both.

Definition 7. When the final grade of $A$ is lower than that of $B$, $r^A < r^B$, and any voter $j$ is of the opposite conviction, $r_j^A > r_j^B$, a social-ranking function is *strategy proof in ranking* if $j$ can neither decrease $B$'s final grade nor increase $A$'s final grade.

Consider a voter $j$ whose utility function $u_j$ depends only on the ultimate ranking of the competitors, that is, only on the order of the final grades. Then if the social-ranking function is strategy proof in ranking, it is a dominant strategy for voter $j$ to assign grades according to his convictions since it serves no earthly purpose to do otherwise.

Theorem 8 (Balinski and Laraki 2010, p. 220). *There exists no social-ranking function that is strategy proof in ranking.*

This is the analog of the Gibbard-Satterthwaite theorem. But the impossibility of perfection does not deny a search for a best possible.

Definition 8. A social-ranking function is *partially strategy proof in ranking* when $r^A < r^B$ and any voter $j$ is of the opposite persuasion, $r_j^A > r_j^B$, then if $j$ can decrease $B$'s final grade he cannot increase $A$'s final grade, and if he can increase $A$'s final grade he cannot decrease $B$'s final grade.

Theorem 9 (Balinski and Laraki 2010, p. 222). *The unique social-ranking functions that are partially strategy proof in ranking are the order functions.*

In elections with many voters (say in the hundreds and above) the majority-gauges $(p, \alpha\pm, q)$ of the candidates almost always determine the majority-ranking since ties among them almost never occur. Observe that it too is partially strategy proof in ranking. To see this, consider the French presidential poll (Table 17). Aubry, with a majority gauge of (38.0%, *good*−, 49.3%), leads Borloo, whose majority gauge is (46.0%, *acceptable*+, 34.4%). How could a voter who prefers Borloo to Aubry manipulate the outcome? Suppose she could increase Borloo's majority gauge. Then she gave to Borloo at most an *acceptable*, so to Aubry a lower grade, implying she can do nothing to decrease Aubry's majority gauge. If, on the other hand, she could decrease Aubry's majority gauge, then she gave Aubry at least a *good*, so to Borloo a higher grade, implying she can do nothing to increase Borloo's majority gauge. The same argument carries over to groups acting together.

**2.2.2. Meaningfulness.** In the spirit of measurement theory social-grading and social-ranking functions must be *meaningful*: the particular representation that is used should make no difference in the ultimate outcomes. By way of an analogy, distance in the absolute and in comparisons should not change the ultimate outcomes when the scale is meters rather than yards. The only meaningful scales of grades in the new model, as has been argued, are ordinal.

Definition 9. A social-grading function $f$ is *language consistent* if

$$f(\phi(r_1), \ldots, \phi(r_n)) = \phi(f(r_1, \ldots, r_n))$$

for any increasing, continuous transformation $\phi$ of the grades of each voter.

For example, when a Franco-American jury assigns grades to students, and each member is asked to give a grade in both of the languages—the French in the range $[0, 20]$ and the American in the range $[0, 100]$—language consistency asks that the aggregate French grades rank the students in the same order as the aggregate American grades. A transformation that does this is not linear, because 50 is failing in the United States whereas 10 is passing in France.

Order functions are clearly language consistent: the $k$th highest grade remains the $k$th highest grade under increasing, continuous transformations. It is well known that the reverse is true as well:

Theorem 10 (Orlov 1981; Balinski and Laraki 2010, p. 201). *The unique social-grading functions that are language consistent are the order functions.*

To be meaningful as a social-ranking function, the analogous property must hold for rankings as well.

Definition 10. A social-ranking function $\succeq_S$ is *order consistent* if the order between any two candidates for some profile $\Phi$ implies the same order for any profile $\Phi'$ obtained from $\Phi$ by any increasing, continuous transformation $\phi$ of the grades of all voters.

The order functions are clearly order consistent. To characterize them requires an additional, eminently acceptable property, namely, that an increase in a candidate's grade necessarily helps.

Definition 11. A social-ranking function $\succeq_S$ is *choice monotone* if $A \succeq_S B$ and a judge increases the grade of $A$ implies $A \succ_S B$.

Note in passing that the traditional model's difficulties with monotonicity are completely eliminated. Majority judgment is at once choice monotone, rank monotone, and strongly monotone. The reason is simple: a change in heart concerning one candidate is expressed by the grade he is given, but that changes nothing in the inputs concerning the other candidates.

Theorem 11 (Hammond 1976; d'Aspremont and Gevers 1977; Balinski and Laraki 2010, pp. 204, 303). *The unique choice-monotone and order-consistent social-ranking functions are the lexi-order functions.*

A *lexi-order social-ranking function* is a permutation $\sigma$ of the order functions $f^\sigma = (f^{\sigma(1)}, \ldots, f^{\sigma(n)})$, that ranks the candidates by

$$A \succ_S B \quad \text{if } (f^{\sigma(1)}(A), \ldots, f^{\sigma(n)}(A)) \succ_{\text{lex}} (f^{\sigma(1)}(B), \ldots, f^{\sigma(n)}(B)).$$

Here $\succ_{\text{lex}}$ means the lexicographic order: the first term where the corresponding grades differ $A$'s is higher. There are $n!$ lexi-order social-ranking functions. The idea is simple: some order function decides; if it does not because there is a tie, a second order function is invoked; if there is a tie in the second order function, a third is called upon; and so on. The importance of Arrow's impossibility becomes crystal clear in this context.

Definition 12. A social-ranking function is *preference consistent* if the order between any two candidates for some profile $\Phi$ implies the same order for any profile $\Phi'$ obtained from $\Phi$ by increasing, continuous transformations $\phi_j$ of the grades of each voter $j$.

For voters' rank orders to be meaningfully amalgamated, there must exist a preference-consistent social-ranking function. But Arrow's theorem tells us that there exists no monotonic preference-consistent social-ranking function. *It says that there is no meaningful way of amalgamating the voters' inputs when they have no common language.* This—in our opinion—is the deep enduring significance of Arrow's theorem (rather than the supposed impossibility of surmounting Arrow's paradox). That should come as no surprise: how can agreement be found among persons who cannot communicate!

Once again, only the order functions will do: they alone are meaningful. But why the majority grade and why the majority value?

**2.2.3. Resists Manipulation.** To manipulate successfully, a voter (or judge) must be able to raise or to lower a candidate's (or competitor's) final grade by changing the grade he assigns. In some situations voters can only change a final grade by increasing his grade, in others only by decreasing it. Voters who can both lower and raise the final grade have a much greater possibility of manipulating: an outsider seeking to bribe or otherwise influence the outcome would surely wish to deal with such voters.

Theorem 12 (Balinski and Laraki 2010, p. 195). *Order functions are the unique social-grading functions for which at most one voter may both increase and decrease a final grade.*

Given a social-grading function $f$ and a profile of a candidate's grades $\mathbf{r} = (r_1, \ldots, r_n)$, let $\mu^-(f(\mathbf{r}))$ be the number of voters who can decrease the final grade, $\mu^+(f(\mathbf{r}))$ be the number of voters who can increase the final grade, and $\mu(f(\mathbf{r})) = \mu^-(f(\mathbf{r})) + \mu^+(f(\mathbf{r}))$. Take the measure of manipulability $\mu$ of a social-grading function $f$ to be the worst that can happen, $\mu(f) = \max_{\mathbf{r}} \mu(f(\mathbf{r}))$, so $\mu(f) \leqslant 2n$. It is easily verified that $\mu(f^k) = n+1$ for any order function $f^k$. By way of contrast, for $f$ a point-summing method $\mu(f) = 2n$.

In fact, the only social-grading functions $f$ for which $\mu(f) \leqslant n+1$ are the order functions. For assume $\mu(f) \leqslant n+1$ and take any $\mathbf{r}$. By monotonicity and anonymity, if some voter can decrease (respectively, increase) the final grade then any voter giving at least (at most) that grade can also decrease (increase) the final grade. So, if more than one voter can both increase and decrease the final grade, $\mu(f) \geqslant n+2$, a contradiction. Therefore, at most one voter can both increase and decrease the final grade, implying $f$ must be an order function.

Take $\lambda$ to be the probability a briber wishes to increase the grade and $1-\lambda$ that he wishes to decrease the grade. Assume that judges have an equal probability to cheat. A social-grading function is sought that minimizes the probability that a voter may be found who can effectively raise or lower the grade in the worst case.

Definition 13. The *probability of cheating* $Ch(f)$ with $f$ is defined to be

$$Ch(f) = \max_{\mathbf{r}=(r_1,\ldots,r_n)} \max_{0\leqslant\lambda\leqslant 1} \frac{\lambda\mu^+(f(\mathbf{r})) + (1-\lambda)\mu^-(f(\mathbf{r}))}{n}.$$

What social-grading functions minimize the probability of cheating?

Definition 14. A social-grading function $f$ is *middlemost* if for $r_1 \geqslant \cdots \geqslant r_n$,

$$f(r_1, \ldots, r_n) = r_{(n+1)/2} \quad \text{when } n \text{ is odd, and}$$

$$r_{n/2} \geqslant f(r_1, \ldots, r_n) \geqslant r_{(n+2)/2} \quad \text{when } n \text{ is even.}$$

When $n$ is odd, there is exactly one such function, $f^{(n+1)/2}$. When $n$ is even, there are infinitely many; in particular, $f^{n/2}$ is the *upper middlemost* and $f^{(n+2)/2}$ is the *lower middlemost*.

To say a social-grading function $f$ *depends only on the middlemost interval* means that $f(r_1, \ldots, r_n) = f(s_1, \ldots, s_n)$ whenever the middlemost interval of the grades $\mathbf{r} = (r_1, \ldots, r_n)$ and the grades $\mathbf{s} = (s_1, \ldots, s_n)$ is the same.

Theorem 13 (Balinski and Laraki 2010, p. 229). *The unique social-grading functions $f$ that minimize the probability of cheating $Ch(f)$ are the middlemost that depend only on the middlemost interval.*

When $f$ is the max or the min order function, or the average function, the probability of cheating is maximized: $Ch(f) = 1$. When $f$ is a middlemost order function, $Ch(f) \approx \frac{1}{2}$. In this sense, the middlemost cut cheating by half.

The unique meaningful social-ranking functions are the lexi-order functions, each a sequence of all $n$ order functions that determines the final ranking of the candidates. Which among the $n!$ of them minimize cheating?

To determine the ranking between any two candidates, the first order function decides, unless there is a tie; in which case the second order function decides, unless there is a tie; in which case the third decides, unless there is another tie; and so on. The need to use each succeeding order function becomes increasingly rarer. Accordingly, it is of the first importance to minimize the probability of cheating in the first order function: by Theorem 11 this is accomplished by choosing an order function that is in the (first) middlemost interval: it is unique if $n$ is odd and one of two if $n$ is even, namely, $f^{(n+1)/2}$ when $n$ is odd and either the upper middlemost $f^{n/2}$ or the lower middlemost $f^{(n+2)/2}$ when $n$ is even. Given that choice, there are now $n-1$ order functions to choose from and the first importance to minimize the probability of cheating is once again to take a middlemost of those that remain: it is either unique or one of two. Given the first two choices, there are $n-2$ to choose from, a middlemost must again be taken, and so on iteratively. To see this more clearly, consider a finite language of number grades going from a high of 10 to a low of 0 and a candidate who receives the seven grades $\{10, 9, 7, 6, 4, 3, 2\}$. The first order function of a lexi-order function that minimizes the chance of cheating is the middlemost; in this case its value is 6. The second that minimizes the chance of cheating is either the upper middlemost or the lower middlemost; in this case its value is 7 or 4. If it is the upper middlemost (its value 7) the next middlemost is unique (with value 4); if it is the lower middlemost (its value 4) the next middlemost is unique (with value 7).

One may consider, by way of a practical illustration, how the judges might try to manipulate the outcome to obtain what they believe is a better order-of-finish by falsifying their grades in the skating competition (see Tables 1, 2, and 13). It is assumed that the grades they gave are honest, and their utility functions on the order-of-finish is lexicographic: what matters most to each judge is the winner, next the second place skater, and so on.

What effective manipulations can judges pursue? $J_1$, for example, would like Savoie in second place, Li in third. He gave Savoie (with majority-grade 10.8) an 11.1: raising Savoie's grade accomplishes nothing. He gave Li (with majority-grade 10.9) a 10.8: lowering Li's grade accomplishes nothing. $J_1$ would like Weiss in fourth place, Honda in fifth. He cannot lower Honda below anyone. He can place Weiss in fourth place by increasing his grade to 10.7; but if he increased it to 10.8, Weiss would leap ahead of Savoie, not at all his intention.

Similar detailed analyses of the possible actions of each judge may be summarized as follows. All judges are contented with the first place of Eldredge. None can change Li's second place; the only effective manipulations concern skaters in third place or below. Two judges can do nothing ($J_3, J_7$); one can realize his preferred order-of-finish by moving his candidate for fifth place from third place to fifth place ($J_2$); four can invert the order of two consecutive skaters in the order-of-finish ($J_1, J_4, J_5, J_9$); one can move his candidate for second place from sixth place to fourth place ($J_6$); and one can move his candidate for third place from sixth place to third place ($J_8$). This comparison with point summing assumes that judges only care about the order-of-finish, which is almost certainly false, for they are likely to give importance to the absolute final scores of the skates if not other considerations as well. Proven in theory, practice confirms that majority judgment is much better at resisting manipulation than point summing and so also in eliciting honesty.

There are some $2^{n/2}$ lexi-order functions that minimize the chance of cheating. Which among *them* should be chosen?

**2.2.4. Heeds the Majority's Will.** The basic idea—a candidate's majority-grade—is firmly based on the majority's will: it is the highest grade $\alpha$ that commands an absolute majority in answer to the question: "Does this candidate merit at least an $\alpha$?" Moreover, the unique social-grading functions that assign a candidate the final grade $\alpha$ if a majority of voters assign her $\alpha$ are the middlemost functions. But when there are many voters and a language of relatively few grades the two middlemost order functions will (almost always) have one value, the majority grade.

Another basic collective decision idea—a kind of "unanimity"—also singles out the majority-grade $f^{\text{maj}}$ among the social-grading functions.

Definition 15. A social-grading function *respects consensus* when all of $A$'s grades strictly belong to the middlemost interval of $B$'s grades implies that $A$'s final grade is above $B$'s final grade.[12]

The rationale for this definition is that when a jury is more united on the grade of one alternative than on that of another, the stronger consensus should be respected by the award of a higher final grade. Or, taking Galton's perspective, respecting consensus means denying crankiness by heeding the middle grades rather than the extreme grades. Recall that the majority-grade $f^{\text{maj}}$ is the lower-middlemost order function.

Theorem 14 (Balinski and Laraki 2010, p. 216). *The majority-grade $f^{\text{maj}}$ is the unique middlemost social-grading function that respects consensus.*[13]

A similar concept singles out the majority-ranking $\succ_{\text{maj}}$ among the social-ranking functions. Consider the input grades $r_1 \geqslant \cdots \geqslant r_n$. The first-*middlemost interval* is the middlemost interval previously defined. The second-*middlemost interval* is the middlemost interval when the defining grades of the first-middlemost interval are ignored. The $k$th-*middlemost interval* is the middlemost interval when the defining grades of the previous middlemost intervals are ignored. For example, with the set of grades $\{10, 9, 7, 6, 4, 3, 2\}$ the first-middlemost interval is $[6, 6]$, the second is $[7, 4]$, the third is $[9, 3]$, and the fourth is $[10, 2]$.

Suppose the grades of $A$ and $B$ are $\mathbf{r}^A = (r_1^A, \ldots, r_n^A)$, $\mathbf{r}^B = (r_1^B, \ldots, r_n^B)$.

Definition 16. A social-ranking function is a *middlemost* if $A \succ_S B$ depends only on the set of grades that belong to the first of the $k$th-middlemost intervals where they differ.

For example, if $A$'s grades are those of the example just given and $B$'s are $\{10, 10, 7, 6, 4, 2, 1\}$, then the first interval where they differ is the third: $A$'s is $[9, 3]$ and $B$'s is $[10, 2]$. This is a natural extension of the idea of a middlemost social-grading function that depends only on the middlemost interval.

Definition 17. Suppose the first of the $j$th-middlemost intervals where $A$'s and $B$'s grades differ is the $k$th. A social-ranking function *rewards consensus* when all of $A$'s grades strictly belong to the $k$th-middlemost interval of $B$'s grades implies that $A$ is ranked above $B$, $A \succ_S B$.

Thus, $A$ is ranked above $B$ for the example just given by a social ranking function (SRF) that rewards consensus. This is a natural extension of the idea of respecting consensus for a social-grading function.

Theorem 15 (Balinski and Laraki 2010, p. 228). *The majority-ranking $\succ_{\text{maj}}$ is the unique middlemost, choice-monotone social-ranking function that rewards consensus.*

The choice of the lower middlemost order function for ranking and electing is the consequence of seeking consensus, and completes the characterization of majority judgment.[14]

But majority judgment fails to satisfy some other properties.

**2.2.5. Condorcet Consistency and the Pure Game of Voting.** Majority judgment is not Condorcet consistent. Consider $2k + 1$ judges who are asked to evaluate two competitors, $A$ and $B$, in a scale of grades $[0, 20]$:

| | $k$ judges | 1 judge | $k$ judges |
|---|---|---|---|
| $A$: | $12, \ldots, 12$ | 12 | $4, \ldots, 4$ |
| $B$: | $16, \ldots, 16$ | 8 | $8, \ldots, 8$. |

A first glance suggests $B$ should be the winner. But $A$'s majority grade is 12, $B$'s is 8: $A$—"preferred" by only one voter according to the traditional paradigm—is the majority judgment winner, whereas $B$ is the overwhelming majority winner with $2k$ votes to $A$'s 1.

But why should the majority's will be counted in *comparisons* rather than in *grades*? The majority says $A$ deserves at least 12 and $B$ at most 8 (in admittedly close votes). How and why is this less compelling than that a majority prefers $B$ to $A$? This is an extremely artificial example: under the impartial culture assumption (uniform distribution) the probability that half the voters give to both candidates more than their majority grade is of the order $1/2^k$; moreover, one judge is supremely decisive: she *alone* can determine $A$'s majority grade to be any grade from 4 to 12 and $B$'s any grade from 8 to 16.

The same type of extreme example can be invented to cast doubts about all of the methods of the traditional model including a majority winner:

| | $k$ judges | 1 judge | $k$ judges |
|---|---|---|---|
| $A$: | $20, \ldots, 20$ | 10 | $0, \ldots, 0$ |
| $B$: | $19, \ldots, 19$ | 9 | $19, \ldots, 19$. |

A first glance suggests $B$ should be the winner. Yet $A$ is the majority winner and Condorcet winner (in an admittedly close $k + 1$ to $k$ vote), whereas $B$ is the clear majority judgment winner with a majority grade of 19 to $A$'s 10.

But the most used or advocated methods *all* fail to elect a majority winner or Condorcet winner when she exists:

- Approval voting. Suppose all voters give ticks to candidates whose grade is 10 or better. $A$ wins though $B$ is the majority winner in the first example, $B$ wins though $A$ is the majority winner in the second example.
- Point summing. In the second example $B$ wins though $A$ is the majority winner.
- Borda's method, first- and two-past-the-post, the alternative vote (or IRV). It is well known that they all can fail to elect a Condorcet winner.

In any case, every Condorcet consistent method necessarily admits the Arrow or the Condorcet paradox. When both paradoxes are sure to be avoided, Theorem 6 shows permuting the grades of candidates cannot change the winner. But consider the first example: when the first $k$ judges' 12's for $A$ are exchanged with the last $k$ judges' 4's for $A$, the candidates' sets of grades are the same, yet $A$ becomes the majority winner instead of $B$.

Thus no method currently advocated is Condorcet consistent when the inputs of voters are *honest*. However, when voters are viewed as *purely rational agents* who seek to maximize their expected utilities and the utilities depend *only* on the identity of the winner, the situation is close to the exact opposite. To describe it, several concepts must be introduced.

A candidate $C$ is a *strong-equilibrium winner* if no coalition of voters can deviate from their strategies and thereby elect a candidate they prefer to $C$ (Aumann 1959). This definition of equilibrium is particularly apt for elections

since groups of voters act in concert. A method of election is *majoritarian* if for any candidate $C$, any absolute majority of voters have a strategy that elects $C$ whatever the strategies of the other voters.

THEOREM 16 (BALINSKI AND LARAKI 2010, PP. 353–355). *A candidate $C$ is a strong-equilibrium winner with a majoritarian method if and only if $C$ is a Condorcet winner*; *moreover, only a majoritarian method can implement the Condorcet winner as a strong-equilibrium winner.*

The methods of Condorcet, first- and two-past-the-post, approval voting, the single transferable vote, majority judgment, and point-summing methods are all majoritarian methods, but Borda's is not. Somewhat more may be claimed for majority judgment: there is a strong equilibrium that elects the Condorcet winner $C$ with her true (honest) majority grade and every candidate is assigned a majority of honest grades (Balinski and Laraki 2010, p. 357). But the upshot is that *any* "reasonable" method elects the Condorcet winner when she exists, showing Condorcet consistency is not a property that convincingly separates the wheat from the chaff, good methods from bad (except for Borda's method).

**2.2.6. The No-Show Paradox.** Majority judgment admits the *no-show paradox*: candidate $A$ wins, a late voter arrives who prefers $A$ to $B$, his vote makes $B$ the winner (alternatively, he leaves). The following example shows it can occur with majority judgment:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $A$'s grades: | 20 | 17 | 15 | 15 | 12 | 11 | 7 |
| $B$'s grades: | 18 | 17 | 16 | 14 | 13 | 10 | 5. |

$A$'s majority grade is 15, $B$'s 14. If a late voter arrives and assigns 6 to $A$ and 4 to $B$ then $A$'s majority grade becomes 12, $B$'s 13, so $B$ wins though the late voter rated $A$ above $B$. The one substantive argument against majority judgment that reviews of the book Balinski and Laraki (2010) have raised is that it admits this paradox (Edelman 2012, Bogomolny 2011, Brams 2011). Is this important? We believe not for many reasons.

First, the no-show paradox is unimportant in a large electorate because it is less likely to occur than a tie in first-past-the-post. And in a small jury the problem does not arise because all judges must participate.

Second, there is an implicit assumption in accepting the very idea of the paradox, namely, that only the winner counts. Note that the voter who assigned 6 to $A$ and 4 to $B$ in the above example exerted an influence: she decreased the final grades of both $A$ and $B$ bringing them closer to her evaluations. Perhaps—since in any case she did not believe either merited high grades—that was more important to her than which of the two wins.

Third, the paradox can occur only when the late voter sees relatively little difference between the two and assigns both low or both high grades:

THEOREM 17 (BALINSKI AND LARAKI 2010, P. 287). *Suppose $A$'s majority grade is $\alpha$, $B$'s is $\beta$, and $A$ is the winner. If a new voter gives $\alpha$ or a higher grade to $A$ and a lower grade than $\alpha$ to $B$—or if she gives $\beta$ or a lower grade to $B$ and a higher grade than $\beta$ to $A$—then $A$ remains the winner and the no-show paradox does not occur.*

Fourth, the no-show paradox is but a particular instance of the violation of a more general property. A method is *participant consistent* if it avoids the no-show paradox. More generally, it is *join consistent* when $A$ is ranked above $B$ in each of two separate parts of an electorate implies $A$ is ranked above $B$ in the combined electorate. Majority judgment is not join consistent (as already shown). But in the context of the traditional model *all* Condorcet-consistent methods violate join consistency (see Balinski and Laraki 2010, pp. 76–79 and Moulin 1988, pp. 237–239). In the new model the only join-consistent methods are point-summing methods (Balinski and Laraki 2010, pp. 294–301) but they must also be rejected since they are meaningless as well as highly manipulable.

Fifth, join consistency is not an overwhelmingly convincing property because when it is violated by majority judgment the two parts of the electorate—that agree on the ranking of two candidates—give the candidates very different distributions of the grades.

DEFINITION 18. A social-grading function $f$ is *grade join consistent* if $f(\alpha_1, \ldots, \alpha_n) \geqslant \gamma$ and $f(\beta_1, \ldots, \beta_k) \geqslant \gamma$ implies $f(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_k) \geqslant \gamma$.

THEOREM 18 (BALINSKI AND LARAKI 2010, P. 289). *The candidates' majority grades* $(\alpha)$ *and signed majority-grades* $(\alpha\pm)$ *are grade join consistent.*

So in particular, when $A$ wins with a majority grade of at least $\alpha$ and the other candidates have lower majority grades, $A$ wins in the combined electorate; and when $A$ defeats $B$ in two electorates—each of them having the same signed majority grade in both electorates—$A$ defeats $B$ in the combined electorate. There is consistency in combined evaluations and most often also a consistency in orders.

Sixth, approval voting advocates have repeatedly based their objection to majority judgment on the no-show paradox (e.g., Brams 2011), claiming AV is free of such drawbacks. But it is not. Why does the paradox shock? Because more support hurts. Generalize the idea and call it the *no-show syndrome*: more support hurts or more support that could help does not help. Approval voting is obviously subject to the no-show syndrome. For suppose $A$ and $B$ are tied in ticks and that a late voter arrives who believes $A$ has higher merit than $B$ but both are worth a tick or both are not worth a tick. Then he could be decisive but is not. In essence this is the same example as that showing MJ admits the no-show paradox; and, as with MJ, if the late voter sees a real difference between $A$ and $B$, in the AV case giving one of them a tick and not the other, then the no-show syndrome does not occur. The counter argument is that a

late AV voter is hardly ever in the position of being decisive. That is obviously true, but a MJ voter is even more rarely in the position of being able to provoke the no-show syndrome (as confirmed in several practical examples below).

**2.2.7. Faithful Representation.** The idea that words, phrases, or numbers can be found that faithfully represent judges' or voters' evaluations of competitors is essential to the successful use of majority judgment. Regrettably there is no direct, practical experiment to test the proposition that voters have common understandings of the scale of grades. Along with others, an AV advocate has expressed the opinion that it is a "tall order" to believe voters will have common understandings of grades in elections (though he accepts judges can and do in other competitions); and moreover that "[In] many political elections, I'm afraid, voters would . . . [give] their favorites the maximum grade and their most serious competitors the minimum grade . . . "[15] (Brams 2011, p. 427). We disagree for many reasons.

A scale of six or seven ordinal grades expressed in words whose meanings to voters in an election are about the same elicits, we believe, much more accurate expressions of their opinions than either single votes, several votes, or rank orders. Of course, no yardstick is perfect: in measuring the length of a table, a doubt may well remain, is it 31.6 or 31.7 inches? Every measure is approximate, as are the words of any language: "blue" is understood by all; yet to one its evocation may elicit the deep blue of the sea on a winter's day, to another the hazy baby blue of the sky on a hot, humid summer day. Whereas a voter might at times hesitate between *very good* and *good*, a confusion between *excellent* and *good* or *very good* and *acceptable* is doubtful. Hesitations between neighboring grades are more likely due to the approximations of the scale than to a lack of understanding of their meanings.

There is ample empirical evidence showing such common scales of evaluation exist in virtually all competitions other than elections: piano, wine, figure skating, gymnastic, film, student, . . . . Usually they use number scales carefully defined in words (and invariably they err by treating ordinal measures as though they were interval measures). Why accept that judges have the ability to assign grades but not voters?

Experiments carried out in parallel with the 2007 and 2012 French presidential elections (Balinski and Laraki 2011, 2010; Favreau et al. 2012), in parallel with the French Socialist Party's presidential primaries (Balinski and Laraki 2013b, Favreau et al. 2012), in French national polls (Balinski and Laraki 2013b, Terra Nova 2011) and on French websites (*Rue89* 2011; *Slate.fr* 2011, 2012) (some of which are described in this paper), have shown voters are perfectly at ease in evaluating candidates in a scale of six or seven grades. Their voting "behavior" is much the same in all of the experiments. Voters very rarely give only high and low grades; they rarely use all the grades so as to distinguish (and rank order) all the candidates; between a quarter and a third of them give their highest grade to at least two candidates; in multiparty elections they are very parsimonious with high grades, indeed the lower the grade the more it is used; in party elections the grades used are well above those of multiparty elections; "gaps" in the grades of a candidate have never appeared. All of this supports the claim that for voters grades have absolute meanings. Critics may ascribe this behavior to the fact that these were experiments, not real elections, so voters expressed their true opinions and not strategically chosen grades. But in the real uses of MJ for which we have relevant data (e.g., the Louis Lyons Award described in detail below), the behavior was essentially the same.

There are good reasons for this behavior. Voters do not care *only* about the winner: they also care about a lot more, such as the candidates' final grades and the order of finish among all. With at least three candidates many voters will surely use more than only the highest and lowest grades not only because of their honest opinions but also for strategic reasons (e.g., if they value $A$ well above $B$ and $B$ well above $C$, giving the highest grade to $A$ and $B$ may harm $A$ and giving the lowest grade to $B$ and $C$ may harm $B$ should $A$ not be elected). This is true even with only two candidates: a voter may not care for either candidate (or care for both) but prefer one over the other, yet there is no reason to believe she would systematically give the highest grade to one, the lowest to the other for she may wish that both candidates realize they are not (or are) held in high esteem.

Wittgenstein's precept (Wittgenstein 1953), "the meaning of a word is its use in the language," clearly verified in the use of scales of evaluation in all other competitions, can be expected to hold in elections as well. Moreover, there is circumstantial evidence supporting the hypothesis that the meanings of grades are shared. First, the results in all of the experiments have made sense and were consistent with themselves and the known facts. Second, the actual usage of the grades has been consistent with the hypothesis of common meanings in that each of the various grades—from *excellent* down to *to reject*—are used with very close to the same frequencies (see Balinski and Laraki 2010, chap. 15 and Balinski and Laraki 2011). Third, in primaries—where partisan voters evaluated their leaders—the grades assigned were markedly above those used in elections with candidates spanning the electoral spectrum (as may be seen below). Quoting Heisenberg (faced with a similar difficulty in arguing for the uncertainty principle), "We believe we have gained *anschaulich* [intuitively intelligible, visualizable] understanding of a [theory] if . . . we can grasp the experimental consequences qualitatively and see that the theory does not lead to contradictions (Heisenberg 1927, p. 127)."

The importance of a scale of several grades to faithfully represent voters' opinions is supported by another experiment (Favreau et al. 2012). Conducted in parallel with the first round of the French presidential election April 22, 2012, voters were asked to vote with AV in Fresnes's bureau #12—"tick the candidates you *approve*"—and with DisAV

**Table 19.** Results,[a] first-past-the-post, AV and DisAV (order of finish in parentheses), French 2012 presidential election, Fresnes, April 22, 2012.

| | First-past-the-post | | | | | |
|---|---|---|---|---|---|---|
| | Bur. #12 (%) | Bur. #14 (%) | AV Bur. #12 (%) | | DisAV Bur. #14 (%) | |
| 1. Hollande | 39.8 | 37.9 | (1) | 58.3 | (1) | 24.3 |
| 2. Sarkozy | 21.3 | 19.0 | (4) | 25.9 | (4) | 50.4 |
| 3. Mélenchon | 13.7 | 15.6 | (2) | 42.6 | (3) | 36.5 |
| 4. Le Pen | 10.6 | 12.2 | (6) | 12.9 | (10) | 75.9 |
| 5. Bayrou | 7.8 | 10.0 | (3) | 33.7 | (2) | 28.7 |

[a]460 voters participated in bureau #12 (60% of those who voted officially); 422 participated in bureau #14 (64% of those who voted officially).

in its bureau #14—"tick those candidates you *disapprove*" (see Table 19).

The first-past-the-post results in the two bureaus give substantially similar results and the same orders of finish (among all 10 candidates, except that in one bureau two minor candidates are tied). However, AV and DisAV give different orders. *Approval* and *disapproval* do not carry opposite meanings: their sum for a candidate should be 100% (about) but is consistently well below. Why? Voters, it seems, have intermediate evaluations that are neither *approve* nor *disapprove* but are unable to express them. Two grades are not faithful, they do not resolve the representation problem: at least three grades are necessary.

Majority judgment as hereto described forces a voter to express an opinion since no expression is counted as *to reject*. It is possible to permit *no opinion* but for most applications—including political elections—we believe voters should be forced to express opinions. This at once encourages voters to make the effort to evaluate all candidates and incites candidates to address all voters. There are, however, situations where *no opinion* is a perfectly acceptable opinion (or when there are juries of different sizes judging one competition, e.g., wines). In such cases the option should be made explicitly, a "grade" of *no opinion* appearing on the ballot. The issue then becomes how such grades are to be counted. One procedure that could be used for a small jury or a large electorate is to replace each of a candidate's *no opinion* by the majority grade of the opinions that are expressed (see Balinski and Laraki 2010, pp. 230–233). This means that the candidate's majority gauge in numbers (not percentages) is exactly that of the expressed opinions. Another procedure for a large electorate is to use the majority gauges of the expressed opinions in percentages (Balinski and Laraki 2010, pp. 248–249). This amounts to interpreting the expressed opinions as a sample of the electorate's opinion. What is best to do in practice depends on the application. Note then when *no opinion* is counted as *to reject*, a completely blank ballot can affect an election—which may be a good thing since it encourages participation—whereas it cannot for the other two procedures.

**Table 20.** Approval voting experiment, French 2002 presidential election.

| | Percent of ballots with ticks (%) | Percent of all ticks (%) | Official vote first-round (%) |
|---|---|---|---|
| Lionel Jospin | 40.5 | 12.9 | 19.5 |
| Jacques Chirac | 36.5 | 11.6 | 18.9 |
| François Bayrou | 33.5 | 10.7 | 9.9 |
| Jean-Pierre Chevènement | 30.3 | 9.6 | 8.1 |
| Jean-Marie Le Pen | 14.6 | 4.6 | 10.0 |

Article 11 of the *Déclaration des droits* (1789) states: "The free communication of thoughts and opinions is one of the most precious rights of man." A faithful representation of opinions requires *giving voters the opportunity to express their opinions as accurately as possible*. This is limited only by the necessity of a language of grades that is common to all voters. Research in cognition suggests seven grades plus or minus two is the optimal number for most situations where ordinary mortals are involved (Miller 1956). In contrast, practical experience where a small number of expert judges evaluate skating, diving, gymnastics, piano performances, or wines, for instance, suggests that as many as 25 or even 40 grades can be distinguished by them. The evidence to date suggests six is the optimal number in political elections (Balinski and Laraki 2010, 2011; Terra Nova 2011).

## 3. Majority Judgment vs. Other Methods in Use

Approval voting (AV) was predicted in 1980 to become "the election reform of the twentieth century" (a prediction recently changed to the twenty-first century (Brams 2010, p. vii)). The three main claims for AV were (1) "while AV often will do no more than confirm plurality winners, in doing so it will confer legitimacy on their victories to the extent that it shows their support to be widespread in the electorate"; (2) "it would help elect the strongest candidate, . . . ," having "a strong propensity to elect . . . Condorcet[-winners]"; and (3) "AV . . . provides the voter with more flexible options and thereby encourages a truer expression of preferences than does plurality voting" (Brams and Fishburn 1983, pp. 8, 10, 4). None of these promises, in our opinion, has stood the test of time.

AV winners are often not shown widespread legitimacy; in particular, they often do not obtain ticks from a majority of the electorate. One example is an experiment conducted in parallel with the French presidential election of 2002 (that we organized with other colleagues at a time when the arguments in favor of AV seemed persuasive; see Balinski and Laraki 2010, §18.4). There were 16 candidates and over 2,500 persons participated (78% of those who voted officially in six voting precincts of two towns, having no pretense of being representative of all of France). In the official first-round plurality vote, five candidates each obtained

**Table 21.** Ballot for the Louis Lyons Award, 2009 (one line for each nominee).

| Having taken into account all relevant considerations, I believe, in conscience, that this nominee, as Louis Lyons Award designate, is: | | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Absolutely Outstanding* | *Outstanding* | *Excellent* | *Very Strong* | *Strong* | *Commendable* | *Neutral* |
| Nominee | | | | | | | |

*Notes.* Check one grade in the line of each nominee. No check is interpreted to mean *neutral*.

at least 8% of the votes (see Table 20); 11 had smaller percentages. The three candidates having at least 10% *all* decreased in their shares of total ticks, whereas the 13 others *all* increased in their shares of total ticks. No candidate came near to obtaining ticks from a majority of the electorate, and it is doubtful that Jospin's margin of 40.5% to Chirac's 36.5% conferred him any more legitimacy than the margin of 19.5% to 18.9%; indeed, it may be argued that his much narrower margins over Bayrou and Chevènement conferred less.

A second example is an election of the president of the Social Choice and Welfare Society (the professional society of the specialists of voting theory). Members were instructed "You can vote for any number of candidates by ticking the appropriate boxes" in an election with three candidates. 71 members voted: the candidates' scores were 32 ticks for $A$, 30 for $C$, and 14 for $B$, no AV majority. Members were also asked to indicate their preference orders to permit the results to be analyzed. Fifty-two complied, showing the profile of preferences to be

$$13\colon A \succ B \succ C \quad 11\colon A \succ C \succ B \quad 9\colon B \succ C \succ A$$

$$11\colon C \succ A \succ B \quad 8\colon C \succ B \succ A.$$

There were 22 ticks for $A$, 20 for $C$, and nine for $B$—no AV majority, no resounding legitimacy—also electing $A$, and denying the Condorcet winner $C$.[16]

A third example concerns French Socialist Party presidential primaries (described in more detail below). Voters favorable to the left used AV in an election with five candidates. Their AV ticks were 87%, 85%, 64%, 53%, 53%, and 26% (the last the only nonsocialist among them). Whereas the previous examples suggested a lack of legitimacy, here all socialist candidates were supported by clear majorities,

**Table 22.** The ballots, Louis Lyons Award, 2009.

| Judge | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | Judge | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $J_1$ | 4 | 5 | 5 | 3 | 3 | $J_2$ | 5 | 5 | 6 | 3 | 4 |
| $J_3$ | 6 | 5 | 5 | 5 | 5 | $J_4$ | 6 | 6 | 6 | 6 | 5 |
| $J_5$ | 6 | 5 | 4 | 4 | 5 | $J_6$ | 2 | 5 | 6 | 1 | 3 |
| $J_7$ | 6 | 5 | 6 | 5 | 2 | $J_8$ | 6 | 5 | 5 | 6 | 5 |
| $J_9$ | 5 | 0 | 0 | 0 | 0 | $J_{10}$ | 5 | 1 | 5 | 6 | 1 |
| $J_{11}$ | 5 | 3 | 4 | 3 | 6 | $J_{12}$ | 6 | 6 | 6 | 5 | 5 |
| $J_{13}$ | 3 | 0 | 0 | 4 | 0 | $J_{14}$ | 6 | 4 | 3 | 0 | 0 |
| $J_{15}$ | 6 | 5 | 0 | 0 | 0 | $J_{16}$ | 3 | 5 | 6 | 2 | 2 |
| $J_{17}$ | 6 | 2 | 2 | 2 | 2 | $J_{18}$ | 6 | 5 | 5 | 5 | 5 |
| $J_{19}$ | 6 | 5 | 4 | 4 | 2 | | | | | | |

the top two with overwhelming majorities that hardly distinguished between them. Does 87% versus 85% confer an unassailable legitimacy? We believe not. Together these examples show claims (1) and (2) for AV are doubtful.

AV has always been presented as an extension of plurality voting and in terms of the traditional paradigm: voters are asked to "tick" candidates, the one with the most ticks wins—as practiced by the Social Choice and Welfare Society or INFORMS whose bylaws state, "each voter may vote for any number of candidates for an office and the individual elected shall be the one receiving the largest number of votes" (INFORMS 2014)—and the analysis is given in terms of comparing candidates. There is no hint that ticks have an absolute meaning; indeed, there is a deliberate intent to give no meaning to a tick at all other than it will be counted, emphasizing thereby its strategic, comparative nature. Had the proponents of AV thought *approve* has absolute meaning they would certainly have claimed that Arrow's paradox is avoided.

When a tick is given absolute meaning we have renamed the method *approval judgment* (Balinski and Laraki 2010) because AV becomes MJ with only two grades, and so inherits all of the desirable properties of MJ. *There is a very fundamental difference in conception—in the model, the analysis and the conclusions— when inputs have absolute meanings* as all of the preceding discussion shows. In the preface of a 2010 book, AV advocates seem to have accepted this view: "the idea of judging each and every candidate as acceptable or not is fundamentally different from either" plurality voting or "allowing voters to rank candidates" (Brams 2010). But why limit the judgment to *accept/not accept* or pass/fail? This could not be considered reasonable for judging wines or skaters, so why does it make sense for judging candidates to political offices? Because there are many more voters in an electorate than judges in a jury, and by some magical "law of large numbers" a scale of two levels is sufficient? Experiments described below belie this hope.

MJ provides the voter with more flexible options and thereby encourages, we believe, a much truer expression of evaluations than does AV: in any case, claim (3) for AV pales in comparison.

*The* 2009 *Louis Lyons Award for Conscience and Integrity in Journalism.* A real use of MJ further clarifies the contrast between the two methods. The nominees for the Lyons prize were ranked, as was mentioned before, using MJ. The Nieman Fellows at Harvard University traditionally decide to whom this award is given. The five nominees were all

**Table 23.** Majority judgment results, Louis Lyons Award, 2009.[a]

| Grades: | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|
| Sixes | 11 | 2 | 6 | 3 | 1 |
| Fives | 4 | 11 | 5 | 4 | 6 |
| Fours | 1 | 1 | 3 | 3 | 1 |
| Threes | 2 | 1 | 1 | 3 | 2 |
| Twos | 1 | 1 | 1 | 2 | 4 |
| Ones | — | 1 | — | 1 | 1 |
| Zeros | — | 2 | 3 | 3 | 4 |
| Majority grade | 6 = *Abs. Outs.* | 5 = *Outs.* | 5 = *Outs.* | 4 = *Exc.* | 3 = *V. Str.* |
| Majority gauge | (−−, 6, 8) | (2, 5−, 6) | (6, 5−, 8) | (7, 4−, 9) | (8, 3−, 9) |
| Majority ranking | First | Second | Third | Fourth | Fifth |

[a]In fact the winner of the award was a group of Afghan journalists.

**Table 24.** Approval voting results, Louis Lyons Award, 2009.

| $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|
| 16 | 14 | 14 | 10 | 8 |

*Note.* *Approve* meaning at least *excellent* (the three highest of seven grades).

**Table 25.** Majority judgment ballot, French Socialist primaries, October 9, 2011 (one line for each candidate).

| Ballot (majority judgment)<br>*Having taken into account all relevant considerations, I judge, in conscience, that this candidate, to be the candidate of the Socialist Party in the presidential election of 2012, would be* | | | | | |
|---|---|---|---|---|---|
| *Excellent* | *Very Good* | *Good* | *Acceptable* | *Poor* | *To reject* |
| Candidate | | | | | |

*Notes.* Check one single grade in the line of each candidate. No check in the line of a candidate means *to reject* the candidate.

very highly regarded, so the fellows decided to use the scale of seven grades given in Table 21.

Judges' grades are in Table 22. The $C_i$'s represent nominees, *absolutely outstanding* is encoded by 6, *outstanding* by 5, down to *neutral* by 0. The numbers of each grade given each nominee are given in Table 23 together with the nominees' majority grades, majority gauges, and the majority ranking.

The clear winner is $C_1$: her grades stochastically dominate every other nominee.[17] She is also the clear MJ-winner, the one candidate with a majority grade of *absolutely outstanding*, and though $C_2$ and $C_3$ are close for second place they are clearly distinguished. The behavior of the 19 judges in this real use of MJ is very similar to that of voters in electoral experiments: only one judge gave different grades to all nominees (so 18 gave the same grade to at least two); three judges did not use the highest grade; four gave their highest grade to at least two nominees; and contrary to the predictions of some critics, judges are far from limiting themselves to the highest or lowest grades.

If AV were used by this jury in this situation it seems reasonable to assume *approve* would mean *excellent* or above. This would give the AV scores in Table 24. $C_1$ wins with a margin of two—though she is in fact far away in the lead—and there is a tie for runner-up. It is only in such situations—a tie or a small difference—that the no-show syndrome can arise, and it is much more likely to occur with AV than with MJ. This is immediately evident since MJ elicits more information so avoids ties more than AV. Suppose here that any one of the six judges $J_2$, $J_5$, $J_6$, $J_7$, $J_{16}$, or $J_{19}$ is the late judge. With AV the remaining judges place $C_1$ first, $C_2$ and $C_3$ tied for second in every case. So each of the six judges can be decisive between $C_2$ and $C_3$ but is not (since their evaluations are indistinguishable with AV): each represents an occurrence of the no-show syndrome. None of the six provokes the syndrome with MJ. Experimental evidence supports this: with 101 voters the chance of AV producing a tie varied from 3.3% to 6.9%, the chance of the majority gauges to produce a tie (not the majority values) was about 0.1% (Balinski and Laraki 2010, p. 343).

Recent experiments associated with the French presidential election of 2012 confirm these points. On October 9, 2011 the Socialist Party held the first round of a primary election to designate its candidate for the French presidency. The runoff between the two leading candidates was held one week later, October 16, François Hollande winning with 56.6% of the votes to Martine Aubry's 43.4%. Under our direction École Polytechnique students conducted experiments in parallel with the official vote designed to compare methods (Favreau et al. 2012). The experiments were held in voting bureaus of Alfortville and Fresnes, two small towns close to Paris.

**Table 26.** Approval voting ballot, French Socialist primaries, October 9, 2011 (one line for each candidate).

| Ballot (approval voting)<br>*For each of the following, candidates to represent the Socialist Party in the presidential election of 2012, I declare that I*: | |
|---|---|
| *Approve of his or her election* | *Disapprove of his or her election* |
| Candidate | |

*Note.* No check in the line of a candidate means to disapprove him or her.

**Table 27.** MJ grades, French Socialist primaries, Alfortville, 2011.

| | *Excellent* (%) | *Very Good* (%) | *Good* (%) | *Acceptable* (%) | *Poor* (%) | *To reject* (%) |
|---|---|---|---|---|---|---|
| Hollande | 40.1 | 34.5 | 12.3 | 7.0 | 2.5 | 3.5 |
| Aubry | 33.1 | 36.3 | 12.7 | 7.4 | 3.5 | 7.0 |
| Montebourg | 12.0 | 27.8 | 23.9 | 13.7 | 11.6 | 10.9 |
| Valls | 10.9 | 17.6 | 26.8 | 16.2 | 12.0 | 16.5 |
| Royal | 12.3 | 14.8 | 25.7 | 18.0 | 12.0 | 17.3 |
| Baylet | 1.4 | 4.6 | 14.4 | 21.1 | 29.6 | 28.9 |

**Table 28.** Results, French Socialist primaries, Alfortville, 2011.

| Majority judgment | Majority gauge | Approval voting (%) | Approval judgment (%) | Reported votes (%) | Actual votes (%) |
|---|---|---|---|---|---|
| 1 Hollande | (40.1%, *Good*+, 25.4%) | 87.3 | 87.0 | 37.7 | 39.7 |
| 2 Aubry | (33.1%, *Good*+, 30.6%) | 85.2 | 82.0 | 29.2 | 28.9 |
| 3 Montebourg | (39.8%, *Accept.*+, 36.3%) | 64.1 | 63.7 | 12.5 | 12.3 |
| 4 Valls | (28.5%, *Accept.*−, 44.7%) | 53.2 | 55.3 | 10.0 | 8.6 |
| 5 Royal | (27.1%, *Accept.*−, 47.2%) | 53.5 | 52.8 | 10.3 | 9.7 |
| 6 Baylet | (41.7%, *Poor*+, 28.9%) | 25.7 | 20.4 | 0.4 | 0.7 |

**Table 29.** Distribution, number of *approve*, French Socialist primaries, Alfortville, 2011.

| No. of *Approve* | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| No. of ballots | 6 | 23 | 69 | 85 | 55 | 46 | 0 |
| Percentage of ballots | 2.1 | 8.1 | 24.3 | 29.9 | 19.4 | 16.2 | 0 |

*Primaries, Alfortville*: *MJ vs. AV*. Two hundred and eighty-four persons (61.3% of those who voted officially) participated. They were asked to vote by two different methods—MJ and AV—on a ballot printed on one page: at the top the MJ ballot (Table 25), at the bottom the AV ballot (Table 26). Instead of the usual neutral instructions, voters were asked to either approve or disapprove candidates.

Finally, at the bottom of the ballot participants were asked to answer two further questions:

1. Which candidate did you vote for in the official vote?
2. With which system do you believe you were better able to express your opinions?

The MJ grades are given in Table 27. The order of the candidates leaves no doubts: from the top down every candidate stochastically dominates the successor with two exceptions: Royal has more *excellent*s than Valls and Montebourg. The MJ and AV results are given in Table 28. "Approval voting" gives the percentages of ballots that *approve* each of the candidates; "approval judgment" gives the percentages of MJ-ballots that are *good* or above; "reported votes" answers question 1or how the 284 participants said they voted in the official election with first-past-the-post; and "actual votes" means how all 463 voters actually voted in the official election.

MJ gives more clear-cut results than AV simply because it is based on more detailed information.[18] With AV the top two candidates are both overwhelmingly approved by close scores: the top candidate does not emerge as a clear winner and a majority approves of *every* socialist candidate (Baylet is a member of another party). But in more competitive elections, majorities often approve of *no* candidate.

The AV ballots of the participants were always consistent with the MJ ballots in that when a candidate with a certain grade $\alpha$ was given *approve* then any candidate with a grade of $\alpha$ or above was given *approve* as well, although the voters' threshold-grades—where they begin to *approve*—differed

**Table 30.** Distribution, thresholds for *approve*, French Socialist primaries, Alfortville, 2011.

| *Excellent* | *Very Good* | *Good* | *Acceptable* | *Poor* | *To reject* |
|---|---|---|---|---|---|
| 24 | 72 | 94 | 54 | 28 | 12 |
| 8.5% | 25.4% | 33.1% | 19.0% | 9.9% | 4.2% |

**Table 31.** Rates of successful manipulation, French Socialist primaries, Alfortville, 2011.

| | 101 ballots (%) | 151 ballots (%) |
|---|---|---|
| Majority judgment | 0 | 0 |
| First-past-the-post | 0 | 0 |
| Approval voting | 33.2 | 75.8 |

**Table 32.** Results, MJ and first-past-the-post, French Socialist primaries, Fresnes, 2011.

| Majority judgment | Majority-gauge | First-past-the-post (%) |
|---|---|---|
| 1. Hollande | (18.2%, *Excellent*−, 49.7%) | 35.7 |
| 2. Aubry | (48.5%, *Very Good*+, 20.2%) | 34.5 |
| 3. Montebourg | (33.7%, *Very Good*−, 39.1%) | 18.5 |
| 4. Royal | (37.5%, *Good*−, 38.9%) | 6.0 |
| 5. Valls | (36.4%, *Good*−, 40.4%) | 5.3 |
| 6. Baylet | (27.2%, *Acceptable*−, 48.2%) | 0.0 |

**Table 33.** Results, Condorcet and Borda methods, French Socialist primaries, Fresnes, 2011.

| Condorcet ranking | Aubry (%) | Hollande (%) | Montebourg (%) | Royal (%) | Valls (%) | Baylet (%) | Borda ranking (%) |
|---|---|---|---|---|---|---|---|
| 1. Aubry | — | 50.2 | 68.5 | 85.0 | 85.9 | 95.5 | 77.0 |
| 2. Hollande | 49.8 | — | 65.3 | 85.4 | 87.1 | 94.8 | 76.5 |
| 3. Montebourg | 31.5 | 34.7 | — | 68.3 | 69.0 | 91.8 | 59.1 |
| 4. Royal | 15.0 | 14.6 | 31.7 | — | 54.7 | 78.2 | 38.1 |
| 5. Valls | 14.1 | 12.9 | 31.0 | 45.3 | — | 78.9 | 36.4 |
| 6. Baylet | 4.5 | 5.2 | 8.2 | 21.8 | 21.1 | — | 12.2 |

*Note.* "Borda ranking" gives the average of a candidate's percentages of the vote against all other candidates.

widely. On average, voters gave 3.7 candidates an *approve*. The distribution of the numbers of *approve* on ballots is given in Table 29.

Statistically, the voters' behavior is roughly consistent with an *approve* meaning *good* or better as shown by the similarity between "approval voting" and "approval judgment" in Table 28. However, the ballots show the thresholds for *approve* varied as in Table 30.

So why choose the most restrictive possible set of grades when the aim is to select the best possible candidate? In answer to question #2, 179 (or 63%) believed they were better able to express their opinions with MJ, 89 (or 31%) with AV, and 16 (or 6%) did not answer.

The database of 284 ballots permits empirical evaluations of the extent to which the various methods combat manipulability. Ten thousand random samples of 101 (or 151) ballots are taken. Among them 30% of those who prefer the runner-up to the winner change their votes, giving the highest possible grade or place in the ranking (depending on the method) to the runner-up and the lowest possible grade or place in the ranking to the winner. The *rate of successful manipulation* is the percentage of times this gives the runner-up the victory (see Table 31).

The 0% rate of manipulation with MJ and first-past-the-post is due in part to the comfortable lead of Hollande with both those methods, whereas with AV the top two finishers are very close.

*Primaries, Fresnes*: *MJ vs. Condorcet and Borda*. Four hundred and fifty-seven persons (or 76.9% of those who voted officially) participated. They were asked to vote by MJ using seven grades, *outstanding* added, above the others; to rank order the candidates by assigning a 1 to the first, a 2 to the second, . . . , and a 6 to the last; and to designate the candidate to whom they gave their vote.

First-past-the-post gives the impression of a very close race between Hollande and Aubry with Montebourg well behind

**Table 34.** Rates of successful manipulation, French Socialist primaries, Fresnes, 2011.

| | 101 ballots (%) | 151 ballots (%) |
|---|---|---|
| Majority judgment | 10.7 | 44.1 |
| First-past-the-post | 17.8 | 50.1 |
| Borda's method | 99.9 | 100.0 |

**Table 35.** Results, MJ and first-past-the-post, French 2012 presidential election, OpinionWay poll, April 12–16, 2012 (773 ballots);[a] and AV scores, Strasbourg et al. experiment.

| Majority judgment | Majority gauge | First-past-the-post score (%) | AV-score (%) |
|---|---|---|---|
| 1. Hollande | (45.05%, *Good*+, 43.28%) | (1) 28.63 | (1) 49.44 |
| 2. Bayrou | (34.06%, *Good*−, 40.71%) | (5) 9.09 | (3) 39.20 |
| 3. Sarkozy | (49.25%, *Acceptable*+, 39.62%) | (2) 27.27 | (2) 40.47 |
| 4. Mélenchon | (42.47%, *Acceptable*+, 40.43%) | (4) 11.00 | (4) 39.07 |
| 5. Dupont-Aignan | (40.57%, *Poor*+, 33.92%) | (7) 1.49 | (8) 10.69 |
| 6. Joly | (36.77%, *Poor*−, 38.53%) | (6) 2.31 | (6) 26.69 |
| 7. Poutou | (26.19%, *Poor*−, 45.73%) | (8) 1.22 | (7) 13.28 |
| 8. Le Pen | (46.13%, *Poor*−, 47.63%) | (3) 17.91 | (5) 27.43 |
| 9. Arthaud | (24.83%, *Poor*−, 49.93%) | (9) 0.68 | (9) 8.35 |
| 10. Cheminade | (48.03%, *to Reject*, −) | (10) 0.41 | (10) 3.23 |

[a]The MJ ranking with all 993 ballots differs as follows: Bayrou is ahead of Hollande and Le Pen—*to reject*—drops to ninth.

**Table 36.** Results, Condorcet and Borda methods, French 2012 presidential election, OpinionWay poll, April 12–16, 2012 (773 ballots).

| Condorcet ranking | Hollande (%) | Bayrou (%) | Sarkozy (%) | Mélenchon (%) | Le Pen (%) | Borda-ranking (%) |
|---|---|---|---|---|---|---|
| 1. Hollande | — | 51.6 | 53.9 | 68.5 | 64.1 | 59.5 |
| 2. Bayrou | 48.4 | — | 56.5 | 59.4 | 70.5 | 58.7 |
| 3. Sarkozy | 46.1 | 43.5 | — | 50.5 | 65.7 | 51.4 |
| 4. Mélenchon | 31.5 | 40.6 | 49.5 | — | 59.7 | 45.3 |
| 5. Le Pen | 35.9 | 29.5 | 34.3 | 40.3 | — | 35.0 |

the two front-runners, whereas MJ reveals that Hollande is the clear victor and Montebourg is not far behind (Aubry Table 32). The Condorcet and Borda methods (Table 33) agree but narrowly elect Aubry—in a direct vote against Hollande she obtains 50.2%—whereas Hollande is the MJ winner. Practice shows what was illustrated earlier in theory: majority voting can elect a candidate who is not judged to be the best according to the evaluations of the electorate.

The rates of successful manipulation, drawing random samples from 457 ballots, are given in Table 34. As confirmed by other experiments, Borda's method is highly manipulable. With every method this rate will augment as the margin of victory diminishes, but MJ consistently better resists than other methods, as confirmed here (see also Balinski and Laraki 2010, Jennings 2010).

*French presidential poll, April* 12–16, 2012. Terra Nova sponsored another national poll conducted by OpinionWay shortly before the official first-round of the election (held April 22). Nine hundred and ninety-three participants voted with three methods: first-past-the-post, direct face-to-face votes among the five principal candidates—permitting calculation of their Condorcet rankings and Borda rankings—and majority judgment. To analyze data close to "reality" a subset of 773 ballots was extracted whose first-past-the-post vote is within 0.1% of the true April 22 vote.

The results are in Tables 35 and 36. Table 35 also gives the figures from an approval voting experiment conducted in several voting bureaus (of Strasbourg, Saint-Etienne, and Louvigny, with 2,340 participants) in parallel with the actual vote on April 22 that were adjusted to the first-round national vote (by a procedure explained by its authors (Baujard et al. 2013)), which permits comparisons of the results. (1) The MJ rankings, Condorcet rankings, and Borda rankings (among the five important candidates) are identical but differ from first-past-the-post and AV, suggesting once again that the latter two measure badly. (2) Whereas MJ, Condorcet, and Borda put Bayrou comfortably ahead of Sarkozy, AV places him behind. (3) With AV two candidates have almost identical scores and three very close scores—increasing the possibility of successful manipulation—whereas MJ (and to a lesser degree Condorcet and Borda) determine unequivocal orders. (4) AV puts the extreme right candidate Le Pen fifth

whereas MJ places her close to the bottom because of her overwhelming number of *to reject* (47.63%).

## 4. In Conclusion

Terra Nova—"an independent progressive think tank whose goal is to produce and diffuse innovative political solutions in France and Europe"—proposed majority judgment in its recommendations for reform of the presidential election system in France (Terra Nova 2011). Majority judgment has also been used in a variety of other contexts: (1) electing members of the British Academy (the UK's national academy for the humanities and social sciences); (2) determining the priority order of applicants for professorial positions in departments of economics and statistics of the University of Montpellier 2 and Paul Valery University;[19] (3) classifying wines—and attributing them gold, silver and bronze medals—at the annual Les Citadelles du Vin competition; (4) voting on the Web in the first-round of France's Socialist presidential primary (designed by others for two different sites (*Rue89* 2011, *Slate.fr* 2011)) and the first round of the French presidential election (*Slate.fr* 2012).

Practice and theory show, we believe, that majority judgment should be adopted as the mechanism to use by juries that judge (figure skaters, wines, pianists, films, ...) and electorates that elect (officers of INFORMS, France, the United States, ...). Others hold opposite opinions. This suggests the need for further investigation, theoretical and practical, and in particular repeated *real* uses of majority judgment in a variety of applications that admit comparisons with other methods.

### Endnotes

1. IIA has several different formulations that lead to the same conclusion (see Balinski and Laraki 2010, §3.2). This is not Arrow's original definition.
2. See (Balinski and Laraki 2010, §20.7).
3. "Impartial" means candidates and voters are treated equally.
4. Llull clearly states this rule. Copeland's rule is usually interpreted as giving $\frac{1}{2}$ for a tie; see, e.g., Saari and Merlin (1996).
5. Any further ties were resolved by Borda's method.
6. Proposed by Warren Smith; see rangevoting.org.
7. Proposed by Sylvain Spinelli; see votedevaleur.info.
8. OIV, Organisation Internationale de la Vigne et du Vin.
9. FINA, Fédération Internationale de Natation.
10. The poll was conducted before the regrettable Dominique Strauss Kahn New York affair.
11. The main result in Moulin (1980) is formulated in terms of the traditional model, with all the voters' preferences for candidates assumed to be single peaked with respect to a single fixed order of the candidates along the real line.
12. A social-grading function *respects dissent* when all of $A$'s grades strictly belong to the middlemost interval of $B$'s grades implies that $A$'s final grade is below $B$'s final grade.
13. Letting $f^{o/\mathrm{maj}}$ be the upper-middlemost order function there is a similar theorem: $f^{o/\mathrm{maj}}$ is the unique middlemost social-grading function that respects dissent.
14. The less convincing choice of the upper-middlemost order function reverses the alternation. In the limit the two procedures converge to the same solution—the majority-gauge— so in a large electorate the results are almost always the same for all middlemost functions (Balinski and Laraki 2010, pp. 236–239).
15. One cannot help but remark that if that were the case then the method would amount to approval voting.
16. One voter approved two candidates, 49 approved one, two approved none.
17. Candidate $A$ stochastically dominates candidate $B$ if for every grade $\alpha$, $A$'s percentage of $\alpha$'s and above is at least as high as $B$'s and is higher for at least one grade.
18. It has been suggested that MJ admits many ties because candidates have the same majority grade. With this perspective AV has even more ties because either a majority of the electorate accords *approve* or not, so in Table 27 five candidates are tied. In fact, MJ measures with the majority gauge and AV measures with the number of *approve*.
19. By law French universities must rank order applicants for professorial positions; in parallel, applicants rank order their preferences for the positions to which they applied; the Ministry of Education then uses a matching algorithm to assign applicants to positions.

### Acknowledgments

The authors are indebted to Claudio A. Kuhlmann for preparing the manuscript for publication. This work was supported in part by a grant administered by the French National Research Agency as part of the Investissements d'Avenir program (Idex [Grant Agreement No. ANR-11-IDEX-0003-02/Labex ECODEC No. ANR—11-LABEX-0047]).

### References

Arrow KJ (1951) *Social Choice and Individual Values*, 2nd ed. (Yale University Press, New Haven, CT).

Aumann RJ (1959) Acceptable points in general cooperative $n$-person games. *Contributions to the Theory of Games*, Annals of Mathematics Studies 40, Vol. 4 (Princeton University Press, Princeton, NJ).

Balinski M, Laraki R (2007) A theory of measuring, electing, and ranking. *Proc. Natl. Acad. Sci. USA* 104:8720–8725.

Balinski M, Laraki R (2010) *Majority Judgment: Measuring, Ranking, and Electing* (MIT Press, Cambridge, MA).

Balinski M, Laraki R (2011) Election by majority judgment: Experimental evidence. Laurent A, Dolez B, Grofman B, eds. *In Situ and Laboratory Experiments on Electoral Law Reform: French Presidential Elections* (Springer, Berlin), 13–54.

Balinski M, Laraki R (2013a) How best to rank wines: Majority judgment. Girand-Héraud E, Pichery MC, eds. *Wine Economics: Quantitative Studies and Empirical Applications* (Palgrave Macmillan, London), 149–172.

Balinski M, Laraki R (2013b) Jugement majoritaire vs. vote majoritaire (via les présidentielles 2011–2012). *Revue française d'économie* 27:11–44.

Balinski M, Jennings A, Laraki R (2009) Monotonic incompatibility between electing and ranking. *Econom. Lett.* 105:145–147.

Bassett GW Jr, Persky J (1999) Robust voting. *Public Choice* 99:299–310.

Baujard A, Gavrel F, Igersheim H, Laslier J-F, Lebon I (2013) Vote par approbation, vote par note. *Revue économique* 64:345–356.

Black D (1958) *The Theory of Committees and Elections* (Cambridge University Press, Cambridge, UK).

Bogomolny A (2011) Majority judgement: Measuring, ranking and electing. Interactive mathematics miscellany and puzzles. Accessed November 28, 2013, http://www.cut-the-knot.org/books/Reviews/MajorityJudgement.shtml.

Borda J-C de (1784) Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 657–665.

Brams SJ (2010) *Handbook of Approval Voting*, preface (Springer, Berlin).
Brams SJ (2011) Grading candidates. Review of majority judgment: Measuring, ranking, and electing. *Amer. Scientist* 99:426–427.
Brams SJ, Fishburn PC (1983) *Approval Voting* (Birkhäuser, Boston).
Copeland AH (1951) A "reasonable" social welfare function. Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, Ann Arbor.
Dantzig GB (1963) *Linear Programming and Extensions* (Princeton University Press, Princeton, NJ).
Dasgupta P, Maskin E (2004) The fairest vote of all. *Sci. Amer.* 290:92–97.
Dasgupta P, Maskin E (2008) On the robustness of majority rule. *J. Eur. Econom. Assoc.* 6:949–973.
d'Aspremont C, Gevers L (1977) Equity and the informational basis of collective choice. *Rev. Econom. Stud.* 44:199–209.
Dodgson CL (1876) A method of taking votes on more than two issues. Pamphlet, reprinted in: Black D (1958) *The Theory of Committees and Elections* (Cambridge University Press, Cambridge, UK).
Edelman PH (2012) Michel Balinski and Rida Laraki: Majority judgment: Measuring, ranking, and electing. *Public Choice* 151(2):807–810.
Favreau B, Gonzalez-Suitt J, Guyon A, Hennion T, Starkloff X, Thibault S (2012) Vers un système du vote plus juste. Projet scientifique collectif, Ecole Polytechnique (student project).
Felsenthal DS, Machover M (2008) The majority judgement voting procedure: A critical evaluation. *Homo oeconomicus* 25:319–334.
Galton F (1907) One vote, one value. *Nature* 75:414.
Gibbard A (1973) Manipulation of voting schemes: A general result. *Econometrica* 41:587–601.
Hägele G, Pukelsheim F (2001) Llull's writings on electoral systems. *Studia Lulliana* 41:3–38.
Hägele G, Pukelsheim F (2008) The electoral systems of Nicholas of Cusa in the *Catholic Concordance* and beyond. Christianson G, Izbicki TM, Bellitto CM, eds. *The Church, the Councils, and Reform* (Catholic University of America Press, Washington, DC).
Hammond PJ (1976) Equity, Arrow's conditions, and Rawls' difference principle. *Econometrica* 44:793–804.
Heisenberg W (1927) Uber die Grundprincipien der Quantenmechanik. *Forschungen und Fortschritte* 3:127.
INFORMS P (2014) https://www.informs.org/About-INFORMS/Constitution-and-Bylaws.
Jennings A (2010) Monotonicity and manipulability of ordinal and cardinal social choice functions. Ph.D. thesis, Arizona State University, Tempe, AZ.
Krantz DH, Luce RD, Suppes P, Tversky A (1971) *Foundations of Measurement: Vol. 1: Additive and Polynomial Representations* (Academic Press, New York).
Kurrild-Klitgaard P (1999) An empirical example of the Condorcet paradox of voting in a large electorate. *Public Choice* 107(1–2):1231–1244.
Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psych. Rev.* 63:81–97.
Moulin H (1980) On strategy-proofness and single peakedness. *Public Choice* 35:437–455.
Moulin H (1988) *Axioms of Cooperative Decision Making, Monograph of the Econometric Society* (Cambridge University Press, Cambridge, UK).
Muller E, Satterthwaite MA (1977) The equivalence of strong positive association and strategy-proofness. *J. Econom. Theory* 14: 412–418.
Orlov A (1981) The connection between mean quantities and admissible transformations of scale. *Math. Notes* 30:774–778.
*Rue89* (2011) Testez une autre façon de voter à la primaire. Accessed March 25, 2014, http://www.rue89.nouvelobs.com/making-of/2011/10/05/testez-autre-facon-de-voter-a-la-primaire-225043.
Saari DG, Merlin VR (1996) The Copeland method I: Relationships and a dictionary. *Econom. Theory* 8:51–76.
Satterthwaite MA (1975) Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *J. Econom. Theory* 10(2):187–217.
*Slate.fr* (2011) Jugement majoritaire: Arnaud Montebourg au second tour. Accessed March 25, 2014, http://www.slate.fr/story/43791/primaire-ps-jugement-majoritaire.
*Slate.fr* (2012) Jugement majoritaire: votre avis sur les candidats a-t-il changé? Accessed March 25, 2014, http://www.slate.fr/story/50193/FRANCE-PRESIDENTIELLE-jugement-majoritaire-sondages.
Terra Nova (2011) Réformer l'élection présidentielle, moderniser notre démocratie. Accessed March 25, 2014, http://www.tnova.fr/essai/r-former-l-lection-pr-sidentielle-moderniser-notre-d-mocratie.
Wittgenstein L (1953) Philosophical investigations, Aphorism 43.
Young HP (1988) Condorcet's theory of voting. *Amer. Political Sci. Rev.* 82:1231–1244.
Zitzewitz E (2006) Nationalism in winter sports judging and its lessons for organizational decision making. *J. Econom. Management Strategy* 15:67–99.

**Michel Balinski** is Directeur de recherche de classe exceptionnelle of the CNRS (emeritus) at the Ecole Polytechnique, France. Founding Editor of *Mathematical Programming*, his research has increasingly focused on electoral systems and collective decision making. He is the 2013 recipient of the INFORMS John von Neumann Theory Prize.

**Rida Laraki** is Directeur de recherche of the CNRS at LAMSADE, Université Paris-Dauphine, and professor at the Ecole Polytechnique, France. Primarily a game theorist, his interests include optimization and social choice.