ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Judge : *Don't Vote!*

Michel BALINSKI
Rida LARAKI

Novembre 2010

Cahier n° 2010-27

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.economie.polytechnique.edu/
/ mailto:chantal.poujouly@polytechnique.edu

Judge : *Don't Vote!*

Michel Balinski[1]

Rida Laraki[2]

Novembre 2010

Cahier n° 2010-27

Résumé: Cet article explique pourquoi (1) le modèle traditionnel de choix social n'est pas réaliste, (2) il ne peut en aucun cas proposer une méthode acceptable pour classer et élire, et (3) qu'un modèle plus réaliste implique inévitablement une seule méthode pour classer et élire ---le jugement majoritaire--- qui satisfait le mieux qu'il se peut les critères traditionnels de ce qui constitue une bonne méthode.

Abstract: This article explains why (1) the traditional model of the theory of social choice misrepresents reality, (2) it cannot lead to acceptable methods of ranking and electing in any case, and (3) a more realistic model leads inevitably to one method of ranking and electing---majority judgment---that best meets the traditional criteria of what constitutes a good method.

Mots clés : Paradoxe d'Arrow, paradoxe de Condorcet, patinage artistique, choix social, jugement majoritaire, manipulation stratégique, vote

Key Words : Arrow's paradox, Condorcet's paradox, majority judgment, skating, social choice, strategic manipulation, voting

Classification JEL: D71, C72

Classification AMS: 91A, 91C, 90B

[1] Economics Department of the Ecole Polytechnique and CNRS, France.
[2] Economics Department of the Ecole Polytechnique and CNRS, France.

Judge : *Don't Vote!*

Michel Balinski and Rida Laraki

The final test of a theory is its capacity to solve the problems which originated it.

George B. Dantzig

Abstract

This article explains why (1) the traditional model of the theory of social choice misrepresents reality, (2) it cannot lead to acceptable methods of ranking and electing in any case, and (3) a more realistic model leads inevitably to one method of ranking and electing—majority judgment—that best meets the traditional criteria of what constitutes a good method.

1 Why *Don't Vote!* in Theory

George Dantzig's limpid, opening phrase of the Preface of his classic work on linear programming and extensions [9] is worth repeating over and over again, for it is far too often forgotten. By his final test, the theory of voting—better known as the theory of social choice—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. The theory has yielded no really decent methods for practical use.

Beginning with the first known written traces (1299) of how candidates are to be elected and ranked, voting has been viewed in terms of *comparing the relative merits* of candidates. Each voter is assumed to rank-order the candidates and the problem is to amalgamate these so-called preferences into the rank-order of society.

This view leads to two unsurmountable paradoxes that plague practice, and so 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 (the same) three candidates, it is entirely possible for A to win, yet with the same voting opinions for B to defeat A when C withdraws.

These paradoxes are real. They occur in practice. They are not the invention of some febrile imagination. Condorcet's paradox is not often seen because voting systems very rarely ask voters to give their rank-orders. It was, however,

observed in a Danish election [20]. 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 "California defeated 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 considers X and Y to be tied (see [4] section 7.8).

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. 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.

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

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

Arrow's theorem ignores the possibility that voters have strategies. Under the assumption that their "true" opinions *are* rank-orders, it does not consider the possibility that their inputs may differ from their true opinions, chosen in order to maximize the outcome they wish. A rule of voting is *strategy-proof* when every voter's best strategy is his true preference-order; otherwise, the rule is *manipulable*. Strategy-proof rules are the most 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 the skaters and judges

and to 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 [3, 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, of course, not true for most voters.

Each voter's input is now a rank-order that is chosen strategically, so it may or may not be her true preference list. 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.

Gibbard and Satterthwaite's Impossibility Theorem ([15, 23]) *There is no rule of voting that is unanimous, non-dictatorial and strategy-proof for all possible preference-profiles (when there are at least three candidates).*

A third result shows that there is an inescapable conflict between designating a winner and determining an order-of-finish among candidates or competitors in 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 of a preference-profile having the preferences

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

(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. Or, instead, this last property may be replaced by: whenever a winner is removed the new ranking must be the same as the original.

Incompatibility Theorem [3, 4] *There is no rule of voting that is winner-loser unanimous, choice- and rank-compatible (when there are at least three candidates).*

This theorem 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. Result: 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 the remain winner but the final ranking among the others should not change. Theorem [2]: there is no unanimous, impartial[1] rule of voting that is both choice- and rank-monotone. Moreover, when some non-winner 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"). 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 that there can be no good method of voting. The traditional paradigm leaves a desperate state of affairs.

But applied mathematics 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, applied mathematics 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 remained the same over seven centuries. The premisses of the model have not been questioned. Comparing candidates has steadfastly remained the paradigm of the traditional model. And yet, both common sense and practice show that voters and judges do not formulate their opinions as rank-orders. Moreover, rank-orders are grossly insufficient expressions of opinion, because a candidate

[1] Impartial means candidates and voters are treated equally; see below section 3.

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. There is ample evidence for this.

In the last two presidential election held in France, there were respectively sixteen and twelve 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 [4, 5]. 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.

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

- The traditional model is inadequate: the input a voter is supposed to have in her mind does not correspond to reality.
- The theory that emerges is inconsistent and contradictory.

The goal of this paper is to give a brief mathematical account of a new paradigm and model for a theory of social choice that comes much closer to capturing the way in which voters naturally express their opinions and that escapes the traditional impossibilities. For a complete account 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 [4].

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, journalists, ..., not only candidates for offices. How? Invariably by evaluating them in a common language of grades. That it is natural to do so is evident since it is so often done. In most *real* competitions (other than elections) the order-of-finish of competitors is a function of number-grades attributed by judges. Most often the functions used to amalgamate judges' grades are their sums, or equivalently, their averages. But this is not always so. The recent changes in the rules used in figure skating offer a particularly interesting case study.

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 ("one-by-one") system in 1998. It is explained in terms of a real problem.

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 twenty-two 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 twenty-two 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.

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

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

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.

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 2. Judges' inputs (indicating rank-orders of the six 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.

To here, the new rule is identical to the old one (for details see [4]). 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 Coleman's [8]—and Cusanus's—best know 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 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,

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

	Eldredge	Li	Savoie	Honda	Weiss	Tamura	Number of wins	Borda score
Eldredge	–	9	9	9	9	9	5	45
Li	0	–	7	7	8	7	4	29
Savoie	0	2	–	5	6	5	3	18
Honda	0	2	4	–	5	4	1	15
Weiss	0	1	3	4	–	6	1	14
Tamura	0	2	4	5	3	–	1	14

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

A more general rule than Condorcet's was proposed in 1299 by Ramon Llull [16]. *Llull's method*: rank the competitors according to their numbers of wins plus ties. It is a more general rule because a Condorcet-winner is necessarily a Llull-winner. Eldredge is the Condorcet- 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 [17]) and later Borda (in 1770, published in 1784 [6]) had an entirely different idea (it is so well-known as Borda's method that we use this designation). 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 the candidates' numbers of wins in all face-to-face confrontations, Cusanus and Borda rely on the candidates' total numbers of votes in all face-to-face encounters.

The *OBO rule* used in skating is this:

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

In this case Borda's rule happens to agree with 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 [11, 10]. They proposed it with elaborate theoretical arguments, calling it "the fairest vote of all." In fact it had already been tried, and discarded.

The OBO rule produces a linear order, so is not subject to Condorcet's paradox, but it is (unavoidably) subject to Arrow's paradox, in this example viciously. 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

Name	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
Honda	2	1	1	1	2	2	3	3	3
Weiss	1	2	2	2	3	3	1	2	1
Tamura	3	3	3	3	1	1	2	1	2

This yields the majority votes, numbers of wins and Borda-scores:

	Honda	Weiss	Tamura	Number of wins	Borda-score
Honda	–	5	4	1	9
Weiss	4	–	6	1	10
Tamura	5	3	–	1	8

so the result

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

For the first four skaters the judges' inputs are

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

yielding

	Savoie	Honda	Weiss	Tamura	Number of wins	Borda-score
Savoie	–	5	6	5	3	16
Honda	4	–	5	4	1	13
Weiss	3	4	–	6	1	12
Tamura	4	5	3	–	1	13

so the result

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

Before Savoie's performance Weiss led Honda, after 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. This chaotic behavior of repeated flip-flops is completely unacceptable to spectators, competitors, and of course common sense. It is inherent to the OBO and Borda methods.

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" [24]. Once again the skating world entered into fierce fights over how to express and how to amalgamate the opinions of judges. Finally—thankfully—it abandoned the idea that judges' inputs should be rank-orders. In so doing, they joined the the growing number of competitions where the rules have judges 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 or aggregation functions are called by some *point-summing methods* by others *range voting*. The judges' scores in the 2001 Four Continents Figure Skating Championships provides an immediate example. Judges' inputs are now 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:

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

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 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 4, 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 to how best manipulate.

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

Table 4. Judge J_2's manipulations that change the order-of-finish to what she wishes (given in the first row). Note that her new grades define the same order.

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 and this can occur only if each voter's input is her true opinion, not some input chosen strategically in the attempt to achieve her ends.

Meaningfulness

Using a point-summing rule raises deep and important questions: Is it at all *meaningful* to sum or average the scores given a competitor? *What* scores? E.g., if finite in number and they go from a low of 0 to a high of 20, should they be evenly spaced or not? Why and under what conditions is it justified to sum them?

How to construct a scale is a science in itself. "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" ([19], p. 1). Given a faithful representation, the type of scale dictates the meaningfulness of the operations by which measurements may be analyzed. Pain, for example, is measured on an eleven point *ordinal* scale going from 0 to 10: sums and averages are meaningless. 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 and inches are *ratio* scales: they are interval scales where 0 has an absolute sense and multiplication is meaningful.

Since a point-summing method sums candidates' scores they must—to be meaningful—be drawn from an interval scale. 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 is the score, implying the scores do not constitute an interval scale, so that sums and averages are meaningless. But in the context of elections the scores of sum-scoring methods are not even defined, they are given no common meaning. *Range-voting* has an infinite scale $[0, 100]$: but one voter's 71 may mean something entirely different from another's 71. *Approval voting*—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—has a finite scale of two scores: but one voter's 1 may mean something entirely different from another's 1. Both methods suggest comparisons since the scores bear no absolute meanings—no meanings whatsoever other than that they will be summed—so both invite strategic voting and both are open to Arrow's paradox (e.g., if some voter's favorite candidate withdraws he may decide to increase the score of some other candidate(s) causing a change in the order-of-finish among the candidates that remain).

A poll held in the context of the French presidential election of 2007 shows the extent of the difference in meanings in dichotomous responses. The question on the right invokes a comparison, that on the left an evaluation: the results

are completely different. When approval voting is used some voters may well have the first question in mind, others the second: nothing justifies the idea of adding such votes.

	Question: Would each of the following candidates be a good President of France?		Question: Do you personally wish each of the following candidates to win the presidential election?	
	Yes	No	Yes	No
Bayrou	60%	36%	33%	48%
Sarkozy	59%	38%	29%	56%
Royal	49%	48%	36%	49%
Le Pen	12%	84%		

Table 5. Polling results, March 20 and 22, 2007 (Bva), French presidential election of 2007.

3 A More Realistic Model

Postulate a finite number of *competitors* or *candidates* $\mathcal{C} = \{A, \ldots, I, \ldots, Z\}$; a finite number of *judges* or *voters* $\mathcal{J} = \{1, \ldots, j, \ldots, n\}$; and a *common language of grades* $\Lambda = \{\alpha, \beta, \gamma, \ldots\}$. The grades may be any strictly ordered words, phrases, levels or categories. Any two levels may be compared, $\alpha \neq \beta$ implies either $\alpha \prec \beta$ or $\alpha \succ \beta$, and transitivity holds, $\alpha \succ \beta$ and $\beta \succ \gamma$ imply $\alpha \succ \gamma$. A language may be finite or a subset of points of an interval of the real line.

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 (for a detailed analysis of this point see [4] where what it takes for the scale to be interval is explained). 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. By way of contrast, ranking competitors is only relative, it bars any scale of evaluation and ignores any sense of shared benchmarks.

A problem is specified by its *inputs*, a *profile* $\Phi = \Phi(\mathcal{C}, \mathcal{J})$: it is an m by n matrix of the grades $\Phi(I, j) \in \Lambda$ assigned by each of the n judges $j \in \mathcal{J}$ to each of the m competitors $I \in \mathcal{C}$,

$$\Phi = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} & \alpha_n \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \beta_1 & \beta_2 & \cdots & \beta_{n-1} & \beta_n \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{pmatrix}.$$

With this formulation of the inputs—the assignment of grades to competitors—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. Voters are able to input detailed expressions of their preferences that are at once simple and cognitively natural (as experience has proven).

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

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

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

(2) *Anonymity*: $A \succeq_S B$ for the profile Φ implies $A \succeq_S B$ for the profile $\Phi\sigma$ for any permutation σ 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*: $A \succeq_S B$ and $B \succeq_S C$ implies $A \succeq_S C$. That is, Condorcet's paradox cannot occur.

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

These four are the rock-bottom necessities. Together they severely restrict the choice of a method of ranking.

> A method of ranking *respects grades* if the rank-order between them depends only on their sets of grades; in particular, when two competitors A and B 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.

The following theorem shows that the new paradigm—voters evaluate competitors—*must* replace the old paradigm—voters compare competitors.

Theorem 1 *A method of ranking is impartial, transitive and independent of irrelevant alternatives in ranking if and only if it is transitive and respects grades.*

Proof. Assume a method satisfies the properties. IIAR implies that to compare two competitors it suffices to compare them alone.

Suppose A is assigned the grades $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and B is assigned a permutation of them $(\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_{\sigma n})$. Then, it is shown, the properties imply A and B must be tied in society's ranking.

Let

$$\Phi^1 = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{\sigma 1} & \cdots & \alpha_n \\ \alpha_{\sigma 1} & \alpha_2 & \cdots & \alpha_1 & \cdots & \alpha_n \end{pmatrix},$$

where the grades of A are in the first row and those of some other competitor A' are in the second row. Suppose $A \succeq_S A'$. Permute the grades of the two judges 1 and $\sigma 1$,

$$\Phi^{1*} = \begin{pmatrix} \alpha_{\sigma 1} & \alpha_2 & \cdots & \alpha_1 & \cdots & \alpha_n \\ \alpha_1 & \alpha_2 & \cdots & \alpha_{\sigma 1} & \cdots & \alpha_n \end{pmatrix}.$$

By anonymity this changes nothing. So the first row of Φ^{1*} ranks at least as high as the second: but by neutrality $A' \succeq_S A$, so that $A \approx_S A'$. Thus $(\alpha_1, \alpha_2, \ldots, \alpha_n) \approx_S (\alpha_{\sigma 1}, \alpha_2, \ldots, \alpha_n)$ and the second list agrees with B's in the first place.

Now let

$$\Phi^2 = \begin{pmatrix} \alpha_{\sigma 1} & \alpha_2 & \cdots & \alpha_{\sigma 2} & \cdots & \alpha_n \\ \alpha_{\sigma 1} & \alpha_{\sigma 2} & \cdots & \alpha_2 & \cdots & \alpha_n \end{pmatrix},$$

and permute judges 2 and $\sigma 2$ to conclude, as in the first step, together with transitivity, that $(\alpha_1, \alpha_2, \ldots, \alpha_n) \approx_S (\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_n)$, where the second list agrees with B's in the first two places. Continuing, it at most n steps, $(\alpha_1, \alpha_2, \ldots, \alpha_n) \approx_S (\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_{\sigma n})$, showing that two competitors with the same set of grades are tied.

Consider any two lists α and β. Each is equivalent to a unique representation in which the the grades are listed from the highest to the lowest. It suffices to compare them to determine which leads the other, so grades are respected.

The converse is immediate. ■

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*—only the sets of grades themselves may 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.

This 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, though sometimes the top and bottom grades, or top two and bottom two grades, are omitted. Such rules present an immediate difficulty because two competitors with different sets of grades may have the same average, and so are tied.

In any case, such functions 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.

Accordingly, a function

$$f : \Lambda^n \to \Lambda$$

that transforms grades given a competitor into a final grade is an *aggregation function* if it satisfies three properties:

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

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

and

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

An aggregation function serves two separate though related purposes: (1) It assigns a final grade to each competitor and (2) it determines the order-of-finish of all competitors. They are analyzed in both their uses, as, respectively, *social-grading functions* and *social-ranking functions*.

A language of grades Λ is usually parameterized as a bounded interval of the nonnegative rational or real numbers $[0, R]$. Obvious examples of aggregation functions are the arithmetic mean or average, any other means such as the geometric or harmonic mean, and the kth order function f^k that is the kth highest grade (for $k = 1, 2, \ldots, n$). Since small changes in the parametrization or the input grades should naturally imply small changes in the outputs or final grades it is natural to assume that an aggregation function is continuous. This assumption is sometimes necessary to establish the characterizations that follow, but not always. However, the characterizing properties hold for arbitrary finite or infinite common languages of grades.

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

4 How Best to Evaluate: Majority Judgment

A method of voting must meet six essential demands:

- Avoid Condorcet's paradox,
- Avoid Arrow's paradox,
- Elicit honest voting,
- Be meaningful,
- Resist manipulation, and
- Heed the majority's will.

One method best meets these demands, majority judgement.

Description

Suppose there are n judges or voters who assign competitors grades. The kth *order function* f^k is the aggregation function social-grading function whose value is the kth highest grade. When the set of grades r of a competitor are ordered from highest to lowest,

$$r = (r_1 \succeq r_2 \succeq \cdots \succeq r_n) \Rightarrow f^k(r) = r_k,$$

a competitor's *majority-grade* f^{maj} is his middlemost or median grade when n is odd, his lower-middlemost when n is even:

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

Interpret the judges' scores 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 5.

	f^1	f^2	f^3	f^4	f^{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

Table 5. Competitors' scores ordered from highest to lowest (identities of judges forgotten). Majority-grade italicized.

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.

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-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, as indicated here

$$\begin{array}{ccccccccccccc} r_1 \succeq \cdots \succeq & r_{t-2} & \succeq & r_{t-1} & \succeq & r_t & \succeq & r_{t+1} & \succeq & r_{t+2} & \succeq \cdots \succeq r_{2t-1} \\ & 5th & & 3rd & & 1st & & 2nd & & 4th & \end{array}$$

so that it is

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

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,

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

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

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

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

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

There are no ties. There can be no tie unless 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 do anything to raising 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.

The Case of Large Electorates

Majority judgment has been tested in several elections [4, 5]. In an experiment conducted on the Web within the last six weeks of the United States presidential election of 2008, members of INFORMS were invited to vote using the ballot given in table 6. Rather than numbers (which have no meaning unless carefully defined), voters were posed a solemn question asking them to evaluate candidates in grades from a scale of six commonly understood words.

Election of the President of the United States of America 2008

To be the President of the United States of America,

having taken into account all relevant considerations,
I judge, in conscience, that this candidate would be:

	Excell-ent	*Very Good*	*Good*	*Accept-able*	*Poor*	*To Reject*	*No Opinion*
Michael R. Bloomberg, Ind.							
Hillary R. Clinton, Dem.							
John R. Edwards, Dem.							
Michael D. Huckabee, Rep.							
John S. McCain, Rep.							
Barack H. Obama, Dem.							
Colin L. Powell, Ind.							
W. Mitt Romney, Rep.							

You must check one single grade or "No opinion" in the line of each candidate.
"No opinion" is counted as *To Reject*.

Table 6. Ballot, U.S. presidential election, INFORMS experiment, conducted September-early October, 2008.

When there are many voters and few grades it is essentially certain that a candidate's middlemost grade will be repeated many times. Thus, a majority of voters assign a candidate at least her majority-grade, and also a majority of voters assign the candidate at most her majority-grade.

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*to Reject*
Obama	35.9%	32.1%	12.2%	08.4%	07.6%	03.8%
Clinton	16.0%	29.0%	21.4%	16.8%	11.5%	05.3%
Powell	10.7%	22.1%	26.0%	26.7%	09.2%	22.1%
Bloomberg	03.1%	14.5%	24.4%	26.7%	09.2%	22.1%
Edwards	01.5%	13.0%	22.1%	30.5%	18.3%	14.5%
McCain	03.1%	07.6%	23.7%	21.4%	30.5%	13.7%
Romney	00.8%	07.6%	10.7%	27.5%	30.5%	22.9%
Huckabee	03.8%	03.8%	06.1%	19.8%	19.1%	47.3%

Table 7. Results, U.S. presidential election, INFORMS experiment, conducted September-early October, 2008.

The results are given in table 7. For example, Clinton's majority-grade is *Good*: $16.0\% + 29.0\% + 21.4\% = 66.4\%$ assign her at least *Good* and $21.4\% + 16.8\% + 11.5\% + 5.3\% = 55.0\%$ assign her at most *Good*.

The procedure for finding the majority-ranking when there are many voters does not necessitate finding the candidates' majority-values. A simpler procedure determines the majority-ranking. Suppose a candidate's majority-grade is α and that $p\%$ of his grades are higher than α and $q\%$ are lower. Then his *majority-gauge* is $(p, \alpha\pm, q)$, where $p > q$ implies α is endowed with a $+$, and otherwise it is endowed with a $-$. Thus Clinton's majority-gauge is (45.0%,*Good*+,33.6%). The majority-gauges determine the rank-order of the candidates. Naturally, $\alpha+$ ranks higher than $\alpha-$, which suffices to rank-order

all the candidates except Bloomberg and Edwards who both have the majority-grade *Acceptable+*. If two candidates have an $\alpha+$, then the one with the larger p ranks higher; and if two candidates have an $\alpha-$, then the one with the higher q ranks lower. So Bloomberg with $p = 42.0\%$ ranks higher than Edwards with $p = 36.6\%$.

	p	$\alpha\pm$	q
Barack H. Obama	35.9%	*Very Good+*	32.0%
Hillary R. Clinton	45.0%	*Good+*	33.6%
Colin L. Powell	32.8%	*Good−*	41.2%
Michael R. Bloomberg	42.0%	*Acceptable+*	31.3%
John R. Edwards	36.6%	*Acceptable+*	32.8%
John S. McCain	33.4%	*Acceptable−*	44.2%
W. Milt Romney	46.6%	*Poor+*	22.9%
Michael D. Huckabee	33.5%	*Poor−*	47.3%

Table 8. Majority-gauges and majority-ranking, U.S. presidential election, INFORMS experiment, conducted September-early October, 2008.

Notice that voters who believed Clinton merited a higher majority-grade than *Good*—and 45.0% were of that persuasion—could do nothing alone or in concert to raise her majority-gauge. Symmetrically, those who believed she merited a lower majority-grade—33.6% of them—could do nothing alone or in concert to lower her 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: honesty is the best policy.

5 Majority Judgment: Theory

When an aggregation 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. Thus two of the six essential demands are necessarily met.

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" [14]. 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 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. Clearly, no social grading function can prevent all voters from lowering the final grade or all voters from raising the final grade.

> 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 judgement 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 Clinton 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 is $u_j(\mathbf{r}^*, \mathbf{r}, f, \mathcal{C}, \Lambda) = -|r_j^* - f(r_1, \ldots, r_n)|$. Then it is a *dominant strategy* for her to assign the grade she believes is merited: that is, it is at least as good as any other strategy and it is strictly better than others in some cases.

The use of a strategy-proof-in-grading function permits an "honest-grade-seeking" judge—one whose objective is a final-grade as close as possible to the grade he believes should be assigned—to discard all strategic considerations and to concentrate on the task of deciding what he believes is the true grade; moreover, he has no need to pay attention to his preference between two grades when one is lower than the true grade and the other higher.

Theorem 2 *The unique strategy-proof-in-grading social-grading functions are the order functions (for a finite or an infinite number of grades).*

Proof. Let $\Lambda = [0, R]$ and $f(r_1, \ldots, r_n) = r$. Unanimity and monotonicity imply the value of r must fall between the worst and best grades, $\max r_j \geq r \geq \min r_j$.

Suppose the judges assigned the grades, $r_1 \geq \ldots \geq r_n$. Then

$$f(r_1, \ldots, r_n) = r_k \text{ for some } k,$$

as will be shown.

To begin, notice that if $f(r_1, \ldots, r_n) = r$,

$$r_j > r \text{ implies } f(r_1, \ldots, r_{j-1}, r_j^*, r_{j+1}, \ldots, r_n) = r \text{ for any } r_j^* \geq r.$$

This is true for two separate reasons. First, when r_j is increased to a higher grade r_j^* the value of f cannot increase since f is strategy-proof-in-grading. Second, when r_j is decreased to a lower grade $r_j^* \geq r$, the value of f can either remain the same or decrease. But if it decreased, then increasing the grade from that point would again contradict the strategy-proofness of f.

Similarly, and for the same reasons, when $f(r_1, \ldots, r_n) = r$,

$$r_j < r \text{ implies } f(r_1, \ldots, r_{j-1}, r_j^*, r_{j+1}, \ldots, r_n) = r \text{ for any } r_j^* \leq r.$$

Define $\mathbf{r} = (r_1, \ldots, r_n)$ with $r_1 \geq \ldots \geq r_n$. If $f(\mathbf{r}) = r = R$, then $r_1 = \max_j r_j = R$, so $k = 1$. Similarly, $f(\mathbf{r}) = r = 0$ implies that $r_n = \min_j r_j = 0$ and $k = n$.

So, it may be supposed $R > f(\mathbf{r}) = r > 0$. Assume, now, that $r \neq r_j$ for all $j \in \mathcal{J}$: this leads to a contradiction. Given that $r_1 \geq \ldots \geq r_n$, it must be that $r_j > r > r_{j+1}$ for some j. Therefore, the previous deductions imply that for any grades r^+ and r^- satisfying $r^+ > r > r^-$,

$$f(\overbrace{r^+, \ldots, r^+}^{j}, \overbrace{r, \ldots, r}^{n-j}) = r \text{ and } f(\overbrace{r, \ldots, r}^{j}, \overbrace{r^-, \ldots, r^-}^{n-j}) = r.$$

But by monotonicity, the value of f on the left is strictly greater than the value of f on the right, a contradiction, proving $r = r_k$ for some $k \in \mathcal{J}$.

Putting the two parts of the argument together establishes:

$$f(r_1, \ldots, r_n) = r_k \text{ when } r_1 \geq \ldots \geq r_n$$

implies

$$f(s_1 \ldots, s_n) = r_k \text{ when } s_1 \geq \ldots \geq s_{k-1} \geq s_k = r_k \geq s_{k+1} \geq \ldots \geq s_n \text{ ;}$$

that is, so long as $s_k = r_k$ and there are $k - 1$ values of s at least as big as r_k and $n - k$ values of s at most as big as r_k, the value of f does not change: it is the kth biggest of these arguments when its value is r_k.

It must still be shown that k is independent of the input $\mathbf{r}$. Define $g(\mathbf{r}) = k$ if $f(\mathbf{r}) = r_k$ on the open set $R > r_1 > \cdots > r_n > 0$. The continuity of f implies the continuity of g on this set. Since g takes only integer values, it must be a constant on this set. So $f(\mathbf{r}) = r_k$ for the same constant k on $R > r_1 > \cdots > r_n > 0$, hence everywhere by the continuity of f, completing the proof when the language is infinite.

When the language is finite, identify the lowest grade with 0, the highest grade with R.

By unanimity, $f(R, ..., R) = R$, and $f(0, ..., 0) = 0$. Since the value of $f(r_1, \ldots, r_n)$ must be one of its arguments, monotonicity implies that there is some k such that

$$f(\overbrace{R, \ldots, R}^{k-1}, R, \overbrace{0, \ldots, 0}^{n-k}) = R \quad \text{and} \quad f(\overbrace{R, \ldots, R}^{k-1}, 0, \overbrace{0, \ldots, 0}^{n-k}) = 0.$$

This k is unique. If the language contains only two grades, the proof ends here.

Suppose the language contains at least 3 grades. Let $R > r_k > 0$ and

$$f(\overbrace{R, \ldots, R}^{k-1}, r_k, \overbrace{0, \ldots, 0}^{n-k}) = r.$$

It will be shown that $r = r_k$. Since f's values are one of its arguments, either $r = 0$, r_k or R. If $r = R$, voter k who believes that the final grade should be lower can decrease the final-grade to the lowest grade (or 0) by lowering his grade to the lowest grade (or 0), violating strategy-proofness. Similarly, if $r = 0$, voter k can increase the final-grade from the lowest (0) to the highest (or R). Thus, $r_k = r$. Therefore, as has already been seen,

$$f(s_1 \ldots, s_n) = r_k \text{ when } s_1 \geq \ldots \geq s_{k-1} \geq s_k = r_k \geq s_{k+1} \geq \ldots \geq s_n.$$

Thus since k is defined uniquely it does not depend on the input $\mathbf{r}$, completing the proof. ■

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^I) where $r_j^I \in [0, R]$, let the vector of final grades be r^I. Suppose the final grades of some two competitors $A, B \in \mathcal{C}$ 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.

> 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 he 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 aggregation 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 3 *There exists no social-ranking function that is strategy-proof-in-ranking.*

It is an immediate consequence of the next theorem. But the impossibility of perfection does not deny a search for the best possible.

> 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 he 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.

To see that the majority-gauge is partially strategy-proof-in-ranking consider the data of the INFORMS experiment (table 8). Obama with a majority-gauge of $(35.9\%, \textit{Very Good}+, 32.0\%)$ leads Clinton whose majority-gauge is $45.0\%, \textit{Good}+, 33.6\%)$. How could a voter who prefers Clinton to Obama manipulate the outcome? Suppose she could increase Clinton's majority-gauge. Then she gave to Clinton at most a *Good*, so to Obama a lower grade, implying she can do nothing to decrease Obama's majority-gauge. If, on the other hand, she could decrease Obama's majority-gauge, then she gave Obama at least a *Very Good*, so to Clinton a higher grade, implying she can do nothing to increase Clinton's majority-gauge.

Theorem 4 *The unique social-ranking functions that are partially strategy-proof-in-ranking are the order functions.*

Proof. Suppose f is a partially strategy-proof-in ranking. It is first shown that this implies f is strategy-proof-in grading.

Suppose $r_1^A > r_2^A > \ldots > r_n^A$ and that $r_j^A > r^A$ for some voter j. Take B's grades to all be different and in the open interval (r^A, r_j^A):

$$r_j^A > r_j^B > r_1^B > \ldots > r_{j-1}^B > r_{j+1}^B > \ldots > r_n^B > r^A.$$

Then since all of B's grades are higher than r^A, so is B's final grade, $r^B > r^A$.

Now suppose judge j reduces B's grade to any value $\hat{r}_j^B$ in the open interval (r^A, r_n^B). Then

$$\begin{array}{cccccccc} (r_j^B, & r_1^B, & \ldots, & r_{j-1}^B, & r_{j+1}^B, & \ldots, & r_{n-1}^B, & r_n^B) > \\ (r_1^B, & r_2^B, & \ldots, & r_{j+1}^B, & r_{j+2}^B, & \ldots, & r_n^B, & \hat{r}_j^B), \end{array}$$

with a strict inequality holding between every pair of corresponding components. So monotonicity implies the final grade determined by the grades on the top is

strictly higher than that determined by the grades on the bottom. Thus judge j is able to reduce B's final grade. Therefore, by partial strategy-proofness, he cannot increase A's final grade. But then any voter who gave A a higher grade than r_j^A cannot increase the final grade as well as voter j.

A completely symmetric argument shows that if $r_j^A < r^A$ for some voter j then he cannot decrease the final grade, nor can any voter who gave a grade lower than r_j^A.

Together, these last two statements show that f is strategy-proof-in-grading. So, by theorem 2, f must be an order function.

Conversely, let $f = f^k$ be the kth-order function and consider the grades of two candidates A and B

$$r_1^A \geq \ldots \geq r_k^A = r^A \geq \ldots \geq r_n^A$$

and

$$r_{(1)}^B \geq \ldots \geq r_{(k)}^B = r^B \geq \ldots \geq r_{(n)}^B,$$

where $\big\{(1), \ldots, (n)\big\}$ is a permutation of $\{1, \ldots, n\}$. Suppose $r^A < r^B$ and $r_j^A > r_j^B$ where $j = (i)$. Judge j would like to increase A's final grade or decrease B's final grade. If he can increase A's final grade then (since $f = f^k$) $r_j^A \leq r^A < r^B$, so that $r_{(i)}^B < r^B$, implying he cannot decrease B's final grade. Symmetrically, if he can decrease B's final grade then $r_{(i)}^B \geq r^B > r^A$, so that $r_j^A > r^A$, implying he cannot increase A's final grade. Thus f^k is partially strategy-proof-in-ranking, completing the proof. ■

Notice that theorem 4 proves theorem 3: since the only partially strategy-proof-in-ranking functions are the order functions and none of them is strategy-proof-in-ranking there can be no such 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.

The same argument shows that the order functions are the only social-ranking functions that are *group partially strategy-proof-in ranking*: if a group acting together can increase (respectively, decrease) some competitor's final grade then they cannot decrease (increase) the final grade of a competitor to whom they gave a lower (a higher) grade.

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 5). Assume the grades they gave are honest, and that 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.

The effective possible manipulations of the judges are:

- J_1 would like Savoie in 2nd place, Li in 3rd. 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 4th place, Honda in 5th. He cannot lower Honda below anyone. He can place Weiss in 4th 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.

- J_2 would like to raise Honda and Weiss above Savoie. He can do nothing to raise either Honda or Weiss; but he can lower Savoie below them by decreasing his grade to 10.6.
- J_3 agrees with J_1, she would like Savoie in 2nd place, Li in 3rd. Raising Savoie's grade and lowering Li's does not reverse their order. Indeed, even in collusion J_1 and J_3 could not together inverse the order of Li and Savoie.
- J_4 would like to push Honda up to 2nd place, Li and Savoie down to 3rd and 4th. Increasing Honda's grade accomplishes nothing; nor does decreasing Li's. By decreasing Savoie's to 10.7 she can place Honda in 3rd and Savoie 4th; but if she decreased it to 10.6, Savoie would vault down to 5th.
- J_5 would like Tamura in 2nd place not 6th, but he cannot raise Tamura above any other skater nor can he lower any other skater below Tamura. The best he can do is to raise Honda to 3rd place by assigning him 10.9.
- J_6 faces a situation similar to J_5's, though she can lower Honda and Weiss below Tamura, her 2nd place skater, thus putting Tamura in 4th place. In fact, acting together J_5 and J_6 can do no more than placing Tamura above Honda and Weiss.
- J_7 would like to push Weiss up to 2nd place and Savoie down to last place. He can do nothing to change the standings.
- J_8 would like to put Tamura in 3rd place ahead of Savoie and invert the positions of Honda and Weiss. He can accomplish the first wish by increasing Tamura's grade to 10.9, but can do nothing about the second.
- J_9 would like to put Honda in 6th place. The best she can do is to put him in 5th by decreasing his grade to 10.6.

All judges are contented with the 1st place of Eldredge. None can change Li's 2nd place; the only effective manipulations concern skaters in 3rd 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 5th place from 3rd place to 5th 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 2nd place from 6th place to 4th place (J_6); and one can move his candidate for 3rd place from 6th place to 3rd 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.

Meaningful

In the spirit of measurement theory an aggregation must be "meaningful" in both its uses, as social-grading and social-ranking functions: the particular language of grades 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.

> A social-grading function f is *language-consistent* if
>
> $$f\big(\phi(r_1),\ldots,\phi(r_n)\big) = \phi(f(r_1,\ldots,r_n)\big)$$
>
> for any increasing, continuous transformation ϕ 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 and the American grading systems, language-consistency asks that the aggregate French grades rank the students in the same order as the aggregate American grades.

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

Theorem 5 *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.

> A social-ranking function $\succeq_S$ is *order-consistent* if the order between any two candidates for some profile Φ implies the same order for any profile Φ' obtained from Φ by any increasing, continuous transformation ϕ of the grades of each voter.

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.

> 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-, rank- 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 6 ([18, 12]) *The unique choice-monotone, order-consistent social-ranking functions f are the lexi-order functions.*

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

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

Here $\succ_{lex}$ 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 doesn't 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 very clear in this context. A social-ranking function is *preference-consistent* if the order between any two candidates for some profile Φ implies the same order for any profile Φ' obtained from Φ by increasing, continuous transformations ϕ_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* is the deep enduring significance of Arrow's theorem (rather than the supposed impossibility of surmounting Arrow's paradox). But this should not be surprising: how can agreement be found among persons who cannot communicate!

Once again, only the order functions will do. But why the majority-grade and why the majority-value?

Resists Manipulation

To manipulate successfully a voter (or judge) must be able to raise or to lower a candidate's (or competitor') 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.

It may be shown that the 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 $\mu^-\big(f(\mathbf{r})\big)$ be the number of voters who can decrease the final grade, $\mu^+\big(f(\mathbf{r})\big)$ be the number of voters who can increase the final grade, and $\mu\big(f(\mathbf{r})\big) = \mu^-\big(f(\mathbf{r})\big) + \mu^+\big(f(\mathbf{r})\big)$. Take the measure of manipulability μ of a social-grading function f to be the worst that can happen, $\mu(f) = \max_{\mathbf{r}} \mu\big(f(\mathbf{r})\big) \leq 2n$. 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 $\mu(f) = n+1$ are the order functions. For assume $\mu(f) \leq n+1$ and take any $\mathbf{r}$. If more than one voter can both increase and decrease the final grade, then, since all other voters can either increase or decrease the final grade, $\mu(f) \geq n+2$, a contradiction. Therefore, at most one voter can both increase and decrease the final grade, implying f must be an order function.

Taking λ to be the probability the briber wishes to increase the grade and $1-\lambda$ that he wishes to decrease the grade, 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.

> *The probability of cheating* $Ch(f)$ with a social-grading function f is
>
> $$Ch(f) = \max_{\mathbf{r}=(r_1,\ldots,r_n)} \max_{0\leq\lambda\leq 1} \frac{\lambda\mu^+\big(f(\mathbf{r})\big) + (1-\lambda)\mu^-\big(f(\mathbf{r})\big)}{n}.$$

What social-grading functions minimize the probability of cheating?

> A social-grading function is *middlemost* if it is defined by a *middlemost aggregation function* f, where for $r_1 \geq \ldots \geq r_n$,
>
> $$f(r_1,\ldots,r_n) = r_{(n+1)/2} \text{ when } n \text{ is odd, and}$$
>
> $$r_{n/2} \geq f(r_1,\ldots,r_n) \geq r_{(n+2)/2} \text{ when } n \text{ is even.}$$
>
> 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*.

An aggregation function f *depends only on the middlemost interval* means that $f(r_1,\ldots,r_n) = f(s_1,\ldots,s_n)$ when 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 7 *The unique social-grading functions that minimize the probability of cheating are the middlemost that depend only on the middlemost interval.*

Proof. Suppose, first, that n is odd. To see that $f^{(n+1)/2}$ minimizes Ch, observe that for any social-grading function f,

$$\max_{\mathbf{r}} \max_{0\leq\lambda\leq 1} \Big\{\lambda\mu^+\big(f(\mathbf{r})\big) + (1-\lambda)\mu^-\big(f(\mathbf{r})\big)\Big\} \geq$$

$$\geq \max_{\mathbf{r}} \Big\{\frac{1}{2}\mu^+\big(f(\mathbf{r})\big) + \frac{1}{2}\mu^-\big(f(\mathbf{r})\big)\Big\} \geq \frac{n+1}{2},$$

the last inequality following from the earlier discussion. Thus it suffices to show that $Ch(f^{(n+1)/2}) \leq \frac{n+1}{2n}$. But that follows because neither $\mu^+\big(f^{(n+1)/2}(\mathbf{r})\big)$ nor $\mu^-\big(f^{(n+1)/2}(\mathbf{r})\big)$ are greater than $\frac{n+1}{2}$ (equality holding when $r_1 > \ldots > r_n$).

To prove the reverse implication when n is odd, suppose f is an aggregation function that minimizes Ch. Then by the observation just made

$$\max_{\mathbf{r}} \left\{ \mu^+\big(f(\mathbf{r})\big) + \mu^-\big(f(\mathbf{r})\big) \right\} = n+1,$$

so f must be an order function. But $Ch(f^k) = \max\{\frac{k}{n}, \frac{n-k+1}{n}\}$, so $k = \frac{n+1}{2}$.

Now suppose n is even, and that f is any aggregation function for which $Ch(f) \leq \frac{n+2}{2n}$, so that

$$\max_{\mathbf{r}} \left\{ \mu^+\big(f(\mathbf{r})\big), \mu^-\big(f(\mathbf{r})\big) \right\} \leq \frac{n+2}{2}.$$

Take $r_1 > \ldots > r_n$, let $f(r_1, \ldots, r_n) = r$, and suppose that some judge j with $r_j < r_{(n+2)/2}$ can change the final grade by increasing his grade not beyond $r_{(n+2)/2}$: then *a fortiori* he can somewhere both increase and decrease the final grade. But that implies that every judge k with $r_k \geq r_j$ can decrease it as well, so that at least $\frac{n+2}{2} + 1 = \frac{n+4}{2}$ judges can decrease the final grade, a contradiction. Therefore no judge j with $r_j < r_{\frac{n+2}{2}}$ can change the final grade by increasing his grade to $r_{(n+2)/2}$. Similarly, no judge j with $r_j > r_{\frac{n}{2}}$ can change the final grade by decreasing his grade to $r_{n/2}$. Therefore,

$$f(r_1, \ldots, r_n) = f(\overbrace{r_{n/2}, \ldots, r_{n/2}}^{n/2}, \overbrace{r_{(n+2)/2}, \ldots, r_{(n+2)/2}}^{n/2}) = r,$$

implying $r_{n/2} \geq r \geq r_{(n+2)/2}$, so f must be a middlemost aggregation function that depends only on the middlemost interval.

To prove the reverse implication, suppose f is a middlemost aggregation function that depends only on the middlemost interval. Its values are in the middlemost interval. If at most one judge is able to both increase and decrease the final grade by changing his grade, then f must be an order function, so f is either $f^{n/2}$ or $f^{(n+2)/2}$, and $Ch(f) = \frac{n+2}{2n}$. Otherwise, since no other judge can both increase and decrease the final grade and f depends only on the middlemost interval, the two judges who give the middlemost grades can both increase and decrease the final grade for some profile of grades $(r_1, \ldots, r_n)$. Since judge $n/2$ can increase it, so can all judges j with grades $r_j < r_{n/2}$; since judge $(n+2)/2$ can decrease it, so can all judges j with grades $r_j > r_{(n+2)/2}$. Therefore, $\mu^+(f) = (n+2)/2$ and $\mu^-(f) = (n+2)/2$, so $Ch(f) = \frac{n+2}{2n}$. ■

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 the first two are tied; in which case the third decides, unless the first three are tied; 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 7 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 is 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- 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).

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

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 α that commands an absolute majority in answer to the question: "Does this candidate merit at least an α?" Moreover, the unique social-grading functions that assign a candidate the final grade α if a majority of voters assign her α are the middlemost aggregation 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^{maj} among the social grading functions.

> A social grading function *respects consensus* when all of A's grades belong to the middlemost interval of B's grades implies that A's final grade is not below B's final grade.

The rationale is evident: when a jury is more united on the grade of one alternative than on that of another, the stronger consensus must be respected by the award of a final grade no lower than the other's. 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^{maj} is the lower-middlemost order function.

Theorem 8 *The majority-grade f^{maj} is the unique middlemost social grading function that respects consensus.*

Proof. A social grading function f respects consensus if and only if $f \leq f^{maj}$. For suppose f respects consensus, and consider any profile $r_1 \geq \ldots \geq r_n$. If n is odd, $f(r_1, \ldots, r_n) \leq f(r_{(n+1)/2}, \ldots, r_{(n+1)/2}) = r_{(n+1)/2} = f^{maj}$. If n is even, $f(r_1, \ldots, r_n) \leq f(r_{(n+2)/2}, \ldots, r_{(n+2)/2}) = r_{(n+2)/2} = f^{maj}$.

Assume now that $f \leq f^{maj}$, and suppose all the grades of A are in the middlemost interval of B's grades. Then, since $f \leq f^{maj}$, the final grade of B according to f is at most the majority-grade of B. But since f is unanimous and monotonic, the majority-grade of B is at most the final grade of A according to f. This shows that f gives a grade to A at least as high as that given to B, so f respects consensus.

But the only middlemost social grading function f for which $f \leq f^{maj}$ is the majority-grade f^{maj}, and it is clear that the majority-grade respects consensus. The theorem and its proof are valid when the language is finite. ■

A similar concept singles out the majority-ranking $\succ_{maj}$ among the social ranking functions. Consider an ordered set of input grades $r_1 \geq \cdots \geq r_n$. The 1^{st}-*middlemost interval* is the middlemost interval previously defined. The 2^{nd}-*middlemost interval* is the middlemost interval when the defining grades of the 1^{st}-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, when the set of grades is $\{10, 9, 7, 6, 4, 3, 2\}$ the 1^{st}-middlemost interval is $[6, 6]$, the 2^{nd} is $[7, 4]$, the 3^{rd} is $[9, 3]$, and the fourth is $[10, 2]$.

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

> 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 $\{10, 10, 7, 6, 4, 3, 1\}$, then the first interval where they differ is the 3^{rd}: A's is $[9, 3]$ and B's is $[10, 3]$. This is a natural extension of the idea of a middlemost social grading function that depends only on the middlemost interval.

> 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 SRF that rewards consensus. This is a natural extension of the idea of respecting consensus for a social-grading function.

Theorem 9 *The majority-ranking $\succ_{maj}$ is the unique middlemost, choice-monotone social-ranking function that rewards consensus.*

Proof. Suppose the ranking $\succeq_S$ satisfies the properties, and consider two candidates, A and B.

If they differ in the 1^{st}-middlemost interval and n is odd, the statement is true. If n is even, suppose the 1^{st}-middlemost intervals of A and B are $[r_-^A, r_+^A] \neq [r_-^B, r_+^B]$. The properties imply

$$A \succ_S B \text{ when } \begin{cases} r_-^A > r_-^B \text{ and } r_+^A > r_+^B, \\ r_-^A > r_-^B \text{ and } r_+^A = r_+^B, \\ r_-^A = r_-^B \text{ and } r_+^A > r_-^B, \\ r_-^A > r_-^B \text{ and } r_+^A < r_+^B. \end{cases}$$

The three first comparisons are implied by choice-monotonicity (starting from $r_-^A = r_-^B$ and $r_+^A = r_+^B$) and the middlemost property (since it may be assumed that all grades outside the 1^{st}-middlemost intervals are minimum or maximum grades). The last comparison is implied by rewarding consensus and the middlemost property (since it may be assumed that all grades of A are in the 1^{st}-middlemost interval of B). This is exactly the output of the majority-ranking. (For the four remaining possibilities $A \prec_S B$.)

If they first differ in the k^{th}-middlemost intervals of their grades for $k > 1$, the proof is the same. ■

The choice of the lower-middlemost order function for ranking and electing is the consequence of seeking consensus.

6 In Conclusion

The above results are bolstered by others that single out majority judgment as the best method that emerges from the new model. This has important practical implications across a very large spectrum of applications that span sporting events, artistic performances, intellectual achievements, political elections, and a host of other instances where competing entities are to be ranked and winners are to be designated.

Among them are the elections organized by the *Monthly*'s publisher, the Mathematical Association of America. The MAA's bylaws stipulate how its officers are to be elected:

> "Each voting member of the Association may vote for as many candidates for each office as he or she desires. For President-Elect, First Vice President, and Second Vice President, the Nominating Committee shall declare elected the person having received the most votes ..."

This is approval voting. It is meaningless and subject to Arrow's paradox *unless* it is changed in a seemingly trivial but actually deep and significant manner. By giving absolute common meanings to 0's and 1's approval voting becomes the special case of majority judgment where the language of grades consists of exactly two words (so is better identified as *approval judgment*). A reasonable way in which to do this is to vote with a ballot such as that used in the U.S.

presidential INFORMS experiment (table 6), defining 1 to be a grade of *Good* or better and 0 to be a grade worse than *Good*. Had it been so defined, that experiment (table 7) would have yielded precisely the same order of finish, Obama receiving 80.2% approvals, Clinton 66.4%, Powell 58.8%, and so on. Had 1 been defined to be a grade of *Excellent*, 0 a grade worse than *Excellent*, only the order of finish of the first three is the same.

But why on earth choose the most restrictive possible set of grades—two grades, an unnatural "pass/fail" dichotomy—when the aim is to select the best possible candidates? In particular, why do so when a richer common scale of evaluation is available?

The most important property of a system of voting is *to give the 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 [22] is the optimal number for most situations where ordinary mortals are involved (e.g., most people can accurately distinguish at most six different pitches in tone). 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 twenty-five or even forty grades can be distinguished by them (much as a person with perfect pitch can accurately identify up to 60 different tones). The evidence from experiments with majority judgment suggests six is the optimal number in political elections [4, 5].

References

[1] Kenneth J. Arrow. 1951 (2nd ed. 1963). *Social Choice and Individual Values*. New Haven CT: Yale University Press.

[2] Michel Balinski, Andrew Jennings and Rida Laraki. 2009. "Monotonic incompatibility between electing and ranking." *Economics Letters* 105 145-147.

[3] Michel Balinski and Rida Laraki. 2007. "A theory of measuring, electing and ranking." *Proceedings of the National Academy of Sciences, U.S.A.* 104 8720-8725.

[4] Michel Balinski and Rida Laraki. 2010. *Majority Judgment: Measuring, Ranking, and Electing*. Cambridge MA: MIT Press, to appear.

[5] Michel Balinski and Rida Laraki. 2010. "Election by majority judgment: experimental evidence." In ed. by B. Dolez, B. Grofman, and A. Laurent, *In Situ and Laboratory Experiments on Electoral Law Reform: French Presidential Elections*. Berlin: Springer, to appear.

[6] Jean-Charles le Chevalier de Borda. 1784. "Mémoire sur les élections au scrutin." *Histoire de l'Académie royale des sciences* 657-665.

[7] Jean Antoine Caritat le Marquis de Condorcet. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris: l'Imprimerie royale.

[8] A. H. Copeland. 1951. "A 'reasonable' social welfare function." Seminar on Mathematics in the Social Sciences, University of Michigan.

[9] George B. Dantzig 1963. *Linear Programming and Extensions.* Princeton NJ: Princeton University Press.

[10] Partha Dasgupta and Eric Maskin. 2008. "On the robustness of majority rule." *Journal of the European Economics Association* 6 949-973.

[11] Partha Dasgupta and Eric Maskin. 2004. "The fairest vote of all." *Scientific American* March.

[12] Claude D'Aspremont and Louis Gevers. 1977. "Equity and the informational basis of collective choice." *Review of Economic Studies* 44 199-209.

[13] Four Continents Championships 2001: Men - Short program. Found July 19, 2010 at: http://icecalc.org/events/fc2001/results/SEG089.HTM

[14] Francis Galton. 1907. "One vote, one value." *Nature* 75 414.

[15] Allan Gibbard. 1973. "Manipulation of voting schemes: a general result." *Econometrica* vol. 41 587-601.

[16] Gunter Hägele and Friedrich Pukelsheim. 2001. "Llull's writings on electoral systems." *Studia Lulliana* 41 3-38.

[17] Gunter Hägele and Friedrich Pukelsheim. 2008. "The electoral systems of Nicolas of Cusa in the *Catholic Concordance* and beyond." In *The Church, the Councils and Reform: Lessons from the Fifteenth Century*, ed. Hg. G. Christianson, T.M. Izbicki, and C.M. Bellitto, 229-249. Washington, D.C.: Catholic University of America Press.

[18] Peter Hammond. 1976. "Equity, Arrow's conditions, and Rawls' difference principle." *Econometrica* 44 793-804.

[19] D. H. Krantz, R. D. Luce, P. Suppes and A. Tversky. 1971. *Foundations of Measurement, Vol. I.* New York: Academic Press.

[20] P. Kurrild-Klitgaard. 1999. "An empirical example of the Condorcet paradox of voting in a large electorate." *Public Choice* 107 1231-1244.

[21] Pierre-Simon le Marquis de Laplace. 1820. *Théorie analytique des probabilités*, 3rd edition. Paris: Mme V^E Courcier, Imprimeur-Libraire pour les Mathématiques.

[22] George A. Miller. 1956. "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological Review* 63 81-97.

[23] Mark A. Satterthwaite. 1973. "Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions." *Journal of Economic Theory* 10 187-217.

[24] Eric Zitzewitz. 2006. "Nationalism in winter sports judging and its lessons for organizational decision making." *Journal of Economics and Management Strategy* 15 67-100.