Majority Judgment

Measuring, Ranking, and Electing

Michel Balinski and Rida Laraki

Majority Judgment

Majority Judgment

Measuring, Ranking, and Electing

Michel Balinski and Rida Laraki

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in 10/13 Times Roman by Westchester Book Composition.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Balinski, M. L.
Majority judgment : measuring, ranking, and electing / Michel Balinski and Rida Laraki.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-01513-4 (hardcover : alk. paper)
1. Social choice. 2. Voting. 3. Ranking and selection (Statistics) I. Laraki, Rida. II. Title.
HB846.8.B354 2011
302′.13—dc22

2010026311

10 9 8 7 6 5 4 3 2 1

Contents

Preface

There was a risk in theorizing. I had witnessed, close up, the fatal, comic effect upon professors and students of hypotheses which had become unconscious convictions. And thus warned, I had thrown overboard, as a reporter facing facts, many of my college-bred notions . . . It was hard to do; ideas harden like arteries; indeed, one theory of mine is that convictions are identical with hardened arteries. But the facts . . . forced me to drop my academic theories one by one; and my reward was the discovery that it was as pleasant to change one's mind as it was to change one's clothes. The practice led one to other, more fascinating—theories.
—Lincoln Steffens

Kenneth Arrow begins his classic, *Social Choice and Individual Values*, with the sentence, "In a capitalist democracy there are essentially two methods by which social choices can be made: voting, typically used to make 'political' decisions, and the market mechanism, typically used to make 'economic' decisions." In the second paragraph he poses the problem of social choice. "The methods of voting and the market . . . are methods of amalgamating the tastes of many individuals in the making of social choices . . . [Any] individual can be rational in his choices. Can such consistency be attributed to collective modes of choice, where the wills of many people are involved?" (1951, 1–2). His celebrated impossibility theorem answers *no!* for voting.

Arrow's conclusion begs refutation.

But Arrow carefully explained the limitations of his analysis. First, he explicitly assumed that "the behavior of an individual in making choices is describable by means of a preference scale [an ordered list]" (11). Second, he deliberately ignored the strategic aspects of voting, in his words, the "game aspects of social choice" (7). Third, he clearly stated his acceptance of the standard though unreasonable view that "individual values . . . are not capable of being altered by the nature of the process itself" (8). This assumes that a voter's or a judge's expression—his or her vote, his or her evaluation—is not altered by the actions of other voters or judges, nor by the mechanism by which all the

expressions are amalgamated. In so doing, Arrow anticipated future developments, showing where to seek an escape from the logical conclusions of his analysis: a brand new model of social choice not bound by these restrictive and unrealistic assumptions. Practice, it turns out, suggests how the model should be formulated.

A mechanism is "an instrument or process, physical or mental, by which something is done or comes into being."[1] Society routinely uses mechanisms left unmentioned by Arrow to collectively designate who or what is best, second-best, and so on, down to worst: it *measures* in one way or another, then ranks in accordance with the measures and declares the winner to be the one with the highest score. Students are graded, then ranked according to their grades; the attributes of wines (e.g., tannicity, finish, bouquet, body) are assigned numerical values, then ranked according to the values; figure skaters, gymnasts, and divers are given marks for a particular exercise, then ranked according to their marks. In each of these instances there may be several or many judges who assign scores that may be quite different, yet each uses a language that is common to all and is understood by all. "Kenneth is an A+ student" is a meaningful statement (or was before the age of grade inflation). The statement that "Sonja's free skating performance is worth 5.9" when the traditional 0–6 scale was used, or that "her skating skills component is 7.75" with the newly adopted scale, or that "Xu Sang's inward flying $1\frac{1}{2}$ somersault was a 9.0" means something specific to figure skating or diving enthusiasts, though all may not agree with the score that was assigned.

The market mechanism—perhaps better thought of as an invisible hand, since its process is but loosely understood—itself provides the world with a measure: *price* expressed in terms of money. Money is, of course, a complex concept that plays many roles in the dynamic workings of an economy, but its units are mere convention, a numéraire expressed in euros, shekels, dinars, or dollars, invented to simplify or solve problems, just as are letter grades or numerical levels of performance. In past times the units differed: grain in ancient Babylon, rice in Japan, cigarettes in concentration camps, though gold and silver—more durable commodities—have enjoyed a longer-lasting use. Price measures, and as a consequence ranks. An instance is the famous classification of 1855, said to be of the Bordeaux wines, though in fact limited to the Médocs and Sauternes and exactly one Graves, Haut-Brion. The Exposition universelle de Paris of that year prompted an official request for a "complete and satisfactory description of the wines of the department" (Debuigne 1970). This ranking

1. *American Heritage Dictionary*. 2d college ed. Boston: Houghton Mifflin, 1982.

of the "grands crus" has steadfastly maintained its importance to this day; it was determined by the prices of the wines prevalent in those years (Markham 1997). Auctions—the traditional English ascending-price auction, the sealed second-price auction of Vickrey, the Dutch descending-price auction—are other well-defined mechanisms that also use price as a measure to determine winners.

Measurement as a means to choose and to rank has accumulated a rich experience and taught lessons well worth learning. Ranking figure skating competitions is a good example: it is an ongoing, dynamic process that reveals many of the defects known in methods of elections. The traditional system often saw the lead of one skater over another reverse as a result of a *third*'s performance—to the obvious consternation of the spectators. This is, of course, nothing but a violation of Arrow's "independence of irrelevant alternatives": no system of ranking should allow the relative positions of two competitors to be influenced by the (irrelevant) performance of a third. Moreover, cheating—when a judge exaggerates the rank of competitors up or down—has provoked major scandals at the Olympics. This is nothing but strategic manipulation, a problem well known in voting theory and central to the theory of games. What did the International Skating Union do? It invented a new method that appears to satisfy Arrow's condition and pretends to combat the possibility of cheating. It is a method that measures. It is a mechanism that has flaws—as do the market mechanism when markets are "rigged" and auction mechanisms that are often plagued by the "winner's curse" [2]—but these flaws may be overcome.

Up to now, the problem of ranking competitions among athletes, goods, or musicians has remained completely separated from the problem of elections. No organized body of knowledge has accompanied ranking competitors. The methods are, by and large, "home-grown," invented by skating enthusiasts for skaters, by œnologists for classifying wines, by piano maestros for discerning prizes to pianists. Nevertheless, the methods show good sense: increasingly, they use measures.

In contrast, the specialists in how to elect and rank have almost exclusively devoted their attentions to elections, either in small committees such as juries or in large electorates such as nations, where a voter submits either a vote for one or several candidates or an ordered list of his preferences among all candidates. The focus of the work has remained resolutely the same for almost a millennium. It is the model analyzed by Arrow: how to transform the so-called preference lists of individuals or judges into a preference list of society

2. The "winner's curse" occurs when the actual value of the good obtained is well below what was paid for it. This happens typically because the winning bidder made the highest estimate of a good of unknown value, as for example when bidding for the right to exploit an off-shore oil field.

or the jury. Despite Arrow's devastating result showing the traditional theory of social choice has no truly acceptable solution to the problem of how to elect and how to rank, that model—and the manner in which voters may express themselves—has remained unquestioned.

We show that this traditional model is fundamentally flawed, for reasons that go well beyond the classical paradoxes. First, it assumes that the preferences of voters and judges are expressed as rank-orderings. This is clearly not true: a voter confronted with (say) twelve candidates knows from personal experience that he does *not* formulate a list of all candidates from first to last. We present evidence that proves that this conception of what judges or voters have in their minds is simply false. Second, the traditional model does not make a clear distinction between the judges' or voters' preferences and their votes, which are the *messages* they are allowed to send by the method of voting that is used. A voter who announces a rank-ordering—or who votes for one candidate among many—is not and cannot be expressing all of his preferences; he is only sending a very limited and strategically chosen message. Third, there is a profound difference between the problem of electing one candidate and the problem of ranking several or many candidates, though this has not been widely appreciated. H. Peyton Young (1986) is the first to have made a clear distinction between them, showing how a same line of reasoning can yield a ranking of the candidates and a winner who is *not* the first-place candidate of the ranking. This, it turns out, is an irreconcilable difference of the traditional model; a new impossibility theorem shows how and why the two problems are incompatible.

In summary, insofar as it concerns ranking and electing, social choice theory hypothesizes a faulty model of reality to produce an inconsistent theory. So why on earth use it?

The analogy with wine, sports, and music opens the door to another view. Lord Kelvin's celebrated warning may be seen on the façade of the Social Sciences Research Building at the University of Chicago: "If you cannot measure, your knowledge is meager and unsatisfactory" (see Kuhn 1961, 178). The economist Frank Knight is reported to have quipped that for social scientists this means, "If you cannot measure, measure anyhow" (Kuhn 1961, 183). His remark makes more sense than most people at first may suppose.

We have studied what is done in practice. We have learned from the insights of œnologists, sportsmen, pianists, and others that the fundamental question should be posed differently. Instead of trying to translate many individual rankings of competitors into a single collective ranking, or many individual lists of preferences into a single collective list of preferences, a common language to measure should be defined, individuals should measure and assign grades to

each competitor in that language, and the many individual grades should then determine the single collective grade of each competitor. In short, the central problem becomes *how to transform many individual grades of a common language into a single collective grade*, when the many individuals have unknown preferences that are too complex to be formulated. Sharing a common language of grades makes no assumptions about a voter's or a judge's utilities or preferences. Utilities measure the satisfaction of a voter or a judge, grades measure the merits of competitors. The basic atoms of Arrow's model are the comparisons between pairs of alternatives, competitors, or candidates. The basic atoms of our model are the grades of a common language assigned to alternatives, competitors, or candidates. Grades yield rankings, but rankings assuredly do not yield grades.

The celebrated market mechanism works because it uses a common measure that facilitates comparisons of goods, services, assets, debts: monetary units. The evaluations in terms of a common language of grades are no more the utilities of judges or voters making a collective decision than the evaluations of items bought and sold in terms of money are the utilities of agents in a market. *The common language of the market is money. The "money" of collective decision is*—in our model—*the common language of grades.*

The change in point of view—in the premises of the underlying model of social choice—changes everything. The method and theory that emerges is simple. It must be, to be practical. We have called it *majority judgment*. It may be used to elect officials, to classify wines, and to rank contestants for international piano prizes and Olympic competitors in skating, diving, and gymnastics. It has been tested in classifying wines and electing a candidate to political office. A simple theory characterizes the methods that satisfy all the "good" properties that were stated in Arrow's axioms. Beyond overcoming Arrow's impossibility, the model makes it possible to address another important question: What mechanisms are the most robust against cheating and strategic manipulation of the judges or the voters? Or, what mechanisms make the judges' and voters' optimal strategies be to give the grades they believe the competitors and candidates merit? Understanding and discovering the psychology or the possible secret cabals of judges is one perfectly reasonable approach to combating strategic manipulation. Another is to design methods that make it impossible for cheating to take place or, if that ideal is unattainable, that minimize the possibility of manipulation. When grades replace orders, "possibility theorems" and "strategy-proof" or "least manipulative" methods replace impossibility theorems. The majority judgment method uniquely best combats strategic manipulation while satisfying the desirable properties of classical social choice theory. In fact, the approach suggests a new mechanism in the

context of the traditional model (the "Borda-majority judgment method"). The aim of this book is to state, to explain, to prove and to apply these theorems.

Brand new ideas are rare indeed. Anyone, at any time, cannot but build upon the accumulation of all of past knowledge, consciously or unconsciously. Confronted by a real problem—how should wines be judged and ranked?—we stumbled on a simple but fundamental idea, realized with some amazement how useful it could be, and set out to explore and develop it. Only when we began to write did we discover that others, most notably Sir Francis Galton and the Marquis de Laplace, had seen one or another aspect of these ideas before. The mass of knowledge of the theory of social choice and welfare—the concepts, theorems, paradoxes, and mechanisms—provided the road signs for the development of a unified theory of measuring, ranking, and electing. The crux of the matter is a new model in which the traditional paradigm—*to compare*—is replaced by a new one—*to evaluate*. In the words of Thomas Kuhn, "Since new paradigms are born from old ones, they ordinarily incorporate much of the vocabulary and apparatus, both conceptual and manipulative, that the traditional paradigm had previously employed. But they seldom employ these borrowed elements in quite the traditional way. Within the new paradigm, old terms, concepts, and experiments fall into new relationships one with the other. The inevitable result is what we must call, though the term is not quite right, a misunderstanding between the two competing schools" (1970, 149).

George Dantzig, the acknowledged father of linear programming and its many generalizations, begins the preface of his magnus opus (1963) with an often forgotten yet naked truth: "The final test of a theory is its capacity to solve the problems which originated it." Our ambition is to establish a theory that meets this test. And so this book is addressed—in addition to the economists, mathematicians, operations research analysts, and political scientists who are the specialists of the theory of social choice—to anyone who is confronted with problems of electing and ranking, including the just plain interested voter (theorems and proofs may be skipped by those who are only interested in *how* to elect and rank).

Acknowledgments

We are indebted to many, friends and colleagues and institutions.

Friends and Colleagues Francis Bloch, B. Curtis Eaves, David Gale, Edi Karni, Yukio Koriyama, Friedrich Pukelsheim, Jérôme Renault, Ludovic Renou, Martin Shubik, Tristan Tomala, and H. Peyton Young supported and challenged us in discussions and communications, and made important suggestions and

critiques. We are deeply indebted to Jack Nagel, Hannu Nurmi, Maurice Salles, and John Weymark, each of whom analyzed the entire manuscript with great care and made very substantial recommendations. It is a particular pleasure to acknowledge the contributions of two exceptionally gifted graduate students in mathematics, Andrew Jennings and Cheng Wan. Andy Jennings read the entire manuscript, checking the proofs, the statements, the reasoning, and the prose, thereby eliminating errors, ambiguities, and infelicities. Cheng Wan carried out all the extensive computer analyses of the 2007 Orsay experiment. Claude Henry and Vincent Renard gave us their unstinting support and encouragement throughout. Anna Kehres-Diaz accepted and espoused the importance of the ideas for real, practical use, and backed the application for a patent and the development of professional software that permitted several experiments to be carried out on the Web. Last, yet also first, we thank Jacques Blouin. Dissatisfied with the existing methods for evaluating wines, he asked us whether we could find a better way; that is how this work began.

Institutions The C.N.R.S. (Centre National de Recherche Scientifique), where both of us have held permanent positions, is wonderful for those who wish to pursue long-term projects: they are free to go ahead and do it! The École Polytechnique, more particlarly, the Laboratoire d'Économétrie, provided the ideal interdisciplinary environment in which to pursue this project that is at once mathematics, economics, political science, operations research, and statistics, with occasional dashes of linguistics, psychology, sociology, and philosophy. The D.R.I.P. (Direction des relations industrielles et des partanariats) of the École Polytechnique has steadfastly supported every endeavor to realize practical applications of our ideas.

The 2007 Orsay experience could not have been realized without the generous support of Orsay's Mayor, Madame Marie-Hélène Aubry, the staff of the Mayor's office, and our friends and colleagues who sacrificed their Sunday (a beautiful spring day) to urging voters to participate and explaining the idea: Pierre Brochot, Stéphanie Brochot Laraki, David Chavalarias, Sophie Chemarin, Clémence Christin, Maximilien Laye, Jean-Philippe Nicolai, Matias Nuñez, Vianney Perchet, Jérôme Renault, Claudia Saavedra, Gilles Stoltz, Tristan Tomala, Marie-Anne Valfort, and Guillaume Vigeral. Thanks to them, the experiment was successful and its expense limited to the costs of ballots, envelopes, and posters.

1 Majority Judgment

For with what judgement ye judge, ye shall be judged: and with what measure ye mete, it shall be measured to you again.
—Matthew 7:2

1.1 Inputs and Outputs

Throughout the world, voters elect candidates, and judges rank competitors, goods, alternatives, cities, restaurants, universities, employees, and students. How? Schemes, devices, or *mechanisms* are invented to reach decisions. Each defines

· the specific form of the voters' and judges' *inputs*, the *messages* used to exert their wills, and

· the procedure by which the inputs or messages are amalgamated or transformed into a final decision, social choice, or *output*.

In piano competitions, a judge's input message is a grade assigned to each competitor—often in the range from 0 (low) to 25 (high)—and the output is the rank-ordering determined by the competitors' average grades over all judges (sometimes by the means or averages of the grades after the one or two lowest and highest grades of each competitor have been eliminated). In the international standard for wine competitions, obtaining a judge's input message is more complex: each of fourteen attributes of a wine is assigned one of seven mentions (*Excellent, Very Good, Good, Passable, Inadequate, Mediocre, Bad*). The *Excellent*s are given a number score of 6 or 8, the *Bad*s a 0, and the others integer scores in between. The sum of the scores over all the attributes determines the judge's input message. The output is one of four medals (grand gold, gold, silver, bronze) or none, assigned to each wine on the basis of its average total score over all judges. In international figure skating contests, still more involved

rules dictate the way the scores given by the twelve judges to each of the many parts of a competitor's performance become the number grades that are their input messages. The output is a rank-ordering determined by first, eliminating the grades of three judges chosen at random; second, eliminating the highest and lowest of the grades that are left; third, ranking the competitors according to the averages of the seven remaining grades. This complicated procedure is intended to combat judges who manipulate their inputs to favor or disfavor one or another competitor (the piano, wine, and figure skating mechanisms are described in detail in chapter 7).

In Australian elections, a voter's input is a complete rank-ordering of the candidates, and the output is a winner. But in most countries a voter's input message is at most one vote for one candidate, and the output is a winner, the candidate with the most votes; or the output is a ranking determined by the candidates' respective total votes. "Approval voting" is a relatively new mechanism used by several professional scientific societies: a voter's input message is one vote or none for every candidate, the output is a winner or a ranking determined by the candidates' respective total votes. These electoral schemes—each a pure invention to elicit the opinions of voters—offer an extremely limited possibility for voters to express themselves (various voting mechanisms used in practice are described in detail in chapter 2; other traditional methods in chapters 3 and 4; approval voting is discussed and analyzed in chapter 18; point-summing methods in chapter 17).

In fact, every mechanism generates information, notably the candidates' total scores, that in many situations may be viewed as constituting a part of the genuine outputs (see chapter 20).

1.2 Messages of a Common Language

Practice—with the notable exception of elections—suggests that letters (e.g., from a high of *A* to a low of *E*), descriptive words or phrases (e.g., from *Excellent* to *Bad*), or numbers (e.g., from 100 to 0) that define an ordered scale provide judges with a *common language* to grade and to rank competitors in a host of different settings. Typically, such languages are invented to suit the purpose, and carefully defined and explained. Their words are clearly understood, much as the words of an ordinary language or the measurements of physics (e.g., temperatures in degrees centigrade or Fahrenheit). A judge's input *message* is a grade or word of the common language. These 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. The common languages used by judges in wine, figure skating, diving, and other competitions are described in chapter 7; their connections with measurement in general are discussed in chapter 8.

Judges and voters have complex aims, ends, purposes, and wishes: their *preferences* or *utilities*. A judge's or a voter's preferences may depend on many factors, including his beliefs about what is right and wrong, about the common language, about the method that transforms input messages into decisions, about the other judges' or voters' acts and behaviors, in addition to his evaluations of the competitors or candidates. But the judges' or voters' input messages—the grades they give—are assuredly not their preferences: a judge may dislike a wine, a dive, or a part of a skater's performance yet give it a high grade because of its merits; or a judge may like it yet give it a low grade because of its demerits. Rules and regulations define how certain performances are to be evaluated, yet votes can be strategic. The fact that voters or judges share a common language of grades makes no assumptions about their preferences or utilities. Utilities are measures of the judges' or voters' satisfaction with the output, the decision of the jury or the society; grades are measures of the merits of competitors used as inputs. A judge's or a voter's input message is chosen strategically: depending on the mechanism for transforming inputs into an output, a judge may exaggerate the grades he gives, upward or downward, in the hopes of influencing the final result.

Arrow's theorem plays an important role in this approach as well. It proves in theory what practitioners intuitively have learned by trial and error: without a common language there can be no consistent collective decision (see chapter 11, theorem 11.6a). Its true moral is that judges and voters *must* express themselves in a common language.

1.3 Majority-Grade

The fundamental problem is to find a *social decision function*: a method whose inputs are the grades of a common language and whose outputs are jury or electoral decisions, namely, final-grades and/or rank-orderings of competitors or candidates. Our theory shows there is one best method of assigning a final grade to each competitor or candidate—the *majority-grade*—and one best method of assigning a "generalized final grade" that ranks the competitors or candidates—the *majority-ranking*, where the winner is the first-place competitor or candidate. The definitions of these two terms are simple and have already

been accepted in practice (e.g., for wine tasting; see Peynaud and Blouin 2006, 104–107). A host of different arguments prove they are the only social decision functions that satisfy each of various desirable properties.

Two supplementary concepts explained below are linked to the majority-ranking. The *majority-value* is a sequence of grades that determines the majority-ranking. The *majority-gauge* is a simplified majority-value that is sufficient to determine the majority-ranking when the number of judges or the electorate is large.

To begin, the aim is to decide on a final grade, given the individual messages of all the judges. Suppose the common language is a set of ten integers {0, 3, 5, 6, . . . , 11, 13} (from worst to best), the system of school grades previously used in Denmark (it is amusing that 1, 2, 4, and 12 are missing, the reasons for which are explained in chapter 8; but this has no influence on the present discussion). Imagine that the grades given to a competitor by all the judges are listed in ascending order from worst to best. When the number of judges is odd, the *majority-grade* α is the grade that is in the middle of the list (the median, in statistics). For example, if there are nine judges who give a competitor the grades {7, 7, 8, 8, **8**, 9, 10, 11, 11}, the competitor's majority-grade is 8. When the number of judges is even, there is a *middle-interval* (which can, of course, be reduced to a single grade if the two middle grades are the same), and the *majority-grade* α is the lowest grade of the middle-interval (the "lower median" when there are two in the middle). For example, if there are eight judges who give a competitor the grades {7, 7, 8, **8**, 11, 11, 11, 13}, the middle-interval goes from 8 to 11 and thus is the set of grades {8, 9, 10, 11}, and the competitor's majority-grade is 8.

The majority-grade α of a competitor is the highest grade approved by an absolute majority of the electors: more than 50% of the electors give the competitor at least a grade of α, but every grade lower than α is rejected by an absolute majority. Thus the majority-grade of a competitor is the final grade wished by the majority. In the first example, {7, 7, 8, 8, **8**, 9, 10, 11, 11}, only two (of nine) judges would vote for a lower grade, and only four for a higher grade. In the second example, {7, 7, 8, **8**, 11, 11, 11, 13}, only two (of eight) judges would vote for a lower grade, and only four for a higher grade.

The choice of the smallest grade of the middle-interval when the number of judges is even is the logical consequence of a principle of consensus. Compare two competitors A and B when there is an even number of judges: if all of A's grades strictly belong to the middle-interval of B's grades, then since there is a greater consensus among the judges for A's grade than for B's grade, A should be ranked at least as high as B. For example, if B's grades are {7, 7, 8, **8**, 11, 11, 11, 13} (the second example) and all of A's grades are

either 9 or 10 (e.g., {9, 9, 9, 9, 9, 10, 10, 10}) and thus strictly belong to B's middle-interval, then A should rank higher than B.

The majority-grade is necessarily a word that belongs to the common language, and it has an absolute meaning. When an absolute majority of the judges give a competitor a particular grade α, then the competitor's majority-grade must necessarily be α, for if the number of judges is odd, the middle grade is necessarily α, and if the number of judges is even, the two middle grades are necessarily α.

It is reasonable to suppose that if a judge wishes that a competitor be accorded a certain grade—say, a 9 in the Denmark school scale—then the more the competitor's final grade deviates from 9, the greater will be the judge's discontent. When this is true, the best strategy for a judge is always to assign the grade that she believes the competitor merits, neither more nor less. For suppose that a judge believes that a candidate merits a grade of 9. If the majority-grade was higher, say, 11, she might be tempted, in anticipation of such an outcome, to assign a lower grade than 9. But doing so would change nothing because the majority-grade 11 would resolutely remain in the middle whatever lower grade she chose to give. If, on the other hand, the majority-grade was lower, say, 7, she might anticipate the outcome and be tempted to assign a higher grade than 9. Again this would change nothing because the majority-grade 7 would stay in the middle. The only other possibility is that the judge assigns a grade of 9 and the majority-grade is 9; in this case the judge is completely content. Thus in any case honesty is the best policy.

1.4 Majority-Ranking

In some applications, a complete ranking among all competitors or alternatives is not sought. For example, most wine competitions only wish to discern one of four medals (grand gold, gold, silver, bronze) or none. In many applications, however, notably in sports competitions, a complete rank-ordering is essential.

When two competitors have different majority-grades, the competitor with the higher grade is naturally ranked higher. Suppose, however, that two competitors A and B have the same majority-grade. For example, A's grades are {7, 9, **9**, 11, 11} and B's are {8, 9, **9**, 10, 11}, so they both have a majority-grade of 9. How are they to be compared? Their common (first) majority-grade is dropped (a single one) because it has already yielded all the information it can give relevant to comparing A with B, and the majority-grades of the grades that remain to each competitor—their second majority-grades—are found. In this example, A's remaining grades are {7, **9**, 11, 11} and B's are {8, **9**, 10, 11}, so

their second majority-grades are both again 9. If one were higher than the other, it would designate the competitor who is ranked higher. If, as here, the second majority-grades are the same, they are discarded, and the third majority-grades of the competitors—the majority-grades of the grades that remain—are found, and so on, until one competitor is ranked ahead of the other. In this case, *A*'s third majority-grade is 11 and *B*'s is 10, so *A* ranks above *B*. One of the two must be ranked ahead of the other unless the competitors have identical sets of grades. This defines the *majority-ranking*.

1.5 Majority-Value

A competitor's *majority-value* is the sequence of his (first) majority-grade, his second majority-grade, his third majority-grade, down to his *n*th majority-grade (if there are *n* judges). Continuing with the example of section 1.4, *A*'s majority-value is the ordered sequence of grades {09, 09, 11, 07, 11} and *B*'s is {09, 09, 10, 08, 11}. The lexicographic order of the majority-values gives the majority-ranking of the competitors. Thus *A* ranks higher than *B* because *A*'s first grade in the sequence where their grades differ is higher than *B*'s. When the common language is a set of integers, as in this case, the majority-value may simply be written as a number—in this example, *A*'s is 09.09110711 and *B*'s is 09.09100811—and the magnitudes of the majority-values determine the majority-ranking of the competitors. Moreover, dividing by 1.01010101 rescales the majority-values so that the minimum is 0, the maximum 13, and a competitor assigned the same grade α by all judges has a rescaled majority-value of exactly α. In this case *A*'s rescaled majority-value is 9.000196 and *B*'s is 9.000098. And if the set of grades were {09, 09, 09, 09, 09}, the rescaled majority-value would be 9.

The *majority-values* summarize all the results of an election or a competition. The first term of the sequence of grades of a competitor is his majority-grade (and when the grades are integers and the majority-value is written as a value or number, its integer part is his majority-grade). The lexicographic order among the sequences of grades that are the majority-values is the majority-ranking (and when the grades are integers and the majority-values are written as values or numbers, the majority-values determine the majority-ranking).

To more clearly understand these assertions, suppose four competitors, *A*, *B*, *C*, and *D* receive the Danish-style grades from nine judges 1, 2, …, 9 (table 1a). Reorder the grades of each competitor from highest to lowest (table 1.1b). The competitors' majority-grades are 10 for *A*, 9 for *B* and *C*, and 8 for *D*. Their majority-values (written as numbers with only the needed precision to

Table 1.1a
Hypothetical Example: Four Competitors, Nine Judges

Judge	*A*	*B*	*C*	*D*
1	13	09	07	08
2	10	09	13	05
3	09	08	11	13
4	10	11	09	08
5	04	10	09	09
6	13	08	00	02
7	11	07	10	13
8	10	11	09	08
9	10	11	11	07

Table 1.1b
Hypothetical Example: Grades Ordered from Best to Worst

A	*B*	*C*	*D*
13	11	13	13
13	11	11	13
11	11	11	09
10	10	10	08
10	**09**	**09**	**08**
10	09	09	08
10	08	09	07
09	08	07	05
04	07	00	02

Note: Majority-grades are shown in boldface.

distinguish their order) determine the majority-ranking among them (where $X \succ_S Y$ means the society or jury S prefers X to Y):

Majority-ranking:	A	$\succ_S$	C	$\succ_S$	B	$\succ_S$	D
Majority-value:	$10.\ldots$	$>$	$9.091009\ldots$	$>$	$9.091008\ldots$	$>$	$8.\ldots$

Suppose, instead, that wines are to be judged, and the common language goes from *Excellent, Very Good, Good, Passable, Inadequate*, and *Mediocre* to *Bad*. The usual standard is five judges, so three wines—a St. Amour, a Bourgueil, and a Cahors—could receive the (already reordered) mentions shown in table 1.2. They all have the same majority-grade: *Good*. Their majority-values (written as a sequence of mentions to the needed precision to distinguish their order) determine the majority-ranking among them:

St. Amour	$\succ_S$	Bourgueil	$\succ_S$	Cahors
Good–Good–Very Good–Passable	$>$	*Good–Good–Very Good–Mediocre*	$>$	*Good–Passable*

Table 1.2
Hypothetical Example: Three Wines, Five Judges

St. Amour	Bourgueil	Cahors
Very Good	*Excellent*	*Excellent*
Very Good	*Very Good*	*Excellent*
Good	***Good***	***Good***
Good	*Good*	*Passable*
Passable	*Mediocre*	*Mediocre*

Note: Majority-grades are shown in boldface.

The first mention where two majority-values differ determines the order: thus the St. Amour and the Bourgeuil have the same mentions in the first three positions, and the fourth determines which ranks ahead of the other.

Two important points may be made immediately. First, if one or several competitors withdraw, the majority-ranking among the remaining competitors necessarily agrees with the majority-ranking among all competitors. So the majority-ranking satisfies Arrow's independence of irrelevant alternatives (IIA) condition: the relative positions of two competitors in the majority-ranking do not depend on the merits of another competitor. This is decidedly not the case with most voting mechanisms used in practice, as the United States learned in the presidential election of 2000 (see chapter 2), nor is it the case with the methods traditionally used to rank figure skaters (see chapter 7).

Second, what is the aim of a competition (or of an election)? It is to reach a *consensual decision*. The jury (or the society) seeks to find agreement. The majority-ranking makes the effect of middle grades more decisive. A few extreme or "cranky" evaluations should have a less decisive effect, though of course, every grade counts. If after the k best and the k worst grades of two competitors are dropped, the grades of a woman rank her ahead of a man by the majority-ranking, then she is ranked ahead of him by the majority-ranking of the jury (or the society). To see this in the first example of this section, where there are nine judges and four competitors, drop the two best and the two worst grades of competitors B and C. C is ranked ahead of B, $C \succ_S B$, because on the basis of the remaining five grades the majority-ranking puts C ahead of B (in this case their grades are all the same except for one, an 8 for B and a 9 for C). It has already been pointed out that the majority-ranking satisfies another such property, namely, when the number of judges is even and all the grades of one competitor A strictly belong to the middle-interval of the grades of another competitor B, then A should be ranked ahead of B. It is proven that the majority-ranking is the *only* method that satisfies these two properties.

1.6 Majority-Gauge

Juries or committees usually have a small number of judges or members: five, nine, perhaps twenty. The method and theory are the same when juries are any numbers of judges or voters. However, the majority judgment for "juries" composed of hundreds or millions of judges—nations electing presidents, cities electing mayors, congressional districts electing representatives, institutions and societies electing officers—has an easier and more compelling description in that context.

Majority judgment was tested in a field experiment on April 22, 2007, in parallel with the first round of the French presidential election, in Orsay, a town close to Paris. In French presidential elections, an elector casts her vote for one candidate (or none). If no candidate receives an absolute majority of the votes, a second round is held two weeks later between the two candidates who had the most votes in the first round. The results of the first round are used to explain the approach.

The experiment took place in three of the twelve voting precincts of Orsay that together had 2,695 registered voters. Of these, 2,383 voters cast official ballots (88% of those registered), of which 2,360 were valid. After voting officially, the voters were asked to participate in the experiment using the majority judgment. They had been informed about it by mail, printed flyers, and posters. It was conducted in accordance with usual French voting practice: ballots were filled out and inserted into envelopes in voting booths with curtains, then deposited in transparent ballot boxes.

The ballot is reproduced in figure 1.1 (the names of the candidates are given in the official order, the result of a random draw). A serious and solemn question was posed to the voters,

To be president of France,
after having taken every consideration into account,
I judge in conscience that this candidate would be

and asked to give an answer for every candidate in a common language of grades—absolute evaluations—common to all French voters:

Très Bien, Bien, Assez Bien, Passable, Insuffisant, or *A Rejeter*.

The first five designations are known to all those who have been school children in France; the last is clear enough. Reasonable translations are:

Excellent, Very Good, Good, Acceptable, Poor, or *To Reject.*

Bulletin de vote:
Élection du Président de la République 2007

Pour présider la France,
ayant pris tous les éléments en compte,
je juge en conscience que ce candidat serait:

	Très Bien	*Bien*	*Assez Bien*	*Passable*	*Insuffisant*	*A Rejeter*
Olivier Besancenot						
Marie-George Buffet						
Gérard Schivardi						
François Bayrou						
José Bové						
Dominique Voynet						
Philippe de Villiers						
Ségolène Royal						
Frédéric Nihous						
Jean-Marie Le Pen						
Arlette Laguiller						
Nicolas Sarkozy						

Cochez une seule mention dans la ligne de chaque candidat.
Ne pas cocher une mention dans la ligne d'un candidat revient à le Rejeter.

Figure 1.1
Ballot, Orsay experiment, 2007 French presidential election.

The meanings of the grades are directly related to the question posed.

The sentences at the bottom of the ballot say, "Check one grade in the line of each candidate. No check in the line of a candidate means To Reject him." We believe that every voter must be required to evaluate every candidate. A voter having no opinion concerning a candidate has not even taken the time to evaluate him and thus has implicitly rejected him (other possibilities are discussed in chapters 13 and 14).

Of the 2,383 persons who cast official ballots, 1,752 participated in the experiment (74%). In fact, the rate of participation was slightly higher because in France a voter is permitted (under certain conditions) to ask someone else to vote in his place, but no one was allowed to vote twice in the experiment. Nineteen ballots were invalid, usually because more than one grade was assigned to a candidate, leaving 1,733 valid majority judgment ballots. The results are given in table 1.3.

Table 1.3
Majority Judgment Results, Three precincts of Orsay, April 22, 2007

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	No Grade
Bayrou	13.6%	30.7%	25.1%	14.8%	8.4%	4.5%	2.9%
Royal	16.7%	22.7%	19.1%	16.8%	12.2%	10.8%	1.8%
Sarkozy	19.1%	19.8%	14.3%	11.5%	7.1%	26.5%	1.7%
Voynet	2.9%	9.3%	17.5%	23.7%	26.1%	16.2%	4.3%
Besancenot	4.1%	9.9%	16.3%	16.0%	22.6%	27.9%	3.2%
Buffet	2.5%	7.6%	12.5%	20.6%	26.4%	26.1%	4.3%
Bové	1.5%	6.0%	11.4%	16.0%	25.7%	35.3%	4.2%
Laguiller	2.1%	5.3%	10.2%	16.6%	25.9%	34.8%	5.3%
Nihous	0.3%	1.8%	5.3%	11.0%	26.7%	47.8%	7.2%
de Villiers	2.4%	6.4%	8.7%	11.3%	15.8%	51.2%	4.3%
Schivardi	0.5%	1.0%	3.9%	9.5%	24.9%	54.6%	5.8%
Le Pen	3.0%	4.6%	6.2%	6.5%	5.4%	71.7%	2.7%

Note: When there was no grade, it was counted as a *To Reject*, as per the instructions on the ballot (so Bayrou's *To Reject* was counted as 7.4%, Royal's as 12.6%, . . . , Le Pen's as 74.4%). There were few.

The Orsay experiment is discussed in greater detail in chapters 6, 15, and 19. It is used here as a vehicle to explain the procedure when there is a large electorate or jury, and to remark on several of its features.

When there are a very large number of voters or judges, as in this case, generically (almost surely) the middle-interval—a single grade if the number of voters is odd, two grades if the number of voters is even—will be one and the same grade. Thus it is safe to simply say that a candidate's *majority-grade* is the median of his grades: it is at once the highest grade approved by a majority and the lowest grade approved by a majority. Alternatively, only a minority would be for a higher grade or for a lower grade. For example, the majority-grade of D. Voynet (see table 1.3) is *Acceptable* because a majority of $53.4\% = 2.9\% + 9.3\% + 17.5\% + 23.7\%$ of the voters judge that she merits at least an *Acceptable*, and a majority of $70.3\% = 23.7\% + 26.1\% + 16.2\% + 4.3\%$ of the voters judge that she merits at most an *Acceptable*. Or, only a minority of 29.7% would be for a higher grade and only a minority of 46.6% for a lower grade.

The *majority-ranking* is calculated more directly in the case of a large number of voters. When the numbers or percentages of grades higher or lower than the candidates' majority-grades are different, which is almost surely true, the majority-ranking is obtained from three pieces of information concerning each candidate:

· p, the number or percentage of the grades better than a candidate's majority-grade;

- α, the candidate's majority-grade;
- q, the number or percentage of the grades worse than a candidate's majority-grade.

The triple (p, α, q) is the candidate's *majority-gauge*. S. Royal's majority-gauge (see table 1.3) is (39.4%, *Good*, 41.5%) since $39.4\% = 16.7\% + 22.7\%$ of her grades are better than *Good*, and $41.5\% = 16.8\% + 12.2\% + 10.8\% + 1.8\%$ are worse than *Good*. If the number or percentage p of the grades better than a candidate's majority-grade α is higher than the number or percentage q of those worse than the candidate's majority-grade, then the majority-grade is completed by a plus $(+)$; otherwise the majority-grade is completed by a minus $(-)$. Thus S. Royal's majority-grade is *Good*$-$. The plus or minus attached to the majority-grade is implied by the majority-gauge, so it is not necessary to show it, but for added clarity it is most often included, so that, for example, Royal's majority-gauge may be written (39.4%, *Good*$-$, 41.5%).

Naturally a *majority-grade*$+$ is ahead of a *majority-grade*$-$ in the majority-ranking. Of two *majority-grade*$+$'s, the one having the higher number or percentage of grades better than the *majority-grade* is ahead of the other; of two *majority-grade*$-$'s, the one having the higher number or percentage of grades worse than the *majority-grade* is behind the other. For example, in table 1.4, S. Royal and N. Sarkozy both have the majority-grade *Good*$-$. Royal has 41.5% worse than *Good*, and Sarkozy has 46.9% worse than *Good*, so Royal finishes ahead of Sarkozy. O. Besancenot and M.-G. Buffet both have the majority-grade *Poor*$+$. Besancenot has 46.3% better than *Poor*, and Buffet has 43.2% better than *Poor*, so Besancenot finishes ahead of Buffet. To see a candidate's majority-gauge, imagine a see-saw or teeterboard with all the voters lined up according to the grades they give, from best to worst. Assuming each voter's weight is the same, the grade given by the voter who stands at the fulcrum where the board is in perfect balance is the majority-grade. Remove, now, all voters who gave the majority-grade and place the fulcrum at the juncture between the better than and worse than majority-grade. If the board tilts to the better grades, a plus $(+)$ is accorded, and the more it tilts, the better is the majority-gauge. If it tilts to the worse grades, a minus $(-)$ is accorded, and the more it tilts, the worse is the majority-gauge.

The majority-gauges and the majority-ranking of the experiment are shown in table 1.4. It may be seen that the majority-ranking is quite different than the order of finish given by the official vote in these three voting precincts (which are *not* representative of the vote in all of France; see the last two columns of table 1.4). This is due to the fact that the majority judgment allows voters to

Table 1.4
The Majority-Gauges (p, α, q) and the Majority-Ranking, Three Precincts of Orsay, April 22, 2007

		p Better Than Majority-Grade	α Majority-Grade	q Worse Than Majority-Grade	Official Vote, 3 Precincts	Official National Vote
1st	Bayrou	44.3%	*Good+*	30.6%	25.5%	18.6%
2d	Royal	39.4%	*Good−*	41.5%	29.9%	25.9%
3d	Sarkozy	38.9%	*Good−*	46.9%	29.0%	31.2%
4th	Voynet	29.7%	*Acceptable−*	46.6%	1.7%	1.6%
5th	Besancenot	46.3%	*Poor+*	31.2%	2.5%	4.1%
6th	Buffet	43.2%	*Poor+*	30.5%	1.4%	1.9%
7th	Bové	34.9%	*Poor−*	39.4%	0.9%	1.3%
8th	Laguiller	34.2%	*Poor−*	40.0%	0.8%	1.3%
9th	Nihous	45.0%	*To Reject*	–	0.3%	1.1%
10th	de Villiers	44.5%	*To Reject*	–	1.9%	2.2%
11th	Schivardi	39.7%	*To Reject*	–	0.2%	0.3%
12th	Le Pen	25.7%	*To Reject*	–	5.9%	10.4%

express their opinions on all the candidates rather than simply singling out one among them.

The reasons for believing in the validity of the experiment are given later, but several salient observations are made here.

· More than one of every three participants gave their highest grade to two or more candidates.

· Only half of the voters used a grade of *Excellent*.

· On average, a voter gave a grade of *To Reject* to over one-third of the candidates.

This proves that voters do not have in their minds rank-orderings of the candidates (and still more evidence supports this claim). A rank-ordering does not allow a voter to express an equal evaluation of candidates, or an intensity of appreciation, or an outright rejection. It also shows that the actual system—cast one vote for one among several or many candidates—forced one-third of the voters to opt for a candidate when in fact they saw no difference among two or more of them. These observations are convincing because voters had no incentive to vote strategically since they were only participating in an experiment. It is precisely such experiments that can elicit the true opinions of voters.

Strategic voting played an important role in the French presidential election of 2007. Voters had in mind what had happened in 2002, when the vote to the left was so widely distributed among some eight candidates that instead of a second

round between Jacques Chirac (the incumbent president and major candidate of the right) and Lionel Jospin (the standing prime minister and major candidate of the left), Chirac was pitted against Jean-Marie Le Pen, the perennial candidate of the extreme right (see chapter 2). It seems safe to assert that in the first round of 2007 a significant number of voters did not vote for the candidate they preferred. Instead of voting for their favorite—an ecologist, a communist, or a Trotskyist—many voters of the left opted for Ségolène Royal, the socialist, the major candidate of the left. And the same phenomenon occurred on the right: it seems that many voters abandoned the extreme right to vote for the major candidate of the right, the U.M.P. candidate, Nicolas Sarkozy. In contrast, the majority judgment encourages a voter to express his convictions (as is proven via several precise criteria that are defined in later chapters).

To see how the majority judgment resists strategic manipulation in the context of elections, take a candidate, say Ségolène Royal, whose majority-grade is *Good* − and whose majority-gauge is

$$(39.4\%,\ \textit{Good},\ 41.5\%).$$

Then, 39.4% of her grades are better than *Good*, 41.5% are worse than *Good*, so 19.1% are *Good*. Who are the voters who can change Royal's majority-gauge by changing the grades they give her, and what are their motivations to change?

Suppose a voter believes a candidate merits a grade of α, and the further the majority-grade is from α, the less he likes it (a reasonable motivation: the voter's preferences in grading are then said to be single-peaked). Then, as was seen, the voter's optimal voting strategy is simply to give the candidate the grade α: the majority judgment is *strategy-proof-in-grading*.

Similar reasoning shows that the majority-grade mechanism is *group strategy-proof-in-grading*. A group of voters who share the same beliefs (e.g., they belong to the same political party) has the same optimal strategy, namely, to give to the candidates the grades it believes they merit. For if the group believed that Royal merited better than *Good*, and all raised the grade they gave her, her majority-gauge would remain the same (p does not change). If all lowered the grade they gave her, her majority-gauge would decrease (q increases), and perhaps her majority-grade would be lowered (not their intent). If the group believed that Royal merited worse than *Good*, and all lowered the grade they gave her, her majority-gauge would remain the same (q does not change). If all raised the grade they gave her, her majority-gauge would increase (p increases), and perhaps her majority-grade would be raised (not their intent).

These strategy-proof-in-grading properties are *not* true of any of the mechanisms currently used today. The strategy of a voter may, however, focus on the

final ranking of the candidates rather than on their final grades. It is impossible to completely eliminate the possibility of strategic manipulation if a voter is prepared for a candidate's final grade to be either above or below what he thinks the candidate merits: there is no mechanism that is strategy-proof-in-ranking. The majority judgment does not escape the Gibbard-Satterthwaite impossibility theorem (see chapters 5 and 13), but it best resists such manipulation.

One means by which it resists is easy to explain. Take the example of Bayrou with a *Good*+ and Royal with a *Good*− (see table 1.4); their respective majority-gauges are

Bayrou: (44.3%, *Good*, 30.6%) Royal: (39.4%, *Good*, 41.5%).

How could a voter who graded Royal higher than Bayrou manipulate? By changing the grades assigned to try to lower Bayrou's majority-gauge and to raise Royal's majority-gauge. But the majority judgment is *partially strategy-proof-in-ranking*: those voters who can lower Bayrou's majority-gauge cannot raise Royal's, and those who can raise Royal's majority-gauge cannot lower Bayrou's. For suppose a voter can lower Bayrou's. Then she must have given Bayrou a *Good* or better; but having preferred Royal to Bayrou, the voter gave a grade of better than *Good* to Royal, so she cannot raise Royal's majority-gauge (cannot raise her p). Symmetrically, a voter who can raise Royal's majority-gauge must have given her a *Good* or worse and thus to Bayrou a worse than *Good*; so the voter cannot lower Bayrou's majority-gauge (cannot increase his q).

Compared with other mechanisms, the majority judgment cuts in half the possibility of manipulation, however bizarre a voter's motivations or whatever her utility function. The majority judgment resists manipulation in still other ways that other methods do not, but to see how requires information found in voters' individual ballots that is not shown in the elections results of table 1.4. For example, significant numbers of voters cannot contribute at all either to raising Royal's majority-gauge or to lowering Bayrou's (28% of those who graded Royal above Bayrou). Moreover, those who can manipulate have no incentive to exaggerate very much in any case, for it does not pay to do so (a more detailed analysis is given in chapter 19).

Some critics have averred that a voter should be forced to "make up his mind" by expressing a clear-cut preference for one candidate. The first-past-the-post system has this property (unless the voter abstains or hands in a blank ballot), as does any mechanism in which the input is a rank-order of the candidates. Both types of mechanism prevent the voter from expressing any intensity of preference: the second-ranked candidate is only that, whatever the voter's

evaluation. But why limit any voter's freedom of expression? Shouldn't someone who sees no discernible difference between two or more candidates be allowed to record this? Shouldn't a voter who believes his second-ranked candidate is merely acceptable or worse be allowed to say so? The majority judgment gives voters complete freedom of expression (within the bounds of the language).

Voters who participated in the Orsay experiment were delighted with the idea that a candidate could be assigned a final grade. The majority-grade is an important signal that expresses the electorate's appreciation of a candidate. Chirac's "triumph" against Le Pen in 2002—a majority of over 80% in the second round—would have been very different had the majority judgment been used: he surely would have won, but with a middling grade—perhaps an *Acceptable* or a *Good* to Le Pen's *To Reject*—that would have given a more sober sense to his reelection. In the election of 2007, Voynet's majority-grade of *Acceptable* and her fourth place in the majority-ranking more clearly express the electorate's concern with environmental issues than her eighth-place finish in the official national vote. Le Pen's last place in 2007 and solid *To Reject* evaluation shows the electorate's strong rejection of his ideas, whereas the official vote makes him one of the four major contenders. Even when there is only one contender, a not infrequent occurrence—in the 2002, 2004, and 2006 U.S. congressional elections, respectively 81, 66, and 59 candidates were elected with no Democratic or Republican opponent—the majority judgment establishes the esteem in which the candidate is held.

U.S. presidential primaries leap to mind as an immediately realistic application. The majority judgment would be relatively easy to implement since the decision to do so may be taken at the state level. It would permit a real expression of the voters' opinions versus sending a message consisting of one name. With as many as five to ten candidates, the first-past-the-post system drastically curtails expression of the voters' opinions: a "big winner" often garners as little as 25% of the total vote, hardly a mandate to be singled out as the principal candidate. Indeed, the luck of the draw may determine the "winner" due to the mere presence of strategic candidates who have no real chance of emerging as real contenders. Finally, and of real importance, the current system is divisive for a political party: it opts for one candidate and throws out the others. With the majority judgment, candidates are not rejected: many, perhaps all, receive good majority-grades, yet one is singled out because he is first in the majority-ranking.

A very small-scale experiment was conducted on the Web in late September, early October 2008. Members of the Institute for Operations Research and the Management Sciences (INFORMS), a scientific society, were asked,

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:

	Excel-lent	*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*.

Figure 1.2
Ballot, U.S. presidential election, INFORMS experiment, September–early October 2008.

Suppose that instead of primary elections in states to designate candidates, then national elections to choose one among them, the system was one national election in which all eligible candidates are presented at once. Or, suppose you are in a state holding a primary where you are asked to evaluate the candidates of all parties (at least one state primary votes on all candidates at once). A possible slate of candidates for President of the United States could be [here followed the names of the eight candidates given in the ballot together with their affiliations.]

They were instructed,

You will be asked to evaluate each candidate in a language of grades. A candidate's majority-grade is the middlemost of her/his grades (or the median grade). The candidates are ranked according to their majority-grades. The theory provides a natural tie-breaking rule.

Then they were invited to vote with the ballot shown in figure 1.2.

This experiment was certainly not representative of the U.S. electorate (nor was it meant to be). A large majority of the members of INFORMS are U.S. citizens, but many members are citizens of other nations. The results, shown in table 1.5, are nevertheless of interest. In this case the winner stands out as the only candidate with a *Very Good*, and the collective opinion of those who voted is quite clear.

Table 1.5
The Majority-Gauges (p, α, q) and the Majority-Ranking, INFORMS Experiment, September–Early October 2008

		p Better Than Majority-Grade	α Majority-Grade	q Worse Than Majority-Grade
1st	Barack H. Obama	35.9%	*Very Good+*	32.0%
2nd	Hillary R. Clinton	45.0%	*Good+*	33.6%
3rd	Colin L. Powell	32.8%	*Good−*	41.2%
4th	Michael R. Bloomberg	42.0%	*Acceptable+*	31.3%
5th	John R. Edwards	36.6%	*Acceptable+*	32.8%
6th	John S. McCain	33.4%	*Acceptable−*	44.2%
7th	W. Mitt Romney	46.6%	*Poor+*	22.9%
8th	Michael D. Huckabee	33.5%	*Poor−*	47.3%

The descriptions of the majority judgment given in this chapter should permit its use in any application—with few judges or many voters—given that a common language of grades has been defined and explained.

1.7 Nomenclature

Majority judgment is the name we have chosen to give to the method we advocate. It encompasses several key ideas, each endowed with a name that is used again and again throughout the book.

· *Majority-grade* The middlemost or median of the grades given to a competitor by the judges or voters; when there is an even number of judges or voters, the lower middlemost of the grades (see section 1.3 and chapter 12).

· *Majority-ranking* The majority judgment ranking of all competitors on the basis of their grades (see section 1.4 and chapter 13). They are ranked according to their majority-values or majority-gauges.

· *Majority-value* The sequence of a competitor's grades consisting of his first majority-grade, his second majority-grade, …, and so on, up to his nth majority-grade (when there are n judges or voters). Competitor A ranks higher than competitor B in the majority-ranking if and only if A's majority-value is lexicographically above B's (see section 1.5 and chapter 13).

· *Majority-gauge* A simplification of the majority-value from which may be deduced the majority-ranking among the competitors in many cases; in particular, when there are many judges or voters such as in most elections. However, when two competitors are tied with the same majority-gauge, they are not necessarilty tied with their majority-values (see section 1.6 and chapter 14).

· *Abbreviated majority-value* An abbreviated but complete expression of the majority-value (see chapter 14).

· *kth-order function*. It singles out a competitor's kth-highest grade. When there are n judges or voters, there are n order functions. The first-order function is the highest grade, the nth-order function is the lowest grade. Among them the majority-grade is the $\left(\frac{n+1}{2}\right)th$-order function when n is odd and the $\left(\frac{n+2}{2}\right)th$-order function when n is even.

1.8 The Thesis

The intent of this book is to show why the majority judgment is superior to any known method of voting and to any known method of judging competitions.

To do so, it presents the fundamentals of the traditional theory of social choice, describing the principal known methods, together with simple proofs of the most important results. It also proposes new characterizations, new incompatibility theorems, and new methods in the traditional model of social choice. Actual voting systems and methods used in practice to judge competitors (e.g., wines, divers, figure skaters) are also described to show how they are in fact subject to all the paradoxes and failures that are identified in theory. Throughout, theory, experiments, and practical evidence in voting and in judging competitions are provided to support the first central point: *the traditional model is a bad model, in theory and in practice.*

The new model is then developed. It is shown that a host of properties uniquely characterize the "order functions" as the only methods that can be used. This is established from a variety of points of view. Practice again plays a central role: experimental evidence is given that shows the majority judgment is a practical method and that common languages in voting and in judging competitions do in fact exist and can be meaningfully defined. Statistical comparisons that depend on real data are made with other methods to show why the majority judgment is better in voting; in particular, approval voting is shown to fail. Often, judging competitions (such as wine, ski jumping, or figure skating) invoke several different criteria: the majority judgment is generalized to such situations. Throughout, theory, experiments, and practical evidence are provided to support the second central point: *the majority judgment is a better alternative to all other known methods, in theory and in practice.*

2 Voting in Practice

Popular election thus practiced, instead of a security against misgovernment, is but an additional wheel in its machinery.
—John Stuart Mill

Voting in practice invokes issues that go well beyond the problem of how to elect one candidate among several or how to determine their order of finish. Many candidates are elected as the representatives of regions (constituencies, congressional districts, states, departments, provinces, or nations) to legislatures (Assemblées nationales, Diets, Houses of Representatives, Knessets, Parliaments, or Senates), or as the representatives of political parties, or of both regions and parties, to legislatures. A multitude of different systems are used; they raise different problems, invoke different information, ask for different inputs, and are resolved with different mechanisms. Nevertheless, several central problems are common to many electoral systems.

First and foremost among them is the subject of this book, the problem of *social choice*: to elect one candidate and rank all candidates. The *apportionment* of a legislature is a second major problem (see Balinski and Young 1982). In one guise, apportionment addresses how to allocate a fixed number of seats among regions according to their respective numbers of inhabitants. The implicit ideal is an allocation of seats that is proportional to the number of inhabitants, though that ideal cannot be realized perfectly. In a second guise, it concerns what is commonly called *proportional representation*: the apportionment of seats to political parties according to their respective votes. In this case the ideal is explicit, but "proportional to votes" has no more intrinsic significance than is to be found in the input messages that are the votes. A third important problem is *districting* or *redistricting*: given a region that has been apportioned k seats in a legislative body, it is to be cut up into k single-member districts of approximately equal population, each of which is a collection of

administratively defined geographic areas (e.g., counties, townships, cantons) that form a connected area.

Theory has had relatively little impact on how elections take place in practice. This is due to two principal reasons. First, voting theory's most famous and most important result on how one candidate is to be selected among several—Arrow's impossibility theorem—says that there is no perfect method, so there has been no definitive, practical, scientifically established advice to be given. The theory of apportionment has given clear-cut answers to the principal problems, in particular when it concerns allocating representation to regions, but it too has an impossibility theorem. Although, unlike Arrow's impossibility, it has no realistic bite, it provides an excuse to say that there is no perfect method. As for the problem of districting, there is no theory, only a wide variety of computational approaches.

The second major reason that theory has had little effect on elections is a consequence of the first: politicians are pleased with this state of affairs. The imperfections of theory, real or supposed, or its absence allows politicians to change or manipulate to their advantage the systems by which they are elected. The abuses are notorious and repeated. The evidence for these claims is plentiful (see Balinski 2004). Politicians have no shame in being at once the players and the referees of the electoral game, and they indubitably eschew outside advice. As A. de Tocqueville so aptly remarked in a letter to his cousin G. de Beaumont in 1851, "How sad it is that everywhere on earth governments are always precisely as roguish as the morals of their subjects permit them to be! Their vices have found but that one limit" (Tocqueville 1967).[1]

Of course, voting, by whatever method, is not the only electoral system known to history. Election by lot was as often as not used in early Greece and Rome, and for a time by the Roman Catholic church. Every potential candidate was assumed to be capable of assuming the office for which he was a candidate. The Venice law of 1268 for the election of the Doge is a wonderful example of this approach (to clarify the sequence of steps, we have added numbering): "The nobility . . . created a masterpiece of electoral technique. [1] The members of the Grand Council more than 30 years old draw at random 30 electors; already this drawing eliminates all possible intrigue and party corruption. [2] In the same manner [they] coopt a smaller group of 9 members who, [3] by a qualified majority of 7 votes, name a new Quarantia, chosen . . . among the most enlightened and best citizens enjoying the general esteem. [4] The Quarantia

1. "Mais quelle triste chose que sur toute la terre les gouvernements soient toujours précisément aussi coquins que les mœurs de leurs sujets peuvent leur permettre de l'être! Leurs vices n'ont jamais trouvé que cette limite là."

draws by lot a dozen members; [5] this dozen names, in turn, by a qualified majority of 8 votes, 25 citizens worthy of esteem; [6] they draw 9 others by lot; [7] these 9, by a majority of 7 votes, elect 11; [8] finally these 11 draw the real Quarantia of 41 members; [9] the Quarantia elects the Doge. With all these convolutions they hoped to purify the election of all party influence or intrigue, distilling to the utmost the quintessence of patriotism and intelligence" (L. Konopczynski, cited in L. Moulin 1953, 115). The rules are not completely unambiguous, and voting does take place, though exactly how is not made clear. And yet, might this be a better procedure than the popularity contests that national presidential elections have become in many democratic nations today? It certainly would cost less.

This chapter describes elements of an assortment of electoral systems, past and present, to give examples of how politicians have manipulated them, to show that the paradoxes of the theory of social choice are real and can have important practical consequences, and to point to the importance of strategic behavior on the parts of both candidates and voters.

2.1 United States of America

Presidents of the United States are elected indirectly. Each of the fifty states is allocated a number of votes in the Electoral College equal to the number of its Representatives plus its number of Senators; in addition, the District of Columbia (the city of Washington) has three electoral votes. The House of Representatives has 435 members and each state has two Senators, so there is a total of 538 Electoral College votes. A voter's input is one vote for at most one candidate. A candidate who receives a plurality of the votes in a state—meaning the most votes, not necessarily a majority—wins all the Electoral College votes of the state in every state but two and in the District of Columbia. The states of Maine (two Representatives by the apportionment of 2000) and Nebraska (three Representatives) have a different but common rule (adopted by Maine in 1972, by Nebraska in 1992): a candidate who receives a plurality of the votes in a congressional district wins one Electoral College vote, and a candidate who receives a plurality of the votes in the state wins two Electoral College votes.

The system has at least two major drawbacks. First, it is entirely possible for a candidate to be elected whose popular vote is below an opponent's. This has happened at least three times,[2] as shown in table 2.1. Had the number of

2. In 1824 the election was decided in the House of Representatives; some claim that John F. Kennedy's election against Richard Nixon is another instance.

Table 2.1
U.S. Presidents Elected with a Minority of the Popular Votes

Year		Popular Votes	Electoral College Votes
1876	*Rutherford B. Hayes*	4,036,298	185
	Samuel J. Tilden	4,300,590	184
1888	*Benjamin Harrison*	5,439,853	233
	Grover Cleveland	5,540,390	168
2000	*George W. Bush*	50,456,002	271
	Albert Gore	50,999,897	266

Note: Presidential election winners are shown in italic.

In 2000 the Electoral College actually elected 538 electors, but one did not vote, which explains why $271 + 266 \neq 538$. In 1960 John F. Kennedy was the Electoral College winner but, some maintain, not the winner of the popular vote. In 1824 John Quincy Adams was elected by the House of Representatives because no candidate had a majority of the Electoral College votes, but he lagged behind Andrew Jackson in the popular vote.

Representatives been apportioned to the states correctly, then *ceteris paribus* Tilden would have won in 1876 with 185 electoral votes.[3]

The second major drawback is that the system is subject to what we call *Arrow's paradox*: the winner or the final ranking of the candidates can change because of the presence or absence of an "irrelevant" candidate.

In the 2000 U.S. presidential election, Ralph Nader had no chance whatsoever to be elected. His national popular vote total was 2,882,955. Yet his presence as a candidate for Florida's 25 Electoral College votes was enough to change the outcome of the election (see table 2.2) because it is practically certain that his votes would have gone primarily to Gore rather than to Bush. Thus, without Nader's candidacy in Florida, Gore would have obtained 291 Electoral College votes to Bush's 246. This is probably not the only time the Arrow paradox has arisen in U.S. presidential elections. For example, in 1992, Bill Clinton was elected with 43.0% of the popular vote (370 electoral votes) to George H. W. Bush's 37.4% (168 votes) and Ross Perot's 18.9% (0 votes); and in 1912, Woodrow Wilson was elected with 41.8% of the popular vote (435 electoral votes) to Theodore Roosevelt's 27.4% (88 votes) and William Howard Taft's 23.2% (8 votes). The losing pairs of candidates were of the "right" in both cases.

In the United States, candidates to the House of Representatives are elected in single-member congressional districts with the first-past-the-post system: an elector can cast one vote for at most one candidate; the winner is the one who has the most votes. How the districts are drawn is of capital importance. The old art of political gerrymandering—"the practice of dividing a geographical

3. There were, however, many voting irregularities, and the final decision depended on a special commission, a unique event in U.S. history.

Table 2.2
U.S. Presidential Election 2000, Florida Popular Votes

George W. Bush	2,912,790
Albert Gore	2,912,253
Ralph Nader	97,488

area into electoral districts, often of highly irregular shape, to give one political party an unfair advantage by diluting the opposition's voting strength" (*Black's Law Dictionary* 2001)—has in the twenty-first century become a science.

It is generally acknowledged that well upwards of 80% of the seats in the House of Representatives are "safe" in the districts established on the basis of the 2000 census. Many claim the districts determine elections, not votes. If an election is deemed "competitive" when the spread in votes between the winner and the runner-up is 6% or less, then 5.5% of the elections were competitive in 2002, 2.3% in 2004, and 9.0% in 2006. Many candidates ran unopposed by a candidate from one of the two major parties in all three elections (19% in 2002, 15% in 2004, 14% in 2006). In Michigan the Democratic candidates together out-polled the Republican candidates by some 35,000 votes in 2002 yet elected only six representatives to the Republicans' nine. In the 2002 Maryland elections, Republican representatives needed an average of 376,455 votes to be elected, the Democratic representatives only 150,708. In all three elections Massachusetts elected only Democrats, at least half without Republican opposition. Ohio elected eleven Republican and seven Democratic representatives in 2006, and yet the Democratic candidates received 211,347 more votes than did the Republican candidates. Every one of California's fifty-three districts returned a candidate of the same party in the three elections: fifty were elected by a margin of at least 20% in 2002, fifty-one in 2004, and forty-nine in 2006, whereas only one candidate won by less than a margin of 6% in any of these elections. Gerrymandering is widespread and decidedly ecumenical: both parties indulge.

How has this situation come about? That is a long and fascinating story culminating in the U.S. Supreme Court's decision of April 28, 2004, that upheld Pennsylvania's actual districting plan (see Balinski 2004). Everyone involved—the attorneys against, the attorneys for, and the Justices—acknowledged that the plan was a blatant political gerrymander. As the chairman of the National Republican Congressional Committee had predicted, "Democrats rewrote the book when they did Georgia, and we would be stupid not to reciprocate . . . [The Pennsylvania redistricting] will make Georgia look like a picnic." In view of the confused and often contradictory precedents of some forty years, four

Justices, led by Antonin Scalia, ruled: "Eighteen years of essentially pointless litigation have persuaded us that *Bandemer* [1986] is incapable of principled application. We would therefore overrule that case, and decline to adjudicate these political gerrymandering claims" (*Vieth v. Jubelirer* 2004). With no theory, there can be no basis on which to decide whether a plan is fair or not. Only one unambiguous criterion has stood the test of time:

> Since "equal representation for equal numbers of people [is] the fundamental goal for the House of Representatives," the "as nearly as practicable" standard requires that the State make a good-faith effort to achieve precise mathematical equality. Unless population variances among congressional districts are shown to have resulted despite such effort, the State must justify each variance, no matter how small. (*Kirkpatrick v. Preisler* 1969)

Every one of Pennsylvania's nineteen districts has a population, according to the census of 2000, of either 646,371 or 646,372; by the mathematics—the one criterion accepted by the Court—the plan was perfect. Similarly, *every one* of Texas's thirty-two districts has a population of 651,619 or 651,620.

How is it possible to determine such "perfect" plans? The answer is simple. First, a fundamental advance in gerrymandering technology has been made; second, a municipality, township, or village is no longer necessarily within one district. The smallest "atom" that is never split is a census tract: the average number of inhabitants of a census tract in Pennsylvania was thirty-eight. Pennsylvania's district plan splits twenty-nine counties and eighty-one municipalities. Computer programs newly developed for the redistricting season following the census of 2000 make it easy to create maps on a screen and to modify them, by transferring a census tract (or other geographic area) from one district to another, with a simple click of the mouse. With each new map a host of information appears concerning the districts: numbers of inhabitants, numbers of votes for Bush and for Gore in the 2000 elections, numbers of African-, Polish-, or Hispanic-Americans, numbers of Catholics and Protestants, distributions of income levels, and much more. Districts in red are Republican, in blue Democratic. To facilitate "kidnapping"—placing two incumbents of the opposition party in the same district—small elephants indicate the residency of Republican incumbents, small donkeys of Democratic incumbents. Technicians, guided by politicians, simply transfer census tracts from one district to another until they find equality. The programs have brought about a fundamental change: gerrymandering has become a science instead of an art. Justice John Marshall Harlan was unusually prescient when in a 1969 dissenting opinion he called for a new system:

> The fact of the matter is that the rule of absolute equality is perfectly compatible with "gerrymandering" of the worst sort. A computer may grind out district lines which

can totally frustrate the popular will on an overwhelming number of critical issues. The legislature must do more than satisfy one man, one vote; *it must create a structure which will in fact as well as theory be responsive to the sentiments of the community* . . . Even more than in the past, district lines are likely to be drawn to maximize the political advantage of the party temporarily dominant in public affairs. (*Wells v. Rockefeller* 1969; our emphasis)

A new structure has been proposed that eliminates the possibility of political gerrymandering. Called fair majority voting (FMV), it combines single-member constituencies, as required by federal law, with proportional representation at the level of the states. Voters vote as they do in the current system: for a candidate of a party in their congressional district. In every state each party is apportioned a number of seats according to its total vote in the state; it must then be decided which candidates of each party are elected. When there are but two parties, the candidates of each party whose percentages of the votes in their districts are the highest are elected (Balinski 2008).

2.2 Zürich, Switzerland

Biproportional apportionment—similar to fair majority voting but in a more complex situation—provides a counterexample to the general rule: it is a rare instance of theory used in practice. Its story is told here to show that, contrary to widely held opinion, citizens are prepared to accept a seemingly complex method if they can be persuaded it is fair.

Following the 2002 Zürich City Parliament election a citizen filed suit in the Swiss Federal Court, claiming that the electoral law had violated his constitutional right to an equal vote because his vote, though counted, had no effect at all. The reason: a system of proportional representation was used in which party lists were presented in electoral districts; his district had only two representatives; the party for which he voted never received sufficient votes to win any seats in his district; and so his vote was worthless. In December 2002 the Court ruled that indeed the electoral law did violate the constitutionally guaranteed proportional representation principle.

The Zürich department of the interior was charged with finding a remedy. It found one on the Internet at the site of Friedrich Pukelsheim, who developed the necessary software and persuaded the department to adopt the method (Pukelsheim 2006; Balinski and Pukelsheim 2006). The theory and its justification had been developed earlier (Balinski and Demange 1989a; 1989b; Balinski and Rachev 1997). It is now the law for the election of members of the parliaments of the City and the Canton of Zürich.

Table 2.3
Results, Zürich City Parliament Election, February 12, 2006

Party Votes-Seats	SP 23180-**44**	SVP 12633-**24**	FDP 10300-**19**	Greens 7501-**14**
Dist.-Seats				
A-**12**	2377-**4**	1275-**2**	1819-**3**	1033-**2**
B-**16**	2846-**7**	1379-**3**	653-**1**	1082-**3**
C-**13**	2052-**5**	629-**2**	349-**1**	786-**2**
D-**10**	2409-**4**	968-**1**	1092-**2**	842-**1**
E-**17**	3632-**5**	1642-**2**	3015-**5**	1499-**2**
F-**16**	2628-**6**	1972-**4**	754-**2**	572-**1**
G-**12**	2938-**4**	1630-**3**	1272-**2**	807-**1**
H-**19**	2976-**6**	2113-**4**	1039-**2**	661-**1**
J-**10**	1322-**3**	1025-**3**	307-**1**	219-**1**
Party-div.	*1.01*	*1*	*1.01*	*1*

The method was inaugurated on February 12, 2006. The results are given in table 2.3. Voters vote as they do in the current system: they cast a vote for at most one of the party lists of candidates in their district. Each district has a predetermined number of seats that have been apportioned as a function of their populations; in table 2.3 districts are identified by A, B, . . . , J, followed by their numbers of seats (e.g., B-**16** means district B has been apportioned sixteen seats).

The parties are apportioned numbers of seats that depend on their total vote in all districts by Webster's or Saint-Laguë's method.[4] In table 2.3 the parties are identified by SP, SVP, . . . , SD, followed by their total votes and numbers of seats (e.g., 5418-**10** under CVP means that party received 5,418 votes in total and was apportioned ten seats). The entry of the row of a district and the column of a party gives the number of votes received by the party list in that district, followed by the number of elected candidates (e.g., 3015-**5** means that in district E the list of the FDP party had 3,015 votes and elected five candidates). The number of candidates elected by each party list is the result of a computation that requires a computer program. However, the solution is easily checked given the *district divisors* and *party divisors* (in table 2.3, the last column and last row, respectively): the vote in an entry is divided by its divisors and rounded to the closest integer (e.g., $3015/(660 \times 1.01) \approx 4.52$ is

4. Webster's or Saint-Laguë's method: If the party votes are $(v_1, \ldots, v_n)$, and the number of seats to distribute is h, then the apportionment is $(a_1, \ldots, a_n)$, where $a_i = [v_i/\lambda]$ and the divisor λ is chosen so that $\sum a_i = h$. $[x]$ is x rounded to the nearest integer for x any real number. When seats are apportioned to regions and the p_i are populations, a minimum of 1 is usually guaranteed each region: in that case the computation is the same except that $a_i = \max\{1, [v_i/\lambda]\}$.

Table 2.3
(cont.)

Party Votes-Seats	CVP 5418-**10**	EVP 3088-**6**	AL 2517-**5**	SD 1692-**3**	Dist. -div.
Dist.-Seats					
A	610-**1**	236-**0**	201-**0**	138-**0**	*600*
B	541-**1**	176-**0**	464-**1**	198-**0**	*432*
C	315-**1**	79-**0**	699-**2**	108-**0**	*400*
D	440-**1**	342-**1**	230-**0**	111-**0**	*660*
E	837-**1**	618-**1**	323-**1**	144-**0**	*660*
F	708-**1**	615-**1**	154-**0**	333-**1**	*473*
G	696-**1**	391-**1**	212-**0**	124-**0**	*650*
H	777-**2**	631-**2**	191-**1**	328-**1**	*470*
J	494-**1**	0-**0**	43-**0**	208-**1**	*400*
Party-div.	*1*	*0.88*	*0.8*	*1*	

rounded to 5). The solution is unique ("up to ties," which are extremely rare). FMV is the same method except that every district is apportioned exactly one seat, which simplifies the theory and the computation.

It should be noted that the number of elected candidates per list may deceive the candidates of some lists. The Greens' district B list elects three candidates with 1,082 votes, whereas the SVP's district E list elects only two with 1,642 votes, and there are other similar instances. This is the price of guaranteeing equity to parties and districts. The solution was accepted with no complaints; indeed, the canton of Aarau has decided to adopt this method, too.

2.3 Mexico

Politicians do not only manipulate electoral systems seeking advantages; they sometimes establish systems that are logically inconsistent. Mexico's system for electing the 500 members of its House of Deputies changed in 1989, 1994, and 1996 (see Balinski and Ramirez 1999). By the 1989 law—the most baroque of all—two political parties could be guaranteed an absolute majority. By the law of 1994 it was possible for a party to gain or lose as many as twenty seats for an error in the count of the votes as small as 0.01% (which represents 0.05 of 500 seats).

The 1996 law was adopted a scant eight months before it was applied in the election of 1997: (1) 300 deputies were to be elected in single-member constituencies with the first-past-the-post system; and (2) 200 deputies were to be elected with a system of proportional representation, the parties presenting lists in each of five 40-member electoral regions for which voters cast a

second, separate vote. These 200 members were to be allocated to the parties in proportion to their total national second votes by the method of Hamilton,[5] except that no party could have more than 300 seats in total, and if a party's percentage of the total vote was $x\%$, then it could have no more than $(x + 8)\%$ of the 500 seats.

There were, on July 6, 1997, five eligible parties, which elected the following numbers of members in the 300 single-member constituencies:

PRI: 165, PAN: 64, PRD: 70, PVEM: 0, PT: 1.

The results of the second votes—for the respective regional party lists—are given in table 2.4a. By the law, the PRI could receive no more than $39.96 + 8\% = 47.96\%$ of $500 = 239.8$, so 239 seats. Having already elected 165, it had to be allocated 74 of the 200 seats, leaving 126 to be distributed among the remaining parties by a first procedure—Hamilton's method—implying the following apportionment of the 200 seats:

PRI: 74, PAN: 57, PRD: 55, PVEM: 8, PT: 6.

The law further stipulated that the seats were to be apportioned among the regional party lists by a second procedure. First, the seats of a party having been limited by one of the constraints are apportioned to its five regional lists by the method of Hamilton (so the PRI's regional lists are apportioned 15, 17, 15, 14, and 13). This leaves $40 - 15 = 25$ in region I, $40 - 17 = 23$ in region II, and so on. Second, Hamilton's method is used to apportion the remaining seats in each region to the other parties. So, for example, the 25 seats remaining in region I are apportioned among the four other parties. The results of this second procedure are given in table 2.4b.

The second procedure allocates to the PRI and the PT the number of seats they should receive according to the first part of the law. However, no other party is apportioned the number of members it should receive: the PAN receives one too few, the PRD two too many, the PVM one too few. The law is logically inconsistent. This is hardly surprising because the second procedure ignores the allocations of the first procedure except for that of the PRI. It should come as no surprise that PRI politicians were the authors of this law. Unknown by the general public, the Consejo General del Instituto Federal Electoral invented an ad hoc rule to correct the error (see the italic numbers in table 2.4b): (1) allocate the seats of the PRI, as before; (2) for *each region* compute the quotas of the

5. Hamilton's method: Suppose h seats are to be apportioned to n parties with votes $(v_1, \ldots, v_n)$. First the integer part of the quota $q_i = hv_i / \sum v_j$ is apportioned to each party i; then any unapportioned seats are assigned to those parties having the largest fractions or remainders, $q_i - \lfloor q_i \rfloor$ (where $\lfloor x \rfloor$ is the real number x rounded down to the nearest integer).

Table 2.4a
Mexican House of Deputies Election, July 6, 1997

Region	PRI	PAN	PRD	PVEM	PT	Total
I	2,379,785	2,504,484	1,019,822	197,098	118,673	6,219,862
II	2,543,570	2,138,564	687,162	112,721	303,794	5,785,811
III	2,354,047	909,386	1,377,933	90,373	126,342	4,858,081
IV	2,098,581	1,237,297	2,385,525	424,672	123,612	6,269,687
V	2,062,736	1,005,807	2,048,461	291,273	83,704	5,491,981
Total	11,438,719 39.96%	7,795,538 27.23%	7,518,903 26.27%	1,116,137 3.90%	756,125 2.64%	28,625,422 100%

Table 2.4b
Mexican House of Deputies Election, July 6, 1997, showing the Law's Inconsistent distribution of 200 Seats and the "Corrections"

Region	PRI	PAN	PRD	PVEM	PT	Total
I	15	16(*17*)	7(*6*)	1	1	40
II	17	15	5	1	2	40
III	15	9	14(*13*)	1	1(*2*)	40
IV	14	8	15	2(*3*)	1(*0*)	40
V	13	8	16	2	1	40
Total	74	56(*57*)	57(*55*)	7(*8*)	6	200

Note: "Corrections" are shown in italics.

remaining seats due each party list and allocate their integer parts; (3) for each party, in the order of the largest to smallest vote total, assign any unapportioned seats to the lists with the largest remainders subject to the further restriction that no region may be assigned more than forty seats. This procedure is completely arbitrary and has no justification whatsoever, e.g., why not start with the party with the smallest vote total?

There exists a method to solve this problem that is justified: the biproportional method used in Zürich. It yields the solution given in table 2.4c.

The Mexican story is not an isolated example of a logically inconsistent law. Silvio Berlusconi's government, facing upcoming elections in April 2006, completely changed the electoral law for electing the members of the Chamber of Deputies on December 14, 2005, a mere four months before the elections. It replaced a mixed system, known as the "Legge Mattarella" or "Mattarellum," whereby 75% of the deputies were elected in single-member districts and 25% on a proportional basis, with a system that is supposedly proportional. Its difficulties are similar to Mexico's: seats are allocated to parties on the basis of the total national vote, but the procedures for apportioning seats to regional party

Table 2.4c
Mexican House of Deputies Election, July 6, 1997, showing the Biproportional Apportionment

Region	PRI	PAN	PRD	PVEM	PT	Total
I	**14**	17	**7**	1	1	40
II	**16**	**16**	5	1	2	40
III	**18**	**8**	**12**	1	**1**	40
IV	**12**	8	**16**	3	**1**	40
V	**14**	8	**15**	2	1	40
Total	74	57	55	8	6	200

Note: Boldface numbers indicate differences from the actual apportionment.

lists do not take proper account of the national apportionments to parties or regions, so fall into similar logical traps (Pennisi 2006). An inconsistency was realized in the 2006 election: the region of Molise was entitled to three seats but ended up with only two, whereas the Trentino Alto Adige was entitled to ten but was awarded eleven (Pennisi et al. 2007).

Another example concerns France. The government wished to change the method for electing members of the European Parliament in 2003 and proposed a system that was meant to guarantee representation proportional to party votes and to department populations. Regrettably, the proposed method of computation was patently absurd, so—ignorant of the biproportional method (which was invented and developed in France)—the government abandoned the idea altogether (Balinski 2004, ch. 7).

2.4 United Kingdom

The plurality or first-past-the-post system is used in the United Kingdom to elect the members of its House of Commons, though a national commission defines the single-member constituencies. The general elections of 2005 were trumpeted to be another decisive victory for Tony Blair's Labour Party. The fact is that it won a very comfortable majority of the seats—356 of 646 seats (55.1%)—with barely more than *one-third* of the votes (35.2%). The Conservative Party, just shy of one-third of the votes (32.3%) won only 198 seats (30.7%), and the Liberal Party, with more than one-fifth of the votes (22.0%), elected less than one-tenth of the members of Parliament (9.6%). This far from representative result is due to single-member constituencies and the existence of three important political parties (there are many small parties as well), so that a great many of the races are won by candidates who have 40% or less of the votes in their districts. A party with 10% of the votes spread fairly evenly

Table 2.5
The Winners of the Last Six British Elections

	1983	1987	1992	1997	2001	2005
Votes	42.4%	42.2%	41.9%	43.2%	40.7%	35.2%
Seats	61.1%	57.8%	51.6%	63.4%	62.5%	55.1%

throughout the country could easily end up with no seats at all. A big advantage accrues to the party that has the highest percentage of votes, even when its margin over the second highest party's percentage is very small.

That a method of voting has the property of producing a "majority" party able to govern is generally viewed as "a good thing," but is this a truly democratic outcome? The six most recent parliamentary elections in the United Kingdom have also seen minorities of the votes translated into (for the most part) large majorities of the seats, as may be seen in table 2.5. The Conservatives benefited from 1983 to 1992, Labour from 1997 to 2005.

2.5 Australia

Australia has a parliamentary form of government similar to that of the United Kingdom's. The Prime Minister is chosen by the House of Representatives from among its membership. Since 1918 it has used what is variously known as the *alternative vote*, *preferential voting*, or *instant-runoff voting* (IRV) to elect each of the 150 members of its House of Representatives in single-member constituencies. Ballots contain a list of the names of the candidates and their party affiliations, each preceded by a box: a voter *must* place the numbers 1, 2, . . . , up to the total number of candidates, indicating their "preferences"; otherwise the ballot is invalid.[6] Citizens are obliged to vote. It is standard practice for parties to hand out cards showing how they wish voters to place their preferences (orders that must be the result of intense negotiations among the parties). Nevertheless, some 4% to 5% of the ballots are invalid (called "informal"). Australia is one of the rare countries—along with Ireland and Malta—where the voters' inputs are rank-orders of candidates in conformity with the basic paradigm of the theory of social choice.

6. Originally candidates were listed on ballots in alphabetical order. As of 1984 the order of the list is determined by a random device. This was introduced to combat "donkey votes": a "1" given to the first on the list, a "2" to the second, and so on down. This encouraged parties to nominate candidates whose names began with an "A". The most egregious example of this was the Senate election in New South Wales of 1937 when the Labor party nominated candidates named Armour, Armstrong, Arthur, and Ashley, all of whom were elected.

Table 2.6
The Winning Coalitions of the Last Six Australian House of Representatives Elections

	1993	1996	1998	2001	2004	2007
First votes	44.9%	47.3%	39.2%	43.0%	46.4%	43.4%
Seats	54.4%	63.5%	54.1%	54.7%	57.3%	55.3%

With the alternative vote, a candidate listed first by a majority is elected; if there is no such candidate, then the candidate *C* who received the fewest first votes is eliminated from all lists, so that the candidates who were second after *C* become first. A candidate listed first by a majority on the emended ballots is elected; otherwise the candidate listed first least often on the emended ballots is eliminated; and the procedure is repeated until a candidate with a majority is found. In the 2007 election there were constituencies having four, five and up to thirteen candidates. The incumbent Prime Minister and Liberal Democratic candidate John Howard, running against twelve other candidates, had 45.5% of the first votes against 45.3% for his Labor opponent; the procedure had to eliminate all eleven other candidates to determine that John Howard had lost with 48.6% of the vote.

The information available about past elections gives the numbers of first votes of the candidates; then the numbers of first votes after one candidate was eliminated; then the numbers of first votes after two candidates were eliminated; and so on, until a majority candidate emerged. For the John Howard defeat, twelve lists of numbers are given. This information is not sufficient to show in practice what undoubtedly occurs and will be pointed out in theory (see chapter 3): Arrow's paradox and violations of monotonicity (the latter occurs when the ballots are changed only in that the winning candidate is listed higher in some rankings, yet this causes her to lose).[7]

The system, based on single-member constituencies, favors the large coalition of parties (see table 2.6) *if* first votes are used as a measure of proportionality. This is due, as in the U.K., to single-member constituencies and a system that ignores how much support a party may enjoy nationally.

The complexity of the count is such that a preliminary and approximate result is given on election night, calculated on the basis of comparing the two

7. The mayor of London, England, is elected with the same system except that a voter is asked only to name his or her first and second choices, whatever the number of candidates. The method violates monotonicity and admits the Arrow paradox (since listing two when there are three candidates is the same as listing all three).

candidates considered most likely to win. Definitive results are not known until the following day.

The Australian Senate has seventy-six members, twelve from each of the six states, two from each of the two territories. It is renewed by halves. Typically, six Senators are elected simultaneously in a statewide vote using the *single transferable vote* (STV) system. It is a multicandidate version of the alternative vote. Again, the voters' rank-orderings of the candidates indicating their "preferences" is the input. But in these elections there can be huge numbers of candidates: in 2004, for example, there were seventy-eight candidates for the six positions in the New South Wales election. A voter can either choose to vote "above the line" or "below the line." "Above" means the voter chooses the rank-ordering (of all the candidates) specified by one political party, the outcome of prior strategic negotiations. "Below" means the voter determines her own rank-ordering. In practice, over 95% vote above the line.

The STV works as follows. First, the ("Droop") *quota* is computed: if b is the number of ballots and s the number of candidates to be elected, $q = \lfloor 1 + \frac{b}{s+1} \rfloor$ (q is the smallest integer that when multiplied by $s+1$ yields a greater number than b, so any smaller q could elect $s+1$ members). The computation is iterative. Any candidate listed first at least q times is elected and dropped from all lists. If this elects fewer than the necessary number, then an elected candidate C's surplus of votes above q is distributed *pro rata* to the candidates that immediately follow C's (currently) first-ranked position. The procedure is repeated until the necessary number of candidates is elected, unless at some step no new candidate is elected. In that case the candidate D with the fewest current first places is eliminated, and her votes are distributed *pro rata* to the candidates that immediately follow D's (currently) first-ranked position. The computation continues in the same manner with the distribution of surplus votes of an elected candidate taking precedence over the distribution of the votes of a candidate with fewest current first places. In the famous New South Wales election—seventy-eight candidates was a record—it took seventy-seven steps to determine all winners. As a multicandidate version of the alternative vote, STV suffers from the same defects. It is not monotonic and it does not avoid Arrow's paradox. It is, in addition, a very complex procedure that takes days to compute and that voters are unable to verify because the complete lists of preferences of the voters are not published. Moreover, as with any method whose inputs are rank-orderings, it seriously constrains the voters' abilities to express intensities of preference: perhaps only one or two or three candidates are acceptable at all, perhaps one candidate is considered best, the second trails well behind, and so on.

2.6 France

French electoral laws have undergone more than two centuries of change. The main manipulative variable of the early years was who was allowed to vote. On the eve of the Revolution, in the summer of 1789, some 5 million (including some women) participated in electing the members of what came to be the Assemblée nationale. Successive regimes imposed conditions of ownership and earnings on the franchise to vote and excluded women. Later in 1789, 4.3 million participated; in 1793 the number went up to 7 million; with the Empire in 1799 the number decreased to 5 million (other controls concerned who was allowed to be elected). But then, with the Restoration in 1814, the number dropped drastically to some 90,000, climbing slowly from 94,000 in 1830 to 241,000 on the eve of the revolution of 1848, when "true" universal suffrage is generally considered to having been achieved. Women did not obtain the franchise until 1944.[8]

A particularly striking example of blatant manipulation was the electoral law of May 1951. The coalition of center parties that ruled France were afraid of two major forces: Charles de Gaulle's R.P.F. (Rassemblement du Peuple Français) and the P.C. (Parti Communiste). The law stipulated that in each department parties should present lists of candidates, the votes then going to party lists. Parties could, however, declare themselves as "grouped." A party that stands alone is treated as a group. (1) If a group of parties obtains a majority of the votes, the group obtains all the seats of the department; otherwise, the seats are first apportioned to the groups. (2) Then the seats of each group are apportioned to its parties. The method of Jefferson (or of D'Hondt) is used in each case.[9] The inventors were certain that the R.P.F. and the P.C. would never form a group. The results of the election in Marne, given in table 2.7, show how well the method worked. Had the method of Jefferson been applied to the parties directly, the P.C. and the R.P.F. would each have had two seats and the parties of the group one in total (last column). The method limited the P.C. and the R.P.F. to one seat each and gave three to the group of parties.

The suburban towns surrounding Paris constituted the P.C.'s fortress, and the R.P.F. was thought to be popular in Paris and several of its more affluent suburbs. The architects of the law feared the electoral system just described would be a disadvantage in these departments. They found a simple expedient:

8. On April 21, 1944, a decree of Charles de Gaulle's provisional government in Algiers established women's suffrage. Earlier in the twentieth century, the left had been, by and large, opposed to granting the vote to women: its members claimed that priests would dictate their votes.

9. The method of Jefferson (or D'Hondt): If the votes are $(v_1, \ldots, v_n)$, and the number of seats to distribute is h, then the apportionment is $(a_1, \ldots, a_n)$, where $a_i = \lfloor v_i/\lambda \rfloor$ and the divisor λ is chosen so that $\sum a_i = h$.

Table 2.7
Elections to the French Assemblée nationale, Marne, Five Seats, June 17, 1951

	Vote	Group Vote	Group Seats	Party Seats	Direct
PC	47,216	47,216	1	1	2
RPF	45,912	45,912	1	1	2
MRP	36,702			2	1
SFIO	18,567			1	0
	+	=75,735	3		
RGR	16,575			0	0
GrC	3,891			0	0
RIF	4,024	4,024	0	0	0

Note: Seven parties presented lists. The MRP, SFIO, RGR, and GrC declared themselves a group. "Party Seats" was the actual apportionment. "Direct" is the apportionment to party lists. No group has a majority. So by the law: $\lambda = 25,000$ apportions the five seats to the groups, and $\lambda = 17,000$ apportions the group's three seats to their parties. $\lambda = 22,000$ apportions the five seats directly to the parties.

by "exception," it was not possible to form a group in the departments of the greater Paris region; Jefferson's rule was applied directly to the votes of party lists (as in the last column of table 2.7). The manipulation was effective. Despite the exception, the P.C.'s 25.9% of the national vote (the highest percentage of any party) returned but 17.8% of the seats. The R.P.F. fared better, having succeeded in forming groups with parties of the right in several departments.

Each Republic has had its own constitution, its own changing electoral systems. Today's regime is the Fifth Republic, established by Charles de Gaulle in 1958. From 1980 to 2007 one or another part of the electoral system has been changed eight times. It is difficult to pretend that any of these modifications were motivated by the wish for a more equitable, more representative democratic system.

Deputies of the Assemblée nationale have been elected in single-member districts with the two-past-the-post system in every election since 1958 except one. (1) An elector can cast one vote for at most one candidate; the winner is the one who wins a majority if it amounts to at least one-quarter of the registered voters. (2) If there is no such candidate, a second round is held among those candidates having at least 12.5% of the votes in the first round.[10] Parties negotiate withdrawals of candidates, so most second rounds are between the two major candidates at opposing ends of the political spectrum. Instead of favoring the center, as did the Fourth Republic's electoral law of 1951, the two rounds of the Fifth Republic favor the major parties of the right and the left.

10. Initially, 5% sufficed, then 10%, finally 12.5%.

The apportionment of seats to departments and the districts defined in 1958 (on the basis of the census of 1954) remained the same for the seven elections spanning 1958 through 1981. The temptation to gerrymander was not resisted. The districts distinctly favored the right. The distortion of representation became much worse across the years, accentuated by the migration from rural to urban areas.

In fear of the outcome of the election of 1986, the socialist government of François Mitterrand changed the electoral law in 1985, reapportioning 577 seats (on the basis of the 1982 census) instead of 474, thereby assuaging many who feared defeat, and installed a system of proportional representation by party lists in departments using Jefferson's method. The evidence suggests this manipulation was a brilliant success: although the parties of the right carried the election, they won by only the narrowest of margins. The main reason for this is that with proportional representation the far right Front National party (F.N.) won thirty-five seats (whereas it usually had either none or one); arguably the maneuver is responsible for having made the F.N. a major contender.

One of the new government's first acts was to revert to the previous system, though maintaining the apportionment of the 577 seats. The Pasqua electoral law of 1986 is unique in having announced three main principles: (1) that there be a reapportionment and redistricting after every second census; (2) that each canton (some forty per department) must belong entirely to one contiguous district (with a very few exceptions of populated cantons and large cities); and (3) that within a department no district's population can differ from the department's average district population by more than 20%. This means the relative difference between any two district populations x and y, $|x - y|/\min\{x, y\}$—which effectively measures the greatest inequality between inhabitants of a department—cannot exceed 50%.

Despite the principles, gerrymandering was the name of the game: the redistricting was decidedly in favor of the right. It was immediately challenged by the socialists in the Conseil constitutionnel (the constitutional court). Superficially, the districting plan gave an impression of being incoherent, even absurd, they wrote, having used a heterogeneous set of criteria:

> To heed old districts or forget them, to impose at all costs a demographic logic or ignore it, to cling to established geographic entities or be unaware of them, to keep cities united or to split them, to separate or mix city and country, mountain and plain, the right and left banks of rivers, affluent and poor neighborhoods, are many alternative choices each with its own logic, all that could be defended. What would not be logical, however, and could not be defended, would be to vary the answers from one place to another.

After giving a rich collection of specific illustrations, their argument concluded,

> All electoral districting plans have three components: demographic, geographic and political. The first two are known and open. The third, hidden, is no less important . . . One cannot be surprised that the Government did not resist the temptation of including [the third]. But in fact it is this precisely that characterizes the arbitrary nature of the plan, and sullies it. From the point of view of demography, of geography, of history, of economy, of social conditions, the criteria were applied in a very heterogeneous manner. But when it comes to the political criterion, it was used in a perfectly homogeneous fashion. (Conseil constitutionnel 1986)

The Conseil let matters stand, ruling that the Constitution did not confer on it the power to decide on the equity of the plan nor the power to suggest other plans. In essence, it followed Justice Felix Frankfurter's famous 1946 dictum, "Courts ought not to enter this political thicket" (*Colegrove v. Green* 1946), upheld in a 2004 U.S. Supreme Court decision written by Justice Antonin Scalia: "[W]e believe the correct standard which identifies unconstitutional political districting has not been met . . . [We] do not know what the correct standard is . . . We would therefore . . . decline to adjudicate these political gerrymandering claims" (*Vieth v. Jubelirer* 2004). The fact is that there is no theory, no well-defined set of criteria, by which to decide which of two districting plans is the more equitable.

France has conducted two censuses since that of 1982, in 1990 and 1999 (but no census will again be taken because a "continuous" estimation procedure was established in 2004). Governments of the left and of the right have ignored the principle that reapportionment and redistricting should have taken place. The Pasqua redistricting has remained in force through the legislative elections of 2007. The distortions have become grotesque. France has one hundred departments that share 570 seats, and 7 seats are allocated to territories. According to the latest available official figures (January 2006)—and compared with the ideal standard of Webster's apportionment—only forty-six departments have the number of seats they deserve, thirty-one have 1 too many, sixteen have 1 too few, six have 2 too few, and one has 3 too few. Table 2.8 gives examples of the inequity of the actual apportionment and shows that there are cases of pairs of departments where the more populated one has fewer seats (in all, there are eighty-two such pairs).

Populations of departments are easy to obtain; those of the legislative districts more up-to-date than the 1999 census were not available (in 2008). According to the 1999 census—and the distortions were undoubtedly worse in 2007—the

Table 2.8
Extract of Apportionment of Seats in the French Assemblée nationale to Departments, 2006 Populations

Department	Population	Equitable	Actual
Seine-et-Marne	1,267,000	11	9
Seine-Maritime	1,245,000	11	12
Haute-Garonne	1,169,500	11	8
Moselle	1,039,000	9	10
Var	974,000	9	7
Ain	565,000	5	4
Saone-et-Loire	546,000	5	6
23 most populated	31,403,000	284	266
77 least populated	31,596,000	286	304
Total	62,999,000	570	570

Note: "Equitable" means Webster's apportionment.

population of the 2d district of Lozère was 34,374 and that of the 2d district of Val d'Oise was 188,200, so that two residents of the first weighed as heavily as eleven residents of the second. Some of this inequality is due to bad apportionment, some to bad definitions of districts. Within the department of Var, independent of the apportionment, two residents of the 1st district weighed about as much as five residents of the 6th district, for the 1st district had 73,946 inhabitants and the 6th district had 180,153. Though these examples are the worst, there is no denying that there are flagrant inequities throughout the country.

No changes at all were made in the representation of France's departments in the Assemblée nationale from 1986 through 2009; this is manipulation all the same—passive manipulation—because the equal, constitutionally guaranteed rights of voters were ignored for the convenience of the deputies whose districts remain unchanged. A new electoral law is, at last, expected in 2010.

French Presidential Elections

Article 7 of the French constitution was amended in 1964 to establish direct popular election of presidents:

> The president of the Republic is elected by an absolute majority of the votes. If it is not obtained in the first round of the election, a second round is held fourteen days later. The only two candidates who may present themselves, after the eventual withdrawal of more favored candidates, are those who have the largest number of votes in the first round.

In each round[11] a voter may cast one vote for at most one candidate, and the order of finish is determined by the candidates' votes. Except for the provision of a run-off between the top two finishers, this is exactly the mechanism used in the U.S. presidential elections: an elector has no way of expressing her or his opinions concerning candidates except to designate exactly one favorite. In consequence—imagine for the moment a field of at least three candidates—his or her vote counts for nothing in designating the winner unless it was cast for the winner, for no expression concerning the remaining two or more candidates is possible.

2002 Presidential Election

The French presidential election of 2002 with its sixteen candidates is a veritable storybook example of the inanity of the two-past-the-post mechanism. Jacques Chirac, the incumbent president, was the candidate of the Rassemblement pour la République (R.P.R.), the big party of the "legitimate" right; Lionel Jospin, the incumbent prime minister, that of the Parti Socialist (P.S.); Jean-Marie Le Pen, that of the extreme right Front National party (F.N.); and François Bayrou, that of the moderate Union pour la Démocratie Française (U.D.F., the ex-president Valéry Giscard d'Estaing's party). Arlette Laguiller was the perennial candidate of a party of the extreme left, the Lutte Ouvrière. The extreme right had two candidates, Le Pen and Bruno Mégret; the moderate right five, Chirac, Bayrou, Alain Madelin, Christine Boutin, and Corinne Lepage; the left and the Greens had four, Jospin, Jean-Pierre Chevènement, Christiane Taubira, and Noël Mamère; and the extreme left had four, Laguiller, Olivier Besancenot, Robert Hue, and Daniel Gluckstein. One group, the "hunters" (Hunting, Fishing, Nature, Tradition party), managed to present only one candidate, Jean Saint-Josse.

The strategic aspects surrounding so many candidates turned the election into something of a farce (see table 2.9). The public expected a confrontation between the dominant candidate of the right, Jacques Chirac, and the dominant candidate of the left, Lionel Jospin. Instead it was offered a choice between Chirac and Jean-Marie Le Pen of the extreme right. Chirac crushed Le Pen, obtaining 82.2% of the votes in the second round. Some 20% of Chirac's votes were obviously *for him*. Most of his votes are more accurately described as

11. There have always been two rounds. The first direct popular election of the president in the Fifth Republic (instituted in 1958) was in 1965: in the first round Charles de Gaulle had 44.64% of the vote, François Mitterrand 31.72%. Together they received 76.36%. In every subsequent election the top two together received a lower percentage. In 2002 the top seven together received 76.04%.

Table 2.9
French Presidential Election, First-Round Votes, April 21, 2002

J. Chirac	19.88%	J. Saint-Josse	4.23%
J.-M. Le Pen	16.86%	A. Madelin	3.91%
L. Jospin	16.18%	R. Hue	3.37%
F. Bayrou	6.84%	B. Mégret	2.34%
A. Laguiller	5.72%	C. Taubira	2.32%
J.-P. Chevènement	5.33%	C. Lepage	1.88%
N. Mamère	5.25%	C. Boutin	1.19%
O. Besancenot	4.25%	D. Gluckstein	0.47%

against Le Pen: the intrinsic value of a vote when there is only one to cast has very different meanings.

Had either Jean-Pierre Chevènement, an ex-socialist, or Christiane Taubira, a socialist, withdrawn, most of his 5.3% or her 2.3% of the votes would have gone to Jospin, and the second round would have pitted Chirac against Jospin. According to most of the polls, Jospin would have beaten Chirac, though by little.[12] In fact, Taubira had offered to withdraw if the P.S. was prepared to cover her expenses, but that offer was refused. It was rumored that the R.P.R. helped to finance Taubira's campaign (a credible strategic gambit backed by no specific evidence). But if Charles Pasqua, an aging ally of Chirac, had been a candidate, as he had announced he would be, then he might have taken away enough votes from Chirac to result in a second round between Jospin and Le Pen. In this event Jospin would surely have been the overwhelming victor, and for the same reason that Chirac emerged the victor: most of his votes would have been against Le Pen.

What does this story show? The first- and two-past-the-post mechanisms invite strategic candidacies, candidates who cannot hope to win but can change the outcome. This is why Arrow's paradox is of great practical significance: the very possibility of its occurrence completely confuses the ultimate outcome. It also shows the very different meanings or values of votes when these mechanisms are used.

2007 Presidential Election

French voting behavior in the presidential election of 2007 was very much influenced by the experience of 2002. There were twelve candidates. Nicolas

12. Sofres predicted a 50%–50% tie on the eve of the first round. In the last eleven predictions, spanning two months before the first round, Sofres polls showed Jospin the winner seven times, Chirac the winner twice, and two ties.

Sarkozy was the candidate of the U.M.P. (Union pour un Mouvement Populaire, founded in 2002 by Chirac), its president and the incumbent minister of the interior; Ségolène Royal, that of the P.S.; Bayrou, again that of the U.D.F. (though he announced immediately after the first round that he would create a new party, the MoDem or Mouvement Démocrate); Le Pen, again that of the F.N.; and Dominique Voynet, the candidate of the Greens. The extreme left had five candidates, Besancenot (again), Marie-George Buffet, Laguiller (again), José Bové, and Gérard Schivardi; the extreme right had two, Le Pen (of course) and Philippe de Villiers; and the "hunters" had one, Frédéric Nihous. The distribution of the votes among the twelve candidates in the first round is given in table 2.10. In the second round Nicolas Sarkozy defeated Ségolène Royal by 18,983,138 votes (53.06%) to 16,790,440 (46.94%).

In response to the debacle of 2002, the number of registered voters increased sharply (from 41.2 million in 2002 to 44.5 million in 2007), and voter participation was mammoth: 84% of registered voters participated in both rounds. Voting is a strategic act. In 2007 voters were acutely aware of the importance of who would survive the first round. Many who believed that voting for their preferred candidate could again lead to a catastrophic second round, voted differently. Such behavior—a deliberate strategic vote for a candidate who is not the elector's favorite ("le vote utile")—was much debated by the candidates and the media, and was practiced.

A poll conducted on election day (by TNS Sofres–Unilog, Groupe Logica-CMG, April 22, 2007) asked electors what most determined their votes. One of the seven possible answers was a deliberate strategic vote: this answer was given by 22% of those (who said they voted) for Bayrou, 10% of those for Le Pen, 31% of those for Royal, and 25% of those for Sarkozy. Comparing the first rounds in 2002 and 2007 also suggests that deliberate strategic votes were important in 2007. In 2002 the seven minor candidates of the left and the Greens (Laguiller, Chevènement, Mamère, Besancenot, Hue, Taubira, Gluckstein) had 26.71% of the vote, whereas in 2007 six obtained only 10.57% (Besancenot, Buffet, Voynet, Laguiller, Bové, Schivardi); in 2002 the five minor candidates

Table 2.10
French Presidential Election, First-Round Votes, April 22, 2007

N. Sarkozy	31.18%	M.-G. Buffet	1.93%
S. Royal	25.87%	D. Voynet	1.57%
F. Bayrou	18.57%	A. Laguiller	1.33%
J.-M. Le Pen	10.44%	J. Bové	1.32%
O. Besancenot	4.08%	F. Nihous	1.15%
P. de Villiers	2.23%	G. Schivardi	0.34%

of the right and the "hunters" (Saint-Josse, Madelin, Mégret, Lepage, Boutin) had 13.55% of the vote, whereas in 2007 two obtained only 3.38% (Villiers, Nihous).

A candidacy can be a strategic act as well, as was shown in the 2002 election. To become an official candidate requires 500 signatures. They are drawn from a pool of about 47,000 elected local and national officials who represent the one hundred departments and must include signatures coming from at least thirty departments but no more than 10% from any one department. Both Besancenot and Le Pen appeared to have had difficulty in obtaining them. Sarkozy publicly announced he would help them obtain the necessary signatures, as a service to democracy.

In the period leading up to the first round of voting, the major candidates of the right and the left—Sarkozy of the U.M.P. and Royal of the P.S.—both argued strenuously against Bayrou, the centrist. Both most feared him in a one-to-one confrontation. The polls show why: as of February 2007 they consistently suggested that Bayrou would defeat either one of them in the second round. Immediately after the first round, Royal and Sarkozy both sought Bayrou's support and tried to incorporate some of his ideas along with theirs. Royal subsequently revealed that she had offered Bayrou the position of prime minister at that time. Once elected, Sarkozy, in naming many political personalities of the left to responsible political positions (ministries, commissions, a coveted international position), put into effect one of Bayrou's principal promises, the appointment of persons from the left *and* the right ("l'ouverture").

Polling results (table 2.11) suggest that François Bayrou was the Condorcet-winner: he would have defeated *any* candidate in a head-to-head confrontation. Moreover, the pair-by-pair confrontations (of March 28 and April 19) determine an unambiguous order of finish (there is no Condorcet-cycle): Bayrou is first, Sarkozy second, Royal third, and Le Pen last. On the other hand, there is a linear interpolation between their percentages of the vote in head-to-head encounters among the principal three on December 15, 2006, and April 19, 2007, where there is a Condorcet-cycle (just under halfway from the first to the second date): Sarkozy defeats Bayrou, Bayrou defeats Royal, and Royal defeats Sarkozy. This suggests that at some time in that period a poll may well have revealed the Condorcet paradox. A more precise actual occurrence of the paradox comes from the 1994 general election of the Danish Folketing. A pre-election poll elicited the voters' preferences among the three main contenders for prime minister—H. Engell, U. Ellemann-Jensen, and P. Nyrup Rasmussen—and found Engell preferred to Ellemann-Jensen (by 50.6%), Ellemann-Jensen preferred to Rasmussen (by 51.1%), and Rasmussen preferred to Engell (by 52.8%) (Kurrild-Klitgaard 1999).

Table 2.11
Polls on Potential Head-to-Head Second-Round Results, French Presidential Election, December 2006–April 2007

	Dec. 15	Jan. 20	Feb. 15	Mar. 15	Mar. 28	Apr. 16	Apr. 19
Bayrou	45%	49%	52%	54%	54%		55%
Sarkozy	55%	51%	48%	46%	46%		45%
Bayrou	43%	50%	54%	60%	57%		58%
Royal	57%	50%	46%	40%	43%		42%
Bayrou					84%		80%
Le Pen					16%		20%
Sarkozy	49%	51%	53%	54%	54%	53%	51%
Royal	51%	49%	47%	46%	46%	47%	49%
Sarkozy					84%		84%
Le Pen					16%		16%
Royal					75%		73%
Le Pen					25%		27%

Source: IFOP, except March 15: TNS Sofres.
Note: A blank indicates no figure is available. Many more Sarkozy/Royal polls were conducted.

The information in table 2.11 (March 28 and April 19) suffices to determine the Borda-scores among the four candidates. When all head-to-head results are known for a set of candidates, a candidate's Borda-score is the sum of his votes against all opponents. It determines the winner and the order among the candidates. On March 28 the Borda-scores were Bayrou 195, Sarkozy 184, Royal 164, and Le Pen 57. On April 19 they were Bayrou 193, Sarkozy 180, Royal 164, and Le Pen 63. Condorcet and Borda agree on the order of finish. These two ideas, though never (or hardly ever) used in practice, have enjoyed a peculiar but tenacious hold on the minds of social choice theorists down to the present day (they are discussed in detail in chapters 3 and 4).

2.7 The Lessons

Practice proves that the many properties of methods imagined and studied by the theorists of social choice and voting are real. Arrow's and Condorcet's paradoxes occur. Borda- and Condorcet-winners can be eliminated in first- and two-past-the-post systems. Throughout history politicians manipulate. They manipulate who has the right to vote, who has the right to be elected, how the district plans are drawn, what systems to use for allocating seats to parties and regions. When it suits their purposes, and the constitutional rules allow it, they change systems, even in the several months preceding an election. Ad hoc

systems are routinely invented. Some are incredibly complex, some are contradictory. The historical record—and the bloggers of today—suggests that most people seem to believe that conceiving an electoral system is a simple matter invented on the back of an envelope, despite the years of effort and thought that have gone into the development of the theory of social choice.

Practice also shows that given a particular system, both the politicians and the voters act strategically. Parties instruct voters how to vote. Minor candidates throw their hats into the ring—perhaps by personal conviction, perhaps urged (or paid) by others—sometimes changing the outcome by their very presence. Voters may cast their one vote not for their favorite candidate but for another; at other times, when asked or forced to give a rank-ordered list of candidates, voters may place last a candidate they prefer to most others but who they fear presents the greatest threat to their favorite. The analysis of electoral systems must account for such strategic behavior.

Has the theory of social choice and voting responded to these real challenges? An account of the theory is given in chapters 3, 4, and 5. Although what emerges gives a largely negative answer, the material provides a rich foundation of ideas, approaches, concepts, and mechanisms. Chapter 6 argues that the theory fails and presents experimental evidence to show why.

3 Traditional Social Choice

Many a man doing loud work in the world, stands only on some thin traditionality, conventionality; to him indubitable, to you incredible.
—Thomas Carlyle

History reveals three rounds of precursors to the development of a full-blown discipline devoted to the study of voting and the problem of social choice.[1] Every revival of interest seems to have begun in ignorance of the previous work. Nevertheless, the basic model for voting has remained the same from the analyses of Ramon Llull in 1299 and Nicolaus Cusanus in 1433 to those of the Chevalier de Borda (1781) and the Marquis de Condorcet (1785), from the studies of Charles L. Dodgson (1873; 1874; 1876; 1884) and E. J. Nanson (1882) to those of Kenneth Arrow (1951), Duncan Black (1958), and all the ensuing work down to the present day. The preferences of individual judges are imagined to be expressed as rank-orders of the candidates; the messages they send—their votes—are determined by the rank-orders; and the jury is to resolve them into a collective rank-order. In the words of Condorcet, "There exists but one rigorous method to know the will of the greatest number in an election. It consists in taking this will from the respective *merits* of all the candidates, compared pair by pair; that may be deduced from lists on which each elector would write their names, following the *order of merit* he attributes to each" (Condorcet 1789; our emphasis). It is significant that Condorcet and his predecessors focused on the merits of candidates and not the "preferences" of voters.

3.1 Traditional Methods and Concepts

Llull seems to be the first to have carefully specified a system of election, and he specified two. The first is a refinement of what Condorcet proposed five

1. See London and McLean (1990) and Hägele and Pukelsheim (2001; 2008).

centuries later (Condorcet 1789, 26); it is known today as Copeland's method (Copeland 1951). Llull displayed a triangular table of all pairs drawn from sixteen candidates coded by letters of the alphabet (noting that other letters and signs might be used if more were needed) and then stated,

> [It] is necessary to ascertain that in the election three things should be considered, of which the first is honesty and holiness of life, the second is knowledge and wisdom, and the third is a suitable disposition of the heart. Each person having a vote in the Chapter should take an oath by the holy gospels of God to consider these three things and to always elect the person in whom they are best [embodied].
>
> [The triangular table having been prepared, the electors] should betake themselves into a hall and begin to conduct the election in the following way.
>
> Firstly, the two persons whose letters or signs appear in the first cell should leave the hall. And afterwards [somebody] inquires of all others on oath which of the two is better suited and worthier . . . [A] dot is placed by the letter assigned to the person who has the most votes . . . If now one [person] has as many votes as another, then a dot is placed by both letters of this cell . . .
>
> Once the examination of all cells has been completed, the dots of each letter must be counted. And if in any of the letters or signs . . . more dots are determined than in any other of the letters, the person for whom this letter or sign stands is elected . . . If now it happens that as many votes are counted for one as for another, lots are thrown over those who had an equal number of votes in the last election, and the one whose lot wins is elected. (Hägele and Pukelsheim 2001)

A candidate who defeats *every* competitor in head-to-head confrontations is called a *Condorcet-winner*. When the input messages merely rank candidates, a Condorcet-winner appears to have a very strong claim to be chosen. It is for this reason one of the dominant concepts in the theory of voting. When, in Llull's scheme, there is a candidate who defeats all others—when there is a Condorcet-winner—he necessarily collects more "dots" than any other because he collects them against all others and thus no other candidate can do as well. But, as many examples show, it is possible for there to be no Condorcet-winner. In that case Llull's (and later Copeland's) proposal was that the winner should be the candidate who wins the largest number of head-to-head confrontations. In fact, Llull was not crystal clear in his account of what to do when several candidates have an equal number of wins against others (a tie). He may have meant to say that the procedure is to be repeated among those who are tied, or that the one who has the plurality in a vote among all tied candidates is to be elected, or perhaps something else. Drawing by lot was proposed as the ultimate tie-breaker.

In subsequent writings Llull retained comparing by pairs, but simplified (and radically changed) the system by ordering the candidates, voting between the first two, then putting the winner up against the third, and so on through to the

end of the list. A clear statistical advantage is given to a candidate who appears later in the ordered list over one who appears earlier, for the simple reason that he is put up against fewer opponents. In some cases different orders determine different winners, so the order itself can determine the winner.

Cusanus proposed a quite different scheme:

> After a solemn introduction into the electoral business they should decide on the list of candidates who because of their outward or inner qualities may be worthy of so majestic an office [the future emperor]. So that the election may be carried out without fear and in complete freedom and secrecy, they swear oaths at the altar of the Lord that they will elect the best man in the just judgement of a free conscience.
>
> The names of all those on the list of candidates are put down by a notary on identical ballots, with only one name on each ballot. Next to each name the numbers One, Two, Three are written, up to as many as there are persons that have been decided upon as worthy candidates [for example, ten]. Every elector receives ten ballots with the ten names.
>
> . . . [Each] elector should go aside alone . . . and consider the name on every ballot. In the name of God he should ponder, directed by his conscience, who among all candidates is least qualified, and place a simple long mark in ink above the number One. Thereafter he should decide who is next least suitable, and mark the number Two with a simple long overline. Thus he continues until he arrives at the best, in his judgement, and there he will mark the number Ten, or generally the number corresponding to the number of candidates. It is a good idea for the electors to use the same ink, identical pens and the same simple marks . . . so that individual handwritings cannot be identified. This preserves maximum freedom for the electors and peace among all.
>
> . . . [Every] elector should bring his ballots forward and throw them with his own hand into an empty sack hanging in the midst of the electors. When all ballots have been deposited in the sack, the priest who has celebrated the mass should be called, as well as a teller with a tablet on which the names of the candidates are listed . . . [The] priest should take the ballots out of the sack in the order in which they come to hand. He then reads out the name and the number marked, and the teller writes the number next to this name in the tablet. When all ballots are recorded, the teller should add up the numbers next to each name. The candidate who has the highest total shall be king.
>
> . . . It is not possible to discover a method which leads to so infallible a decision more safely. Indeed, all sorts of comparison among all candidates and all confrontations and arguments likely to be made by every elector are included in this system . . . You may well believe that no more perfect method can be found. (Hägele and Pukelsheim 2008)

Cusanus neglected the possibility that some electors might not follow the instructions and give, say, Ten points or One point to several candidates.

In their exhortations, both Llull and Cusanus insisted on "honesty," "holiness," "disposition of the heart," "oaths at the altar of the Lord," "conscience," and so on. Were they concerned with the possibility that voters might manipulate? One cannot help but wonder if asking today's voters to solemnly swear

oaths to elect, with free consciences, the best candidate would in any way alter the outcomes of elections.

Cusanus proposed what today is known as Borda's method (assuming instructions are followed). Borda's 1784 account of the proposal begins by showing that when every elector casts one vote (as they usually do in most countries), and there are (at least) three candidates, A, B, and C, candidate A may well receive the most votes when the electors prefer both B and C to A. He postulated twenty-one electors having the following *profile of preference-orders* (or *preference-profile*),[2] meaning their rank-orderings based on their assessments of the relative merits of the candidates ($A \succ B$ means A is preferred to B):

$$1 : A \succ B \succ C \qquad 7 : A \succ C \succ B \qquad 7 : B \succ C \succ A \qquad 6 : C \succ B \succ A.$$

So, for example, seven voters rank A higher than C, and C higher than B, and therefore A higher than B. It is as usual implicitly assumed that individual preference-orders are rational, meaning transitive, and strict. The English and U.S. plurality or first-past-the-post system elects A with eight votes, while B receives seven votes and C receives six votes, yet thirteen voters prefer both B and C to A. The French system eliminates C in the first round and elects B in the second round. And yet the Condorcet-winner is C. With the same preference-orders, three different systems give three different winners. Voters' inputs are treated as strict preferences, though it is easy enough to allow for indifferences (as is done quite often in the literature); this leads to no fundamentally different analyses or conclusions.

Borda explained that the trouble with plurality voting is that some of the preferences of the voters are ignored. He concluded that "in order for a system of election to be good, it must give the electors the means to pronounce on the merits of each candidate, compared successively with the merits of each of his competitors" (1784, 659–660). His emphasis on the merits was, of course, quite right. He went on to propose Cusanus's method: A's score, with 8 firsts and 13 thirds, is $3 \times 8 + 1 \times 13 = 37$; B's score, with 7 firsts, 7 seconds, and 7 thirds, is $3 \times 7 + 2 \times 7 + 1 \times 7 = 42$; and C's score, with 6 firsts, 14 seconds, and 1 third, is $3 \times 6 + 2 \times 14 + 1 \times 1 = 47$; so C is the *Borda-winner*.

Equivalently, instead of giving 1 point for a last place, 2 points for a second-to-last place, and 3 points for a first place, 0 points could be given for a last, 1 for a second, and 2 for a first; this simply reduces the score of every candidate by the number of electors (in this example, 21) and thus yields exactly the same

2. We use the word "preferences" in the chapters devoted to the traditional model, although they are but a pale reflection of the real thing.

ordering. Explained in these terms, a voter accords k *Borda-points* to a candidate if k opponents are ranked below him; and a candidate's *Borda-score* is the sum of his Borda-points over all voters. This description makes it easier to see a second interpretation of the method, pointed out by Borda: a candidate's score is equal to the sum of the votes he receives in all pair-by-pair votes (see table 3.2b for an example). This is because an elector votes for a candidate whenever he is confronted by a less preferred opponent. Accordingly, to calculate every candidate's Borda-score it suffices to know the tallies of every pairwise vote. The Borda-scores may be used to determine two different outputs: a collective rank-ordering among the candidates, the higher-placed candidates having the higher scores, called the *Borda-ranking*, and a *Borda-winner*, a candidate with the highest Borda-score (there may be several of each when there are ties). For Borda's example the Borda-ranking is $C \succ_S B \succ_S A$ (*the subscript S is systematically used to indicate a society's or a jury's ranking, whatever method is used to establish it*). Borda's method has continued to exert a strong appeal to theorists down to the present day, though it has seen very limited use in voting (it is used, together with an ad hoc procedure to break ties, to rank marching bands in Texas; see chapter 7).

On the other hand, beware: every set of pairwise tallies does not necessarily result from a profile of preference-orders; moreover, a set of pairwise tallies may devolve from more than one profile. The seemingly innocent set of preferences shown in table 3.1 corresponds to no profile of rational preferences. To see why, note that the 80% of the voters who prefer A to C have one of the preference-orders $A \succ B \succ C$, $B \succ A \succ C$, or $A \succ C \succ B$; at most 39% have the first two preference-orders since they all prefer B to C; therefore at least 41% must have the last preference-order. But then at least 41% prefer A to B, a contradiction. Thus a numerical example of pairwise votes that is not accompanied by a profile of preferences is per se suspicious.

Of course, such pairwise votes could occur in practice, but that would mean that some electors had voted irrationally or strategically, for example, for A against B, for B against C, but also for C against A.

In summary, Cusanus and Borda count the *number of votes* of a candidate in all paired confrontations, whereas Llull and Copeland count the *number of wins* in all paired confrontations.

It seems surprising that Llull never noticed what is today known as the *Condorcet paradox*. Condorcet gave many examples of this paradox including one with sixty electors:

$$23: A \succ B \succ C \qquad 2: B \succ A \succ C \qquad 17: B \succ C \succ A$$
$$10: C \succ A \succ B \qquad 8: C \succ B \succ A.$$

Table 3.1
Pairwise Votes Corresponding to No Profile of Preference-Orders

vs.	*A*	*B*	*C*
A	–	40%	80%
B	60%	–	39%
C	20%	61%	–

Note: In this example, *A* obtains 40% of the vote versus *B*.

When A is pitted against B, she wins with 33 votes to B's 27, so this society of electors ranks A ahead of B, or $A \succ_S B$; when B stands alone against C, he receives 42 votes to C's 18, so the society ranks B higher than C, or $B \succ_S C$; and when C runs against A, she amasses 35 votes to A's 25, so $C \succ_S A$. In short, by Llull's and Condorcet's criterion of pair-by-pair comparisons, a society S may have no unchallenged winner and no consistent rank-ordering of the candidates. In symbols, $A \succ_S B \succ_S C \succ_S A$. Not surprisingly, the search for a method that always elects a Condorcet-winner, when he exists, has been the aim of many. Llull (and Copeland after him) proposed exactly such a method, resolving any remaining ties among candidates having the same number of wins by lottery. There have been others as well.

Condorcet discovered that Borda's method could elect a candidate other than the Condorcet-winner and so disdainfully rejected it. An example of this possibility may be seen in tables 3.2a and 3.2b. Borda's method elects C for every valid ϵ (namely, $0 \leq \epsilon \leq 2$), whereas for $1 < \epsilon \leq 2$ the Condorcet-winner is A.

Had Condorcet looked more closely, he might have noticed another damning but more subtle property of the method. Borda's method determines C to be the winner and $C \succ_S B \succ_S A \succ_S D$ to be society's rank-ordering (for any value of ϵ in the interval). But if candidate D withdraws, Borda's method makes A the winner and determines society's rank-ordering to be the exact opposite of what it was alleged to be before: $A \succ_S B \succ_S C$. This strange and unacceptable behavior is Arrow's paradox. A winner and society's rank-ordering of any set of candidates should not change when any other candidate enters or leaves the fray. A good procedure for ranking should be (in the jargon of voting theory) *independent of irrelevant alternatives* (IIA): society's ranking between any two candidates and its designation of a winner should remain the same in the presence or absence of any other candidates. This is particularly real in sports where tentative standings are announced throughout the competition (see chapter 7).

A Condorcet-winner, when he exists, remains a Condorcet-winner whatever other candidate withdraw, so that mechanism satisfies the IIA condition. This is another major reason that theoreticians defend Condorcet.

Table 3.2a
Preference-Profile, $0 \leq \epsilon \leq 2$

$\epsilon\%$	4%	5%	29%	3%	3%	$2-\epsilon\%$	2%	32%	6%	5%	4%	5%
A	A	A	A	A	A	B	B	B	B	C	C	C
B	B	C	C	D	D	A	A	C	D	A	B	B
C	D	B	D	B	C	C	D	D	A	B	A	D
D	C	D	B	C	B	D	C	A	C	D	D	A

Note: In this example, 29% have the preference $A \succ C \succ D \succ B$.

Table 3.2b
Pairwise Votes for Example of Table 3.2a

vs.	A	B	C	D	Borda-Score
A	–	$49+\epsilon\%$	54%	57%	$160+\epsilon$
B	$51-\epsilon\%$	–	49%	65%	$165-\epsilon$
C	46%	51%	–	82%	179
D	43%	35%	18%	–	96

Borda's method is highly vulnerable to manipulation. Any voter who changes her preference-order changes the Borda-scores. Moreover, a coalition of voters acting in concert can change the outcome. In the example of tables 3.2a and 3.2b, 32% of the voters state their preference-orders to be $B \succ C \succ D \succ A$. It is sufficient for half of them to state instead that their preference-orders are $B \succ D \succ C \succ A$ (although they may in fact prefer C to D) to elect B (for this decreases C's Borda-score by 16 and increases D's by the same amount).

Nanson (1882) proposed another method. An elector's message or vote is his preference-order over the candidates, the Borda-scores are calculated, and every candidate whose score is below the average score is eliminated. Repeat the procedure until one candidate remains. Were the votes those of tables 3.2a and 3.2b with $\epsilon = 2$, candidate D would be eliminated (since the average Borda-score is 150); repeating the procedure to decide among A, B, and C eliminates C (the average score is 100, C's score is 97); and A, the Condorcet-winner, is the victor against B. *Nanson's method* always elects the Condorcet-winner, if she exists, since a majority in her favor against every opponent guarantees that her Borda-score is above the average. But Nanson's method also fails where Borda's does: the withdrawal of one or more candidates can change the outcome—it is not independent of irrelevant alternatives. For suppose the preferences of the voters are

$10\% : A \succ B \succ C \qquad 25\% : A \succ C \succ B \qquad 25\% : B \succ A \succ C$
$40\% : C \succ B \succ A.$

The Borda-scores are 95 for A, 100 for B, and 105 for C; A is eliminated and C wins. But if B withdraws, A is the winner. Donald Saari advocates a slightly different method that he calls *instant-Borda-runoff*: "This means (as suggested by E. J. Nanson) that the candidates are first ranked with a Borda-score; the bottom candidate is dropped, and the remaining candidates are reranked with the Borda-score. This continues until one candidate—the winner—remains" (2001b, 103). The following example shows that the methods are different:

$10 : A \succ B \succ D \succ C \qquad 11 : C \succ A \succ D \succ B \qquad 10 : B \succ D \succ C \succ A.$

Nanson's rule first eliminates C and D, and elects A. Saari's rule first eliminates D, then B, and elects C. Instant-Borda-runoff also always elects a Condorcet-winner, when he exists, for the same reason and also fails independence of irrelevant alternatives. Why one of these methods should be used rather than the other is unclear, and not explained.

Australia uses a similar system, the *alternative vote*, to elect the members of its House of Representatives (see chapter 2). This procedure behaves very badly. To begin, consider the following example in which A is eliminated and B wins:

$4\% : A \succ B \succ C \qquad 47\% : B \succ A \succ C \qquad 49\% : C \succ A \succ B.$

It admits Arrow's paradox: the withdrawal of C changes the outcome to A. It happens that A is the Condorcet-winner.

Moreover, the alternative vote fails another test. It is not (to use the technical term) *monotonic*: placing higher may cause a winner to lose. Monotonicity is, we believe, a very important property; any method that is not monotonic should be disqualified. Suppose that a candidate B wins in a race against several others; then B should certainly win again if the preference-orders were the same except that one or more electors ranked B higher. But consider this preference-profile:

$4\% : A \succ B \succ C \qquad 28\% : A \succ C \succ B \qquad 38\% : B \succ C \succ A$
$14\% : C \succ B \succ A \qquad 16\% : C \succ A \succ B.$

C is eliminated in the first round, and B defeats A in the second round. But if the 4% move B to first place, producing the profile

$4\% : B \succ A \succ C \qquad 28\% : A \succ C \succ B \qquad 38\% : B \succ C \succ A$
$14\% : C \succ B \succ A \qquad 16\% : C \succ A \succ B,$

Table 3.3a
Preference-Profile: B the Nanson and Instant-Borda-Runoff Winner

22%	11%	11%	23%	22%	11%
A	A	B	B	C	C
B	D	A	C	A	B
D	C	D	D	D	D
C	B	C	A	B	A

Table 3.3b
Preference-Profile: A the Nanson and Instant-Borda-Runoff Winner

22%	11%	11%	23%	22%	11%
A	A	B	B	C	C
B	B	A	C	A	B
D	D	D	D	D	D
C	C	C	A	B	A

then A is eliminated in the first round, and C defeats B in the second round. The same example shows that the French system is not monotonic. Interpreting it in the opposite direction shows that the alternative vote and the French system are both manipulable: they elect C with the second preference-profile, but by lowering B from first to second place, the 4% who put C in the last position can make B the winner. For example, in the French 2007 election, if Sarkozy's first-round vote had sufficiently increased at the expense of Royal's, then Bayrou would have been his opponent in the second round, and Sarkozy would have lost (according to the overwhelming evidence).

Nanson's method and instant-Borda-runoff are not monotonic either; the example of tables 3.3a and 3.3b shows that a winner, if ranked higher, may lose. Suppose the preference-profile is as given in table 3.3a. D's Borda-score of 111 is the only one below the average of 150, so D is eliminated; among the remaining three, A's 99 is the only score below the average of 100, so A is eliminated; thus with either method B defeats C.

Now suppose that B becomes the second-ranked candidate of all those who had ranked the candidates $A \succ D \succ C \succ B$, so that the preference-orders become those of table 3.3b. Then D and C with Borda-scores of 100 and 145 are eliminated at the first step with Nanson's method; and A wins against poor B, whose only fault was to have become more attractive to the electorate. Instant-Borda-runoff gives the same result: it first eliminates D, then C, and again A defeats B. Thus neither method is monotonic.

An alternative interpretation is that the 11% of voters whose true preference-orders are $A \succ D \succ C \succ B$ manipulate by sending a message other than their true order of preference in which their least preferred candidate is moved up to second place, and thus they elect their first choice A. When a method is not monotonic, there is an opportunity to manipulate. A method that admits Arrow's paradox creates still more transparent opportunities to manipulate.

3.2 IIA and Arrow's Impossibility Theorem

Perhaps the most astonishing aspect of the theory of voting is that despite Arrow's celebrated impossibility theorem and others, the search for a best method of election has continued within the basic framework of the model first conceived by Llull and Cusanus almost a millennium ago, namely, How should a voting system rank all candidates—or determine a winner—given every individual's ranking of them?

The idea that a Condorcet-winner, when she exists, *must* be the winner has well nigh become universally accepted dogma. In many situations a Condorcet-winner exists, but sometimes, as in the previous Condorcet example and in the examples of tables 3.3a and 3.3b, there is no such candidate. There are real examples where none exist as well, though identifying them is not a simple matter because of the lack of information concerning the preference-orders of voters (see chapter 2).

Any two candidates may be compared with the majority-rule: when a candidate A is higher than a candidate B in a majority of the voters' preferences, then A should be ranked higher than B, that is, $A \succ_S B$. The Condorcet paradox is simply an instance showing that the majority-rule does not always give a transitive rank-ordering. Table 3.3a is an example: $A \succ_S B \succ_S C \succ_S A$, there is no Condorcet-winner, but each of the candidates A, B, and C defeats D with the majority-rule. Call the *majority-rule-ranking* the binary relation for which $A \succ_S B$ when a candidate A is ranked higher than a candidate B by a majority of the voters. In the example of tables 3.2a and 3.2b, the majority-rule-ranking happens to be transitive for $1 < \epsilon \leq 2$: $A \succ_S C \succ_S B \succ_S D$ and $X \succ_S Y$ if X is a candidate to the left of Y in the given order. There are preference-profiles that have a Condorcet-winner but for which the majority-rule-ranking is not transitive.

Arrow attacked the problem via an entirely new, axiomatic approach but in the terms of the same basic model. There are many alternative, essentially equivalent statements of Arrow's theorem. One of the points that varies concerns strict preferences $A \succ B$ versus weak preferences $A \succeq B$, that is, $A \succ B$ or $A \approx B$, the latter meaning A and B are tied in a voter's preferences. Heretofore,

strict preferences of the inputs have been implicit, and the possibility of ties in the outputs has largely been ignored. Several candidates may be tied for the highest or any Borda-score, so that there may be several possible Borda-winners and several possible Borda-rankings; and every other method that has been presented may result in ties. For the sake of simplicity the following account assumes strict preferences $\succ$ of inputs and outputs, and outputs are unique, so some arbitrary tie-breaking rule breaks ties (e.g., the lexicographic order of the names of candidates decides). But it must be understood that Arrow's theorem and the theorems and proofs that follow are essentially the same if weak preferences $\succeq$ are used instead and outputs are multiple.

Arrow assumes every voter's input message is a preference-order that is exactly the expression of the voter's preferences (thus excluding any strategic considerations) and that this message is unaffected by the nature of the decision process itself. A *ranking function* maps the voters' inputs into an output that is society's rank-ordering for a fixed set of candidates (in the literature, a ranking function is often called a social welfare function). Arrow considers the set of *all* possible ranking functions and reasons that to be satisfactory such functions must behave properly:

1. There must exist a solution for every possible preference-profile. The function's *domain of profiles is unrestricted*.[3]

2. When every voter i prefers A to B, $A \succ_i B$, then so does society, $A \succ_S B$. The function must respect a *unanimous decision* (in the literature, this property is also called Pareto optimality).

3. Whether society places a candidate A higher or lower than a candidate B can depend only on the relative positions of A and B in the preference-orders of the voters. This is the *independence of irrelevant alternatives* (IIA) property.

4. No one voter's preference-order can always determine society's rank-ordering whatever the preference-orders of all the others. The function must be *nondictatorial*.

Arrow's Impossibility Theorem *There is no ranking function that satisfies the properties (1)–(4) when there are at least three candidates.*

Alternatively, the only system that satisfies properties (1), (2), and (3) is *dictatorial*: one voter's preference-order determines society's preference-order.

The truth of this result has been repeatedly observed for all the proposals discussed so far: each fails to satisfy at least one of the properties. The majority-

3. In the $\succeq$ case it suffices to assume that all profiles of strict preferences $\succ$ are possible.

rule-ranking fails transitivity when the domain is unrestricted; all the others fail IIA. This is a very clean and sparse result. It asks only that three properties be satisfied for all possible profiles, and that is too much. And yet, there are other important properties that good ranking functions should satisfy as well, notably, monotonicity and resistance to strategic manipulation.

Proof The clever proof that follows is surprisingly short and easy (Geanakoplos 2005). First, let B be any candidate, and take a preference-profile in which every voter places B either at the top or at the bottom. Then, it will be seen, a function satisfying properties (2) and (3) must place B either at the top or at the bottom. For, if not, there are different candidates A and C for which $A \succ_S B$ and $B \succ_S C$. Suppose every voter changes by placing C above A while keeping B either first or last. Then IIA implies $A \succ_S B$ and $B \succ_S C$ for the new profile as well (since the voters' orders between A and B, and between B and C remain the same). But for the new profile unanimous decision implies $C \succ_S A$, contradicting the transitivity of the output, society's preference-order.[4]

Second, there is a voter who by changing his preference-order can, for some profile, move a candidate B from the bottom to the top of society's preference-order. Take any profile in which every voter's input has B at the bottom. By unanimous decision B must be at the bottom of the output, society's preference-order. Change the voters' inputs one by one, moving B from the bottom to the top of their lists, until some voter $i = i(B)$ causes B to change position in the output (this must happen sometime because when all voters change, unanimous decision would place B first in the output). Call Φ the profile before i's change, Φ' the profile after i's change. Since B changed from the bottom of the output with the profile Φ, and B is everywhere at the top or at the bottom in Φ', the first argument shows that B must be at the top of the output with profile Φ' and at the bottom of the output with profile Φ.

Third, the voter $i = i(B)$ is a dictator concerning all pairs of candidates A, C other than B. To see that i can cause the output to be $A \succ_S C$, construct the profile Φ'' from Φ' as follows: i moves A above B, so that i's input satisfies $A \succ_i B \succ_i C$; all other voters determine in any way they wish the order between A and C but leave B at the top or bottom position. IIA implies $A \succ_S B$ because all the voters' orders between A and B are as in Φ (where i and society had B

4. The argument for $\succeq$ instead of $\succ$ in the inputs is the same. Suppose not, i.e., $A \succeq_S B$ and $B \succeq_S C$; place $C \succ_i A$ for all voters i; so by transitivity $A \succeq_S C$, by unanimous decision $C \succ_S A$. Similar changes may be made in the next several steps of the proof as well. But Pareto optimality (unanimity) must be extended to indifferences, and the concept of dictatorship must be modified as follows. In comparing any two candidates, there is a sequence of dictators $\mathcal{D}$: the first decides; if he is indifferent, the second decides; if he is indifferent, the third decides; . . . ; if all $\mathcal{D}$ are indifferent, so is society.

at the bottom). IIA also implies $B \succ_S C$ because all the voters' orders between B and C are as in Φ' (where i and society had B at the top). Transitivity of the output implies $A \succ_S C$. But by IIA this implies that $A \succ_S C$ holds whenever $A \succ_i C$.

Fourth, the voter $i = i(B)$ is a dictator concerning B and any other candidate A as well. Let C be some third candidate, and let C take on B's role in the second argument. The third argument shows there must be a voter $j = j(C)$ who is a dictator for all pairs of candidates D, E other than C, in particular A, B. But i alone changed the output from $A \succ_S B$ to $B \succ_S A$ in going from Φ to Φ', so j must be i.

Thus there is no method that satisfies all four conditions. ■

Designate by $\mathcal{C}$ the set of candidates and by $\mathcal{C}'$ any subset of the candidates, $\mathcal{C}' \subset \mathcal{C}$. A *ranking rule* F associates to each profile $\Phi^{\mathcal{C}'}$ on candidates $\mathcal{C}'$ society's rank-ordering of the candidates $\mathcal{C}'$. Versus a ranking function, a ranking rule is defined for any subset of candidates. Arrow's theorem may be expressed in terms of ranking rules.

When Φ is the preference-profile, ϕ_i is the rank-order of voter i. Replace the unanimous decision property by

$2'$. When every voter i has the same rank-order, $\phi_i = \phi$ over the candidates, then society has the rank-order ϕ.

When $\phi_i \equiv \phi_i^{\mathcal{C}}$ is the rank-order of i over all the candidates $\mathcal{C}$ and $\mathcal{C}' \subset \mathcal{C}$, then $\phi_i^{\mathcal{C}'}$ is the rank-order obtained by simply dropping the candidates that are not in $\mathcal{C}'$; and $\Phi^{\mathcal{C}'}$ is the corresponding profile.

Reinterpret the IIA condition as a consistency property:

$3'$. $F(\Phi^{\mathcal{C}}) = \phi^{\mathcal{C}}$ implies $F(\Phi^{\mathcal{C}'}) = \phi^{\mathcal{C}'}$, that is, the order in the output among the candidates $\mathcal{C}'$ alone agrees with the order in the output among all the candidates $\mathcal{C}$. In other words, the ranking between two candidates is not changed by the presence or the absence of another candidate. This is the IIA property that is often violated in practice.

Arrow's Impossibility Theorem (Second Version) *There is no ranking rule that satisfies (1), (2′), (3′), and (4) when there are at least three candidates.*

This version is immediate because (3′) implies (3), and (2′) and (3′) imply (2).

A *choice rule* f has as input a preference-profile $\Phi^{\mathcal{C}'}$ and as output a winner $C \in \mathcal{C}'$. As before, it is defined for any subset of candidates. A *choice function* is defined on a fixed set of candidates (in the literature, a choice function is usually called a social choice function). Several candidates may, of course, be tied as winner in theory and in practice. In practice *generically unique* rules and

functions are sought: outputs are unique save for an exceptional, very small set of profiles. In practice some additional rule specifies the winner among the tied candidates (obviously rare when first-past-the-post or Borda is used in a large electorate but important when the electorate is small). France, for example, in a law of 1999 established by a left-leaning government, broke ties among *conseillers régionaux* (representatives in regional assemblies) by electing the youngest candidate; a right-leaning government changed this in 2003 to electing the oldest candidate.

In all the examples of Arrow's paradox the withdrawal of a nonwinning candidate changed the winner. To analyze this property is the reason that the notion of a *rule* (versus a function) has been introduced. Another version of Arrow's theorem shows that this is true of all mechanisms. Replace the unanimous decision property by

$2''$. When every voter i ranks a candidate C first, society declares C the winner.

Reinterpret the IIA condition as

$3''$. If C is the winner over some subset of candidates $\mathcal{C}'$, $f(\Phi^{\mathcal{C}'}) = C$, and some nonwinner D is dropped, then C remains the winner, that is, letting $\mathcal{C}'' = \mathcal{C}' - D$, $f(\Phi^{\mathcal{C}''}) = C$. This condition may be attributed to Nash (1950) or to Chernoff (1954).

Arrow's Impossibility Theorem (Third Version) *There is no choice rule that satisfies (1), (2″), (3″), and (4) when there are at least three candidates.*

Proof To see this, suppose that there was an f that satisfied the conditions. Then define the ranking rule F recursively, by putting the winner of f first, the winner of f among the remaining candidates second, . . . , or more precisely,

$$F(\Phi) = \{C_1 \succ C_2 \succ \cdots \succ C_n\},$$

where

$$C_i = f(\Phi^{\mathcal{C}^i}) \quad \text{for } \mathcal{C}^i = \mathcal{C} - \{C_1, C_2, \ldots, C_{i-1}\}.$$

As Arrow suggested, "Knowing the social choices [that is, the choice function] made in pairwise comparisons in turn determines the entire social ordering and therewith the social choice function [that is, the ranking function] $C(S)$ for all possible environments" (1951, 28).

It is straightforward to verify that F satisfies properties (1) and $(2')$. To prove that $(3')$ holds, it is shown that if a candidate C_k is dropped, the order among the remaining candidates stays the same. If C_k is dropped, property $(3'')$ implies

that C_1 is the winner of $\Phi^{C^1-C_k}$, so that $C_1 = f(\Phi^{C^1-C_k})$ and thus is in the first position, that C_2 is the winner of $\Phi^{C^2-C_k}$ and thus is in the second position, . . . , that C_{k-1} is the winner of $\Phi^{C^{k-1}-C_k}$ and thus is in the $(k-1)th$ position. The remaining candidates remain in the same order by the definition of F. Since properties (1), (2′), and (3′) hold, F is dictatorial, implying that f is, too. ■

In fact, as shown in chapter 4, the situation for large classes of well-known methods is even more chaotic than Arrow's theorem depicts.

A potential confusion should be clarified here. The independence of irrelevant alternatives (IIA) property has been given a number of conceptually distinct though related definitions, and a considerable literature has been devoted to clarifying them; details are beyond the needs of this book. Some definitions concern only choice functions and not ranking functions, some concern a fixed number of alternatives, others a variable number of alternatives. The IIA property of importance in elections (e.g., the 2000 election of Bush, the 2002 election of Chirac; see chapter 2) and in judging competitions (e.g., skating; see chapter 7) is valid for any method of voting or judging whatever the inputs and implies most of the alternative concepts. It is stated here for rankings (where the first-ranked alternative is the choice and the number of alternatives or competitors is variable).

A ranking function is *strongly independent of irrelevant alternatives* (IIA) if the ranking between any two alternatives does not change when another alternative is either added or dropped.

As a consequence, to compare any two alternatives it is sufficient to consider the inputs concerning them only. This definition encompasses the Arrow, Chernoff, and Nash formulations. It permits giving a precise definition of an already familiar idea.

The *Arrow paradox* is an instance of the violation of strong independence of irrelevant alternatives.

The second version of Arrow's theorem says that there is no ranking rule that is strongly IIA. From this point on the adjective *strong* is dropped: *IIA means strong independence of irrelevant alternatives.*

This formulation of IIA may be considered too strong in other contexts, such as choosing allocations in an economy. For example, with divisible goods, Arrow's axiom precludes using individuals' marginal rates of substitution. Weaker forms of IIA that avoid Arrow's impossibility have been proposed (e.g., Fleurbaey and Maniquet 2008). This book addresses different problems.

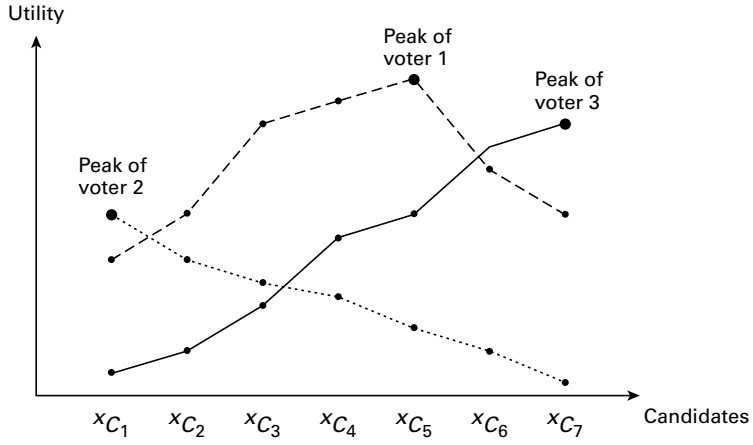

Figure 3.1
Single-peaked preferences.

3.3 Restricting the Domain

In what voting situations is it certain that there always is a Condorcet-winner? And—a slightly more demanding question—in what situations is it certain that the majority-rule-ranking is always transitive? Specifically, is there a restricted domain of preference-profiles where these concepts work?

Duncan Black was the first person to offer answers to these questions: "We will assume that in a committee m [candidates] are put forward . . . and that an ordering can be found for the points to represent these [candidates] on the horizontal axis, such that the [voters'] preference curves become single-peaked" (1958, 14–15). Identify the different candidates C with points x_C on the x-axis, and represent any voter's preference by a graph $y = g(x)$, where $g(x_{C_i}) > g(x_{C_j})$ means the voter ranks candidate C_i higher than C_j. A graph g is *single-peaked* if it either decreases from beginning to end or increases from beginning to end or first increases and then decreases (see figure 3.1).

Whenever the domain is restricted to preference-orders that are single-peaked for *every* voter *relative to one common order*, there must exist a Condorcet-winner, and the majority-rule-ranking is transitive (when the number of voters is even a tie-breaking mechanism may be necessary). Suppose there are an odd number of voters. Call the voter whose preferred candidate C^* is the median or middlemost of the preferred candidates of all voters, the *median-voter*. The Condorcet-winner is C^*, for C^* has an absolute majority of the votes against any other candidate C. If C is to the left of C^*, then C^* obtains the votes of the

median-voter and all voters whose preferred candidates are to the right of C^*; and if C is to the right of C^*, then C^* obtains the votes of the median-voter and all voters to the left of C^*.

Essentially the same reasoning establishes the transitivity of the majority-rule-ranking because if any candidate is dropped, the voters' preference-orders among the remaining candidates are still single-peaked, so if the Condorcet-winner is put aside there is a (second) Condorcet-winner among the remaining candidates; continuing, the majority-rule-ranking is found. Notice, however, that this argument fails when there is an even number of voters; for example, if

$$50\% : A \succ B \succ C \qquad 50\% : B \succ C \succ A;$$

then by majority-rule A and B are tied, $A \approx_S B$, and also A and C are tied, $A \approx_S C$, but B defeats C, $B \succ_S C$, so the majority-rule-ranking when ties are possible (with $A \succeq_S B$ meaning either $A \succ_S B$ or $A \approx_S B$) is not transitive because $A \succeq_S B \succ_S C \succeq_S A$. In an election with thousands or millions of voters such ties with the majority-rule are so unlikely that they may reasonably be dismissed (as is done in a model that follows), so Black's result may be accepted in general; it is true generically.

Is it reasonable to believe that there are elections among candidates that satisfy Black's condition? The stereotyped classical political cleavage of left versus right is the obvious possibility, but the empirical work on spatial voting models has clearly discarded the idea that one dimension suffices (e.g., Enelow and Hinich 1984). A recent electoral experiment proves that the condition is far from satisfied in practice (see chapter 6), although the data yield a probabilistic sense of left to right. Black himself had doubts, for he advocated another, hybrid method. *Black's method*: "[The] Condorcet criterion should first be used to pick out the majority candidate if there is one; and if no majority candidate exists, that candidate should be chosen who has the highest Borda-score" (1958, 66).

Black's method does not behave continuously: small changes in the messages of the voters or their numbers can lead to big abrupt changes in the results. Specifically, suppose f is either a choice function or a ranking function and $\xi \in f(\mathcal{P}_i)$ for every preference-profile $\mathcal{P}_i$ in a sequence approaching the profile $\mathcal{P}$, $\mathcal{P}_i \to \mathcal{P}$. Then $\xi \in f(\mathcal{P})$ if f is continuous. In the example of tables 3.2a and 3.2b, the Borda-ranking is $C \succ_S B \succ_S A \succ_S D$ whenever $0 \leq \epsilon \leq 2$. When $0 \leq \epsilon \leq 1$ there is no Condorcet-winner, and when $1 < \epsilon \leq 2$ the majority-rule-ranking is transitive with $A \succ_S C \succ_S B \succ_S D$. Therefore Black's method suddenly vaults A from third place to first place as ϵ increases from below 1 to above 1. Any hybrid method is likely to exhibit this type of

behavior. The two methods are almost opposed: the Condorcet-winner is resolutely in the next-to-last position in the Borda-ranking. This is no isolated phenomenon. Suppose there are $m+1 \geq 4$ candidates and $2m+1$ voters with the preference-profile

$$m+1 : A_1 \succ A_2 \succ \cdots \succ A_{m-1} \succ A_m \succ A_{m+1}$$
$$m : A_m \succ A_{m-1} \succ \cdots \succ A_2 \succ A_{m+1} \succ A_1.$$

Then A_1 is the Condorcet-winner but next-to-last in the Borda-ranking (when $m+1=4$ A_1, is tied for next-to-last).

Black's restriction to single-peaked preference-profiles naturally led to the question, What are not only sufficient but also necessary restrictions on the preference-profiles of a society to guarantee the existence of a Condorcet-winner and/or a transitive majority-rule-ranking? The essence of the answer to the second part was established in 1969 (Inada 1969 for ranking functions, Sen and Pattanaik 1969 for choice functions). It is described in the context of the model of Partha Dasgupta and Eric Maskin (2008), who postulated a continuum of voters rather than a finite number and asked that the majority-rule-ranking be *generically transitive*, meaning transitive except for ties, when exactly half the voters prefer one candidate to another—hardly a likely event when there are many voters.

Consider any three candidates A, B, and C. The preference-orders of voters can contain two different *Condorcet-cycles* among them:

$$A \succ B \succ C \qquad B \succ C \succ A \qquad C \succ A \succ B,$$

or

$$A \succ C \succ B \qquad C \succ B \succ A \qquad B \succ A \succ C.$$

Theorem *The majority-rule-ranking is transitive on a domain of preference-profiles if and only if the domain is restricted to profiles that contain no Condorcet-cycle among any triple of candidates.* (Dasgupta and Maskin 2008; the same result with a finite, odd number of voters is given in Sen and Pattanaik 1969.)

This restricts the preference-orders of voters to those that omit at least one of the orders of each of the Condorcet-cycles on every three candidates. It is a simple matter to verify that single-peaked preferences satisfy this condition. There are, however, profiles for which the condition does not hold and yet there is a transitive majority-rule-ranking and so a Condorcet-winner (see section 4.3). This occurs because a restriction on the domain says nothing about the number of voters involved (e.g., in a large electorate, extremely few violations

of the condition may have no influence). As mentioned earlier, there are also real-life, historical examples where there is no Condorcet-winner. The following preference-profile is a real example that violates the restriction on the domain. It comes from the election of a president of the Social Choice and Welfare (SCW) Society (Brams and Fishburn 2001; Saari 2001a):

$13 : A \succ B \succ C \qquad 9 : B \succ C \succ A \qquad 11 : C \succ A \succ B$
$11 : A \succ C \succ B \qquad 8 : C \succ B \succ A.$

The first three preference-orders belong to a Condorcet-cycle and so violate Dasgupta and Maskin's restriction. With this profile C is the Condorcet-winner and $A \succ_S C \succ_S B$ is the Borda-ranking. Borda's method does not satisfy IIA because when B withdraws, C is first in the Borda-ranking. (Note that if only the first thirty-three voters had expressed themselves, Condorcet's paradox would have been realized.)

Dasgupta and Maskin proposed yet another hybrid model that champions the primacy of a Condorcet-winner and more generally the majority-rule-ranking. They asked whether there is a ranking function that works well for a less restricted domain of preference-profiles than does the majority-rule. A function works well if it treats all voters equally and all candidates equally ("impartiality") and is unanimous and independent of irrelevant alternatives. They answered *no:* "Specifically, we establish ... that if a given voting [function] F works well on a domain of [preference-profiles], then majority rule works well on that domain too. Conversely, if F differs from majority rule, there exists some other domain on which majority rule works well and F does not" (2008, 953).

Dasgupta and Maskin then singled out the Borda-ranking as a second-best method because it is unanimous and impartial and gives a transitive ranking of the candidates. Its manipulability and violation of independence of irrelevant alternatives are its glaring failures.

Theorem *The Borda-ranking satisfies independence of irrelevant alternatives on a domain of preference-profiles if and only if the domain is restricted to profiles for which one among any three candidates is either unanimously first among the three, or unanimously second among the three, or unanimously last among the three*. (Dasgupta and Maskin 2008)

The theorem invokes a very restrictive condition that can hardly be assumed to occur; for example, the preceding example of the SCW Society violates it.

They concluded that the simplest way to overcome the possible absence of a Condorcet-winner is what we call the *Dasgupta-Maskin method*: "If no one obtains a majority against all opponents, then among those candidates who

defeat the most opponents in head-to-head comparisons, select as winner the one with the highest [Borda-score]" (2004, 97). This is a hybrid method, a refinement of Copeland's (or Llull's) method: if several candidates are tied as Copeland-winners, then apply Borda's (or Cusanus's) method to decide among them. It yields exactly the same outcomes in the example of tables 3.2a and 3.2b as does Black's method, and so it is not continuous. Attempting to define a method of voting that guarantees electing the Condorcet-winner, when she exists, based on some other rationale is bound to behave in this manner. The method is, of course, manipulable. The example of tables 3.2a and 3.2b with $\epsilon = 2$ shows it. A, the Condorcet-winner, wins. But if A is dropped to third place by the 5% of the voters with the ranking $C \succ A \succ B \succ D$, there is no Condorcet-winner; each of the three candidates A, B, and C beat two others; and C, the Borda-winner, wins, much to the satisfaction of those voters. Surprisingly, the method was actually used well before it was formally defined (see chapter 7).

What is fascinating in the centuries-old history of the debates that have raged over how to elect and how to rank is the persistent dominance of the concepts of the Condorcet-winner and the Borda-ranking. In fact, they are not compatible ideas.

4 Electing versus Ranking in the Traditional Model

By my troth, this is the old fashion; you two never meet but you fall to some discord: you are both, in good troth, as rheumatic as two dry toasts; you cannot one bear with another's confirmities.
—William Shakespeare

Condorcet himself defined a method of ranking that behaves continuously, always places the Condorcet-winner first (when he exists), and always agrees with the majority-rule-ranking (when it is transitive). Moreover, it almost satisfies IIA. These facts were largely ignored until H. Peyton Young read and analyzed Condorcet's famous *Essai* with care (Condorcet 1785; Young 1988). This oversight was due in part to the first person to have looked into the history of the methods used for voting, Duncan Black. Black (1958) discussed Borda's and Condorcet's ideas at length, but much of his commentary—most notably that concerning Laplace's and Galton's contributions but also those of Condorcet—was seriously misleading. Black paid a good deal of attention to the papers of Charles L. Dodgson (better known as Lewis Carroll), reprinting three of his pamphlets (Dodgson 1873; 1874; 1876). Black established that Dodgson could have had no inkling of his precursors' contributions. Yet, realizing the inherent injustice of the first-past-the-post system, Dodgson reinvented Borda's method and he published it. It was immediately used to decide on an Oxford University appointment. The winner having won by only one point, the committee held a runoff between the top two of the Borda-ranking. Arrow's paradox raised its ugly head—the order between them was reversed—so Dodgson dropped the idea, advocating instead the Condorcet-winner if there was one and otherwise no winner. For he could find no satisfactory rule for resolving what he aptly called *cyclical majorities*, situations such as $A \succ_S B \succ_S C \succ_S D \succ_S A$, where each candidate defeats his successor. Later, however, Dodgson seems to have concluded that "an election [is] more of a game of skill than a real test of the wishes of the electors," and he argued that this is what makes cyclical majorities

important: "Suppose A to be the candidate whom I wish to elect, and that a division is taken between B and C; am I bound in honour to vote for the one whom I should really prefer, if A were not in the field, or may I vote in whatever way I think most favourable to A's chances? Some say 'the former,' some 'the latter.' I proceed to show that, whenever [there is no Condorcet-winner] and there are among the electors a certain number who hold the latter course to be allowable, the result *must* be a case of cyclical majorities" (cited in Black 1958, 232, 265).

Black casually dismissed Nanson's method despite its election of the Condorcet-winner (when he exists), claiming that it would be unintelligible to the average voter and too laborious. He discussed Condorcet's "probabilistic" approach and correctly reported that Condorcet had subsequently abandoned it, but he failed to see that it defined a method of ranking. For this he should be forgiven because Condorcet's description was confused, complicated, and erroneous.

4.1 Condorcet's Method of Ranking

For Condorcet an election is primarily a consultation of the members of society in the quest of the best possible winner or the truest possible ranking according to merit. In the same spirit as his celebrated jury theorem, Condorcet reasons that electors may commit errors of judgment, but they have an independent, equal, and better-than-even chance $p > \frac{1}{2}$ of choosing the better of every pair of candidates. Given the results of votes on every pair of candidates, Condorcet posed the problem of finding the ranking among all candidates that has the highest probability of being "correct" in view of the electors' opinions. This is known today as the maximum likelihood estimate over all possible rankings (Young 1988).[1] How to calculate this estimate is easy to explain. A voter contributes k *Condorcet-points* to an arbitrary rank-ordering if her preference-order agrees with it in k pair-by-pair comparisons of the candidates. The *Condorcet-score* of a rank-ordering is the sum of its Condorcet-points over all voters. Equivalently, letting $v(A_i \succ A_k)$ be the number of voters that rank candidate A_i ahead of A_k, the Condorcet-score of a rank-ordering $A_1 \succ A_2 \succ \cdots \succ A_n$ is

$$\sum_{i=1}^{n-1} \sum_{k>i}^{n} v(A_i \succ A_k).$$

1. This model is not entirely satisfactory, however, because asking for a voter's ranking of the candidates implies that the pair-by-pair comparisons are not independent. How to model this, let alone how to do the ensuing computations, is not at all clear.

Table 4.1
Pairwise Votes for Example of Tables 3.2a and 3.2b

vs.	*A*	*B*	*C*	*D*	Borda-Score
A	–	$49+\epsilon\%$	54%	57%	$160+\epsilon$
B	$51-\epsilon\%$	–	49%	65%	$165-\epsilon$
C	46%	51%	–	82%	179
D	43%	35%	18%	–	96

The Condorcet-score of the ranking $A \succ_S C \succ_S B \succ_S D$ in the example of table 4.1 is

$$v(A \succ C) + v(A \succ B) + v(A \succ D) + v(C \succ B) + v(C \succ D) + v(B \succ D),$$

or $54+49+\epsilon+57+51+82+65 = 358+\epsilon$, and that of $B \succ_S A \succ_S C \succ_S D$ is $358-\epsilon$.

The *Condorcet-ranking* is the rank-ordering that maximizes the Condorcet-score (there may be several when there are ties). The calculation has a convincing ring in and of itself. It establishes a linear order among all the possible rankings of the candidates. Condorcet proposed that the first-place candidate of the Condorcet-ranking should be declared the winner. This is also known as *Kemeny's rule* (Kemeny 1959; 1962). Kemeny's idea (developed in ignorance of Condorcet's writings) was to find a "consensus" ranking that is "least distant" from the electors' rankings.

When $\epsilon = \frac{1}{2}$, the Condorcet-ranking of the candidates in table 4.1 is $A \succ_S C \succ_S B \succ_S D$; and when $\epsilon = 0$, there is a tie and $B \succ_S A \succ_S C \succ_S D$ is also a Condorcet-ranking. Both are very different from the Borda-ranking $C \succ_S B \succ_S A \succ_S D$. When there is a Condorcet-winner, she must be at the top of the list, and when there is a *Condorcet-loser*—a candidate who loses to every other candidate in head-to-head confrontations—he must be at the bottom. To see this, notice that D is a Condorcet-loser (for any $0 \leq \epsilon \leq \frac{1}{2}$), and suppose D is not the last in the Condorcet-ranking, for example, $C \succ_S D \succ_S A \succ_S B$. Changing this ranking by switching the positions of D and A changes nothing in the total Condorcet-score except that the number of votes for A against D is substituted for the number of votes for D against A, so the total Condorcet-score of the new ranking can only be larger. It follows that since a Condorcet-loser is defeated by every other candidate he must be the last in a Condorcet-ranking. Similar arguments show that a Condorcet-winner (if she exists) must be first in a Condorcet-ranking and that when the majority-rule-ranking is transitive, it is the Condorcet-ranking. The Condorcet-ranking thus supports the idea of the Condorcet-winner and the majority-rule-ranking.

On the other hand, the Condorcet-ranking's winner is manipulable, as are all the choice functions previously mentioned. When $\epsilon = \frac{1}{2}$, the Condorcet-ranking is $A \succ_S C \succ_S B \succ_S D$. If half of the 4% of voters with preferences $C \succ B \succ A \succ D$ change to $B \succ C \succ A \succ D$, then the Condorcet-ranking becomes $B \succ_S A \succ_S C \succ_S D$, electing B, whom they prefer to A.

Young develops another argument that seriously undermines the legitimacy of the Condorcet-winner. He shows that when $p > \frac{1}{2}$ but is close to $\frac{1}{2}$, the maximum likelihood estimate of the winner is the Borda-winner and not the first-place candidate of the Condorcet-ranking. He goes on to argue that when there are many voters and p is comfortably greater than $\frac{1}{2}$, the Borda-winner, the majority-rule winner, and the winner with any reasonable method will almost certainly be one and the same. From all of these perspectives, the first-place candidate of the Condorcet-ranking—and by association, the Condorcet-winner—certainly does not have an unchallenged claim to be declared the winner.

It is instructive to look more closely at the tie that occurs between exactly two Condorcet-rankings when $\epsilon = 0$, as in the last example. The two orders are

$$A \succ_S C \succ_S B \succ_S D \quad \text{and} \quad B \succ_S A \succ_S C \succ_S D.$$

By implicitly or explicitly studying the problem of choice rather than the problem of ranking, most of the past work would reject such ties. Or, if such a tie did exist, then it would be interpreted as meaning that there is a three-way tie for first place—B tied for first surely qualifies C as tied for first, too—so that society's order should be $A \approx_S B \approx_S C \succ_S D$. This point of view is too restrictive. What explains the curious nature of the two orders is the fact that the outputs of the traditional model are rankings of the candidates, whereas the outputs of Condorcet's method are rankings of rankings of candidates.

Condorcet himself seems to have realized that the maximum likelihood estimate of who the winner should be among the candidates depends (in a complicated fashion) on the value of $p > \frac{1}{2}$ and is not necessarily the Condorcet-winner. He then abandoned this approach and the Condorcet-ranking altogether to champion the one simple idea of the Condorcet-winner. Young (1986) was the first to see the distinction clearly: "Even if we reject the specific probabilistic model by which this conclusion is reached, there are still strong a priori grounds for asserting that the [Condorcet-ranking and the Borda-winner] are the optimal rules for ranking and choice respectively."

A choice function or a ranking function is *anonymous* if it treats all voters equally, *neutral* if it depends only on the numbers of the different preference-orders of a profile (or treats all candidates equally), and *impartial* if it is both anonymous and neutral. These are clearly essential properties of choice and

ranking functions. Imagine now that a society is split into two separate parts, each with its preference-profile, and that a choice or a ranking function f determines a set of winners $\mathfrak{W}_1$ and $\mathfrak{W}_2$ of each part or a set of rankings $\mathfrak{R}_1$ and $\mathfrak{R}_2$ of each part. f is a *join-consistent* choice rule when f determines the set of winners of the entire society to be those candidates that are winners in both $\mathfrak{W}_1$ and $\mathfrak{W}_2$ (whenever they have a winner in common, $\mathfrak{W}_1 \bigcap \mathfrak{W}_2 \neq \emptyset$). f is a *join-consistent* [2] ranking rule when f determines the set of rankings of the entire society to be those rankings that are in both $\mathfrak{R}_1$ and $\mathfrak{R}_2$ (whenever they have a ranking in common, $\mathfrak{R}_1 \bigcap \mathfrak{R}_2 \neq \emptyset$). This idea seems reasonable: the solutions in common to both groups (if such exist) should be the solutions of the two groups reunited.

Suppose that f is a ranking function that selects the ranking

$$A_{j_1} \succ_S \cdots \succ_S \overbrace{A_{j_k} \succ_S A_{j_{k+1}}} \succ_S \cdots \succ_S A_{j_m}$$

for some preference-profile, where m is the number of candidates and $A_{j_{k+1}}$ immediately follows A_{j_k}. Then f is *near-majority-rule* if the number of voters who rank A_{j_k} higher than $A_{j_{k+1}}$ is at least as great as the number who rank $A_{j_{k+1}}$ higher than A_{j_k}; moreover, if those numbers of voters are equal, then there is a tie and f must also select the ranking

$$A_{j_1} \succ_S \cdots \succ_S \overbrace{A_{j_{k+1}} \succ_S A_{j_k}} \succ_S \cdots \succ_S A_{j_m},$$

where the only change in the ranking is that the neighboring A_{j_k} and $A_{j_{k+1}}$ are reversed. Majority-rule does not in general yield a transitive order; this condition simply asks that it be satisfied among the immediate successors of a solution. It has already been shown that the Condorcet-ranking does enjoy this property.

Theorem *The unique ranking rule that is impartial, join-consistent, and near-majority-rule is the Condorcet-ranking*. (Young and Levenglick 1978)

Note that earlier a rule (as opposed to a function) meant that the number of candidates varies; here it is taken to mean that the number of voters varies. Elsewhere, the context will determine the sense of rule versus function.

4.2 Borda's and Sum-Scoring Methods

Consider now the problem of choice rather than that of ranking. To begin, notice that when the profile is a *Condorcet-component*—equal numbers of voters

2. Young sometimes calls this "reinforcement," at other times "consistency."

having each of the preference-orders of a Condorcet-cycle—such as $3k$ voters with the profile

$$k : A \succ B \succ C \qquad k : B \succ C \succ A \qquad k : C \succ A \succ B,$$

impartiality requires that all candidates must be in a vast tie as winners. But there are other profiles for which the same ought to be true, namely, when for *every* pair of candidates A and B the number of voters who rank A higher than B equals the number who rank B higher than A. f has the *cancellation property* if in this case it selects all the candidates (there is again a gigantic tie among them all). There can be many such situations. An example is equal numbers of voters having each of the two preference-orders

$$k : A \succ B \succ C \succ D \qquad k : D \succ C \succ B \succ A.$$

Make the innocuous assumption that f is *faithful*: if there is only one voter, then his highest-place candidate is f's choice.

Theorem *The unique choice rule that is impartial, join-consistent, faithful, and satisfies the cancellation property is the Borda-winner.* (Young 1975)

The method of Borda is a special case of a more general class. A *scoring scheme* assigns a real number score s_i to each voter's ith-place candidate. A *sum-scoring method* assigns a candidate the sum of her scores over all voters and elects the one having the largest total score. Among the sum-scoring methods Borda's is characterized as any one for which $s_1 > s_2$ and $s_{i+1} - s_{i+2} = s_i - s_{i+1}$ for $i = 1, \ldots, m-2$. Clearly, the only reasonable scoring schemes are strictly monotone, that is, satisfy $s_1 > \cdots > s_m$.

Suppose f is a choice rule that designates candidate A as the unique winner for the preference-profile $\mathcal{P}$; in symbols, $f(\mathcal{P}) = A$. f *respects the choice of large electorates* if A is the winner when the preference-profile is any $\mathcal{Q}$ together with the profile $\mathcal{P}$ replicated a sufficient number of times; in symbols, there is an integer n large enough to guarantee that

$$f(\mathcal{Q} + \overbrace{\mathcal{P} + \cdots + \mathcal{P}}^{n}) = A,$$

where $+$ means taken together in one profile.

Theorem *The unique choice rules that are impartial, join-consistent, and respect large electorates are the sum-scoring methods.* (Young 1975)

There is a parallel result that addresses rankings. A ranking rule f is *pairwise join-consistent* if $A \succ B$ for the preference-profile $\mathcal{P}$ and $A \succeq B$ for the preference-profile $\mathcal{Q}$, then $A \succ B$ for $\mathcal{P} + \mathcal{Q}$; and if $A \approx B$ for both profiles,

then the same holds in the joint profile. *f respects the pairwise rankings of large electorates* if $A \succ B$ in $\mathcal{P}$, then for any $\mathcal{Q}$ there is an integer n large enough to guarantee that $A \succ B$ for the profile

$$\mathcal{Q} + \overbrace{\mathcal{P} + \cdots + \mathcal{P}}^{n}.$$

Theorem *The unique ranking rules that are impartial and pairwise join-consistent and that respect the pairwise rankings of large electorates are the sum-scoring methods*. (Smith 1973)

Borda's method may not elect the Condorcet-winner (when he exists). Much more may be said, for there is a basic opposition between Condorcet-winners and sum-scoring-winners, of which one manifestation is the following.

Theorem *The Condorcet-winner of some profiles is elected by no strictly monotonic sum-scoring method*. (Fishburn 1984)

The proof is in the seven-voter example

$$3 : C \succ A \succ B \qquad 2 : A \succ B \succ C \qquad 1 : A \succ C \succ B \qquad 1 : B \succ C \succ A.$$

C is the Condorcet-winner, but any sum-scoring method with $s_1 > s_2 > s_3$ elects A because A's score is $3s_1 + 3s_2 + s_3$ whereas C's is only $3s_1 + 2s_2 + 2s_3$.

In fact, as Donald Saari (2000) has conclusively shown, *anything can happen* when sum-scoring methods are used. His idea is to find a coordinate system for the space of all possible profiles so as to identify *everything* that can go wrong in using sum-scoring methods and methods that depend on pair-by-pair comparisons of candidates. This theory is too elaborate to explain here, except to relate its most striking conclusion. Given a set of n candidates, the output of a method of election is a list that rank-orders every possible subset of the candidates. Saari (1989; 1992) sets out to study all possible inputs and outputs, and he calls this a "dictionary." He shows that with any sum-scoring method M the following is possible. Given any outputs $\mathfrak{O}$—consisting of a rank-ordering of the n candidates and rank-orderings of all subsets of the candidates however chaotic or contradictory—there is an input preference-profile for which M's output is $\mathfrak{O}$. For instance, an earlier example (tables 3.2a, and 3.2b) shows that the Borda-ranking is $C \succ_S B \succ_S A \succ_S D$, but when D drops out, the Borda-ranking is reversed to $A \succ_S B \succ_S C$. Notice that when first-past-the-post or Borda's rule is used, it is clear how the rules are applied to subsets of candidates; with sum-scoring methods in general, how the rule applies to subsets of candidates must be specified.

Nevertheless, Saari's conclusion is to opt for Borda's method: "What provides hope from these dictionaries is that the Borda Count ... is the unique rule (when used with every subset of candidates) that significantly minimizes the number and kinds of allowed paradoxes. Thus the Borda Count enjoys the maximum number of positive properties; e.g., only Borda always ranks a Condorcet-winner over a Condorcet loser" (2009, 4). Borda's method may minimize misbehavior among all sum-scoring methods, but it certainly does not eliminate it (as is shown by experimental evidence in chapters 6 and 19 as well as the example of tables 3.2a and 3.2b). Moreover, it fails in other important dimensions (as will be seen anon): it is highly manipulable and extremely biased in favor of centrist political candidates or in favor of good but unexceptional competitors.

4.3 Objections to Condorcet-Consistency

Borda's method is characterized by conditions that concern selecting a *candidate*, Condorcet's by conditions that concern selecting a *ranking* of the candidates. That is why Borda's method is singularly suited to choosing a winner (but not a ranking) and Condorcet's to choosing a ranking (but not a winner). This point is strikingly evident in an example with eighty-one voters that Condorcet invented to argue Borda's method was bad, but that Saari (2001b) used even more effectively to show the questionable legitimacy of a Condorcet-winner in the context of the traditional model:

$$\begin{array}{lll} 30 : A \succ B \succ C & 1 : A \succ C \succ B & 29 : B \succ A \succ C \\ 10 : B \succ C \succ A & 10 : C \succ A \succ B & 1 : C \succ B \succ A. \end{array}$$

The Borda-score ranks the candidates $B \succ_S A \succ_S C$, yet A is the Condorcet-winner. Intuitively, the reason for the Borda outcome is that A wins over B by the narrowest of margins and defeats C comfortably, whereas B trounces C by so much that he emerges as the over-all winner. But look closer. Thirty of the eighty-one voters have preferences that constitute a Condorcet-component

$$10 : A \succ B \succ C \qquad 10 : B \succ C \succ A \qquad 10 : C \succ A \succ B.$$

These voters cancel each other out when it comes to designating a winner because there is a perfect symmetry among all the candidates; said differently, these voters together say that the candidates A, B, and C are tied. Indeed, the same is true for another Condorcet-component that is contained in the preference-profile:

$$1 : A \succ C \succ B \qquad 1 : C \succ B \succ A \qquad 1 : B \succ A \succ C.$$

Since these thirty-three voters cancel each other out, their preference-orders may be dropped, and the election outcome should be decided by the remaining forty-eight voters, whose profile is

$20 : A \succ B \succ C \qquad 28 : B \succ A \succ C.$

B is the clear winner. The Condorcet-winner (when he exists) seems *not* to be the candidate who should win in *every* case. And by the same token, the Condorcet-loser is not the candidate who should be last in every case.

There is an alternative, perhaps more positive, explanation of why the Condorcet-winner is not the candidate who should always win. If a candidate is ranked first by a majority of voters in some profile, then surely he should be the winner when a Condorcet-component is adjoined to the profile. Consider the same example. B is the first-place candidate for a majority of voters in the forty-eight-voter preference-profile. Therefore, the first-place candidate of the seventy-eight-voter profile obtained by adjoining the thirty-voter Condorcet-component should be B; but for that profile A is the Condorcet-winner.

It is this fundamental observation that Saari generalized to build his elaborate but telling geometric theory. It clearly shows the various classical voting paradoxes and makes it easy to manufacture situations where they occur. Its focus is scoring and pairwise voting methods (meaning those, such as Borda's and Condorcet's, that depend only on the matrix of pairwise votes), and it studies them as methods of choice *and* of ranking in a vector space whose coordinates are all possible preference-orders. Saari identifies what he calls profile deviations, which are those parts of profiles that cause "all election difficulties" (2000, 4). They primarily come from Condorcet-components of profiles as observed in the last example but more generally from parts of profiles such as those identified by Young in his concept of cancellation (when the number of voters placing a candidate A higher than a candidate B equals the number placing A lower than B).

Saari concludes the following:

- "[All] profile deviations reflect some procedure's inability to recognize certain kinds of informational symmetry" (2000, 58).
- "The [Borda-score] outcome for *all* n-candidates is the unique *ranking* which avoids all of the indicated problems" (2000, 57; our emphasis).

Saari claims that Borda's dominates all other scoring methods in avoiding difficulties because in addition to the Condorcet-components it "cancels out" all the symmetries that cause noxious deviations.

· "The encouraging news is that the measures developed for 'what the voters really want' isolate a unique procedure, the Borda-score, which meets all expectations" (2001b, 110).

These conclusions ignore the distinction between ranking and choice; moreover, Borda's method and the more general scoring methods are all highly manipulable. Saari recommends the following:

· Instant-Borda-runoff should be used "to combine the consistency of Borda outcomes while frustrating manipulative voting" (2001b, 103).

Instant-Borda-runoff is not monotonic and is manipulable, as seen in the example of tables 3.3a and 3.3b, though it is perhaps slightly less obviously manipulable than Borda's method, and it always elects the Condorcet-winner (when he exists). Saari himself argues against the reranking procedure used by instant-Borda-runoff: "This result [a corollary stating that changes in the Borda-ranking when candidates are dropped are due to Condorcet-components of the profile] provides a strong argument to ignore the [Borda-rankings] of subsets and to place value on the [Borda-ranking] of all n candidates . . . the reranking process dismisses valuable information about the voters . . . we should keep and use the original [Borda-]ranking of all n-candidates" (2000, 40).

Moreover, any instant-runoff sum-scoring method, including instant-Borda-runoff, violates monotonicity. The proof is given by the following example (Fishburn 1982):

$$6 : A \succ B \succ C(:9) \qquad 3 : A \succ C \succ B(:3) \qquad 4 : B \succ A \succ C(:1)$$
$$6 : B \succ C \succ A(:6) \qquad 6 : C \succ A \succ B(:8) \qquad 2 : C \succ B \succ A(:0).$$

Take $s_1 \geq s_2 \geq s_3 = 0$. A's score is $9s_1 + 10s_2$, B's is $10s_1 + 8s_2$, C's is $8s_1 + 9s_2$, so C is eliminated and A is elected. If three of the four voters with $B \succ A \succ C$ and both with $C \succ B \succ A$ move A up one slot, giving the profile of preferences indicated in parentheses, B is eliminated and C is elected.

Saari's interesting and elaborate theory provides additional proof that the traditional model is, as he said, chaotic. He chooses Borda's method as the least chaotic in the context of the traditional model. We believe the results show the model must be discarded.

Instant-Borda-runoff may be seen as a choice rule that, along with Llull's, Copeland's, Nanson's, Black's, Dasgupta-Maskin's and others', seeks to elect the Condorcet-winner whenever there is one.

A choice rule is *consistent with Condorcet* if the rule elects the Condorcet-winner whenever she exists.

Saari's objection to Condorcet-winners made on the basis of Condorcet's eighty-one-voter example is easily generalized to all choice rules that are consistent with Condorcet. A choice rule *cancels properly* if A is the winner for the profile $\mathcal{P}$ implies A is the winner for the profile $\mathcal{P}+\mathcal{C}$, where $\mathcal{C}$ is a Condorcet-component.

Theorem *There is no choice rule consistent with Condorcet that cancels properly.*

Proof Take any rule consistent with Condorcet, and suppose there are $n \geq 3$ candidates, $A_1, \ldots, A_n$. For n even, define $\mathcal{P}$ to be the $(n+1)$-voter profile

$$\begin{array}{rl} \frac{n}{2}: & A_1 \succ A_2 \succ A_3 \succ \cdots \succ A_n \\ \frac{n}{2}+1: & A_2 \succ A_1 \succ A_3 \succ \cdots \succ A_n \end{array}$$

(and when n is odd, replace $\frac{n}{2}$ by $\frac{n+1}{2}$). A_2 is the Condorcet-winner and thus the rule's winner. Take $\mathcal{C}$ to be the Condorcet-component

$$\begin{array}{ll} 1: & A_1 \succ A_2 \succ A_3 \succ \cdots \succ A_n \\ 1: & A_2 \succ A_3 \succ \cdots \succ A_n \succ A_1 \\ \vdots & \vdots \\ 1: & A_n \succ A_1 \succ A_2 \succ \cdots \succ A_{(n-1)}. \end{array}$$

In the $(2n+1)$-voter profile $\mathcal{P}+\mathcal{C}$, the number of voters who prefer A_1 to A_k versus A_k to A_1 are (for n even)

$$\begin{array}{rlrl} (\frac{n}{2})+(n-1): & A_1 \succ A_2 & (\frac{n}{2}+1)+1: & A_2 \succ A_1 \\ (n+1)+(n-2): & A_1 \succ A_3 & 2: & A_3 \succ A_1 \\ & \vdots & & \vdots \\ (n+1)+1: & A_1 \succ A_n & n-1: & A_n \succ A_1. \end{array}$$

A_1 is preferred more often against each opponent and thus is the Condorcet-winner and therefore the rule's winner (when n is odd, $\mathcal{P}+\mathcal{C}$ is a $(2n+2)$-voter profile, and it is easy to check that A_1 is the rule's winner). The rule does not cancel properly. ∎

Join-consistency says that if a candidate A is the winner with a rule in each of two separate electorates, then A must be the winner with that rule in the united electorate. Another objection to Condorcet-winners is the following.

Theorem *There is no choice rule consistent with Condorcet that is join-consistent.* (Due to Young; see to H. Moulin 1988, 237–238.)

Proof The argument follows that given in H. Moulin (1988). Take any join-consistent rule that is consistent with Condorcet, and suppose there are $n \geq 3$ voters. Choose a preference-profile $\mathcal{P}$ with n voters that has no Condorcet-winner for which the rule makes A the winner. This implies there must be another candidate B who is preferred to A by k voters, where $k > \frac{1}{2}n$. Letting Q be the preference-profile

$$2k : A \succ B \succ C \succ \cdots \qquad n : B \succ A \succ C \cdots,$$

consider the preference-profile $2P + Q$ of $3n + 2k$ voters. Since the rule is join-consistent, A must be elected by $2\mathcal{P}$. A is the (unique) Condorcet-winner of Q, so the join-consistent rule uniquely elects A. However, of the $2k + 3n$ voters, $2k + n$ prefer B to A, so more prefer B to A. Moreover, at least $2k + n$ prefer B to any other candidate, so B is the Condorcet-winner, a contradiction. ■

Again—now with regard to join-consistency—there is an opposition between sum-scoring methods and rules that are consistent with Condorcet.

For any profile of preferences, a candidate A is preferred to each other candidate B by a certain number of voters, $v(A \succ B) \geq 0$. A's *Simpson-score* is the minimum of those numbers, $s(A) = \min_X v(A \succ X)$. The *Simpson-winner* is the candidate with the highest Simpson-score. *Simpson's method* is clearly consistent with Condorcet. Apply it to the fifteen-voter profile

$$\begin{array}{ll} 3 : A \succ D \succ C \succ B & 3 : A \succ D \succ B \succ C \\ 5 : D \succ B \succ C \succ A & 4 : B \succ C \succ A \succ D. \end{array}$$

Then

$$\begin{array}{l} s(A) = \min\left\{v(A \succ B), v(A \succ C), v(A \succ D)\right\} = \min\{6, 6, 10\} = 6, \\ s(B) = 4, \qquad s(C) = 3, \qquad s(D) = 5, \end{array}$$

so A is the Simpson-winner, though he is not a Condorcet-winner.

If four additional voters with the same preferences

$$4 : C \succ A \succ B \succ D$$

participated in the vote, all preferring A to B, then the nineteen-voter Simpson-winner is B. This is the well-known *no-show paradox*: these voters would have done better for themselves not to vote.

Suppose a choice rule selects candidate A as a winner. Then the rule is *participant-consistent* if when an additional voter casts his ballot, either A is a winner or a candidate preferred to A by that voter is a winner. Participant-consistency is essentially a special case of join-consistency where one part of the electorate consists of one voter. Once again, the Condorcet approach fails.

Theorem *There is no choice rule consistent with Condorcet that is participant-consistent when there are at least four candidates.*[3] (H. Moulin 1988, 238–239, 251).

Proof The proof is slick. It uses the fifteen- and nineteen-voter profiles just given. Suppose there is such a rule. Take a profile $\mathcal{P}$ of n voters with a pair of candidates A and B for which $s(B) \geq v(A \succ B) + 1$ and $n - 2s(B) + 1 > 0$ (the fifteen-voter profile shows such pairs exist). Consider the augmented profile

$$\mathcal{P} \qquad n - 2s(B) + 1 : A \succ B \succ \cdots$$

For this profile B is the Condorcet-winner. There are $2n - 2s(B) + 1$ voters in all, so it suffices to show that B is preferred to any other candidate by at least $n - s(B) + 1$ voters. From $s(B) \geq v(A \succ B) + 1$ it follows that $s(B) + v(B \succ A) \geq n + 1$, or $v(B \succ A) \geq n - s(B) + 1$, so B has a majority against A. B is preferred to any other candidate C by $v(B \succ C) + n - 2s(B) + 1 \geq n - s(B) + 1$ voters since $v(B \succ C) \geq s(B)$. But then A cannot be a winner in $\mathcal{P}$ because participant-consistency would imply that A is the winner in the augmented profile. This proves the following lemma.

Lemma *If $s(B) \geq v(A \succ B) + 1$ and $n - 2s(B) + 1 > 0$, then A cannot be a winner with a participant-consistent rule that is consistent with Condorcet.*

As a consequence, A is the only possible winner with a rule that satisfies the two properties in the fifteen-voter example because $s(A)$ is greater than any other candidate's Simpson-score. Adjoin the four voters with identical preferences, all of whom prefer A to B, to obtain the nineteen-voter profile. In that profile, $s(A) = 6, s(B) = 8, s(C) = 7$, and $s(D) = 5$. The same lemma implies that the only possible winner is B, a contradiction. ■

Proper cancellation, join-consistency, and participant-consistency are three closely related properties that argue against Condorcet-winners. Each in essence characterizes sum-scoring methods, so it can be concluded that there is a fundamental opposition between Condorcet-winners and sum-scoring-winners in the traditional model.

4.4 Borda-Winners and Condorcet-Rankings

Putting aside for the moment that the Borda-*ranking* is wide open to strategic manipulation, is it nevertheless a good method for designating a *winner*? In that

3. Simpson's method is participation-consistent with three candidates.

case, declaring the Borda-score winner in first place, the Borda-score winner among the remaining candidates in second place, and so on, another ranking—instant-winner-Borda-runoff—is defined. Is this a good ranking? It is sure to rank a Condorcet-loser last, and it seems better suited to designating a winner. But all three orders—Saari's instant-(loser-)Borda-runoff ranking, the instant-winner-Borda-runoff ranking, and the Borda-ranking—may differ.

Is the Condorcet-*ranking* a good method for designating a *winner*? In that case, declaring the Condorcet-score winner in first place, the Condorcet-score winner among the remaining candidates in second place, and so on, another ranking—instant-winner-Condorcet-runoff—is defined. Or one could declare the Condorcet-score loser in last place, the Condorcet-score loser among the remaining candidates in next-to-last place, and so on, to obtain the instant-loser-Condorcet-runoff ranking. Are these good rankings? Nothing at all changes; these two rankings both coincide with the Condorcet-ranking (in this sense, Condorcet's method almost satisfies IIA).

To see why suppose, more generally, that the first several candidates and the last several candidates were dropped and that the Condorcet-ranking among the remaining candidates changed. This implies that the change would have increased the total count among the remaining candidates; but then the same change would increase the total count by the same amount in the original order, a contradiction. This is reassuring. On the other hand, the Condorcet-ranking *always* ranks the Condorcet-winner W first (as does instant-loser-Borda-runoff, when W exists) and always ranks the Condorcet-loser L last (as does instant-winner-Borda-runoff, when L exists), which is unacceptable for selecting either a winner or a loser, as Young and Saari have argued.

The contrast between Borda's and Condorcet's approaches is stark. The method of Borda is more appropriate for designating winners and losers, Condorcet's more appropriate for finding a best ranking among the candidates. Very simple new characterizations of them makes this point even clearer.

Given any preference-profile, a *candidate-scoring method* assigns a non-negative score to every candidate. A candidate-scoring method should (1) *assign a zero to the worst possible candidate*, that is, give a score of 0 to a candidate who is last on every voter's list; and (2) *correctly reward a minimal improvement*, that is, when one voter inverts two successive candidates of his list, the score of the candidate who has placed higher increases by 1.

Theorem 4.1 (Borda Characterization) *The Borda-score is the unique candidate-scoring method that assigns a zero to the worst possible candidate and correctly rewards minimal improvements.*

Proof This is easy to see. The Borda-score clearly satisfies the two properties. On the other hand, the argument that follows shows that if the properties hold, then the score of any candidate must agree with the candidate's Borda-score. Consider any preference-profile, and choose any one candidate C. Put C at the bottom of every voter's list. C's score must now be 0. Take one voter and raise C back to his previous position one step at a time. At each step he gains a point, so if he rises k times, he gains k points from this voter. But this is precisely the number of points contributed by this voter to the Borda-score of C, so doing the same for every voter proves the theorem. ∎

An immediate consequence of theorem 4.1 is that if a candidate A moves up (moves down) in one or several voters' rankings, then A's Borda-score increases (decreases), and the scores of all other candidates either remain the same or decrease (increase), so A cannot be ranked lower (higher) in the Borda-ranking. Thus the Borda-ranking is *choice-monotone*.

Choice-monotonic: If A ranks at least as high as B, that is, $A \succeq_S B$, and either A moves strictly higher or B moves strictly lower in some one or more voters' rankings, then society ranks A strictly ahead of B, that is, $A \succ_S B$.

Given any preference-profile, a *rank-scoring method* assigns a non-negative score to every ranking. Given any ranking $A \succ B \succ C \succ D \succ \ldots$, its *opposite* ranking is $A \prec B \prec C \prec D \prec \ldots$. A rank-scoring method should (1) *assign a zero to the worst possible ranking*, that is, give a score of 0 to a ranking if every voter's preference is the opposite ranking; and (2) *correctly reward a minimal improvement*, that is, when one voter inverts two successive candidates of his list, the score of every order that agrees with the change increases by 1.

Theorem 4.2 (Condorcet Characterization) *The Condorcet-score is the unique rank-scoring method that assigns a zero to the worst possible order and correctly rewards minimal improvements.*

Proof The Condorcet-score clearly satisfies the two properties. The following argument shows that any rank-scoring method that satisfies the two properties must give a score to a ranking that is equal to its Condorcet-score.

Consider an arbitrary ranking $\mathcal{R} = A \succ B \succ C \succ D \succ \ldots$, and a particular voter v whose preference-order is $\mathcal{R}_v$. If v's preference $\mathcal{R}_v$ is the opposite of $\mathcal{R}$, then she contributes 0 to the score of $\mathcal{R}$, exactly the same as her contribution to the Condorcet-score of $\mathcal{R}$. Make the inductive assumption that when a voter contributes k to the Condorcet-score of $\mathcal{R}_v$ she makes an equal contribution to the score of $\mathcal{R}$, and consider a voter's preference $\mathcal{R}_v$ that contributes $k+1$ to the Condorcet-score of $\mathcal{R}$. This means that $k+1$ pairs of candidates are

ordered in $\mathcal{R}_v$ as they are in $\mathcal{R}$. Check to see if the bottom pair of candidates in $\mathcal{R}_v$ is ordered in accordance with its order in $\mathcal{R}$; if it is not, check the next pair of successive candidates, and so on, until a pair is found that is ordered in accordance with its order in $\mathcal{R}$. There must be such a pair because $\mathcal{R}_v$ contributes $k+1 > 0$ to the Cordorcet-count of $\mathcal{R}$. Invert the order of that pair in $\mathcal{R}_v$; that must decrease voter v's contribution to the Condorcet-score by 1. By the induction hypothesis, the altered $\mathcal{R}_v$ contributes k to the score of $\mathcal{R}$, and by property (2), reestablishing the order of $\mathcal{R}_v$ adds 1 to the score of $\mathcal{R}$, showing that the contribution of $\mathcal{R}_v$ to the score and to the Condorcet-score are exactly the same. ■

An immediate consequence of theorem 4.2 is that when the Condorcet-ranking $\mathcal{R}$ has A in first place, and A moves up in some one or several voters' rankings, then A remains in first place. It suffices to see that this is true if A moves up one place in some one voter's ordering. When this occurs, the Condorcet-score of $\mathcal{R}$ increases by 1, whereas the Condorcet-score of every other ranking either increases or decreases by 1, so $\mathcal{R}$ must still have the highest score. This in fact shows more, namely, that the Condorcet-ranking is *rank-monotonic*:

Rank-monotonic: If $\mathcal{R}$ is the ranking of society whose first-place candidate is A, then $\mathcal{R}$ is the ranking of society when A moves up in one or more voters' rankings.

This is not true of the Borda-ranking. The analogous property holds for the last-place candidate as well. For a detailed discussion of the many different concepts of monotonicity, see Nurmi (2004).

A large class of methods is similar to Borda's in that each assigns scores to candidates. Given a preference-profile $\mathcal{P}$, a *value choice rule* assigns to each of the n candidates a real number $f(\mathcal{P}) = (v_1, \ldots, v_n)$, and elects the candidate(s) k having the largest number v_k. Consider two separate societies having preference-profiles $\mathcal{P}_1$ and $\mathcal{P}_2$. A value choice rule f is *sum-consistent* if $f(\mathcal{P}_1 + \mathcal{P}_2) = f(\mathcal{P}_1) + f(\mathcal{P}_2)$ (where again $\mathcal{P}_1 + \mathcal{P}_2$ means the two profiles taken together as one profile), that is, a candidate's number for the joint profile $\mathcal{P}_1 + \mathcal{P}_2$ is the sum of his numbers in $\mathcal{P}_1$ and in $\mathcal{P}_2$. It is straightforward to verify the theorem that the unique impartial and sum-consistent value choice functions are the sum-scoring functions. Among them Borda's is the one that satisfies the cancellation property.

Another large class of methods is similar to Condorcet's in that each assigns scores to rankings. They do not seem to have been studied at all, let alone defined. Given a preference-profile $\mathcal{P}$, a *value ranking rule* assigns to each

ranking a real number and selects the ranking(s) having the largest real number. Sum-consistency may be defined for value ranking functions in the same way as it is for value choice functions, and an analogous theorem characterizes those whose values equal sums of values over the rankings of voters.

The arguments of this section all suggest that Borda-*rankings* and Condorcet-*winners* should be discarded. The properties imply that Borda's approach makes sense only for designating a *winner* and Condorcet's only for designating a *ranking*.

4.5 Incompatibility of Electing and Ranking

All of this naturally leads to a fundamental question. Is a voting procedure supposed to *elect a winner* (or a *loser*) or to *rank all candidates*? It has usually been implicitly assumed (with the notable exception of Young) that it should do both. Ranking procedures are routinely used to find the winner either by taking the first-place candidate or by iteratively finding and eliminating the last-place candidates until one candidate remains. Symmetrically, given a method to elect a winner, all candidates are ranked by taking the winner as the first, then applying the same method to designate a winner among the remaining candidates, and repeating until a complete ranking is obtained. Or, given a method to designate a loser, all candidates are ranked by taking the loser as the last, reapplying the same method to designate a loser among the remaining candidates, and repeating to obtain a complete ranking.

Imagine a sports competition where the first-ranked candidate is not the winner! It is surely reasonable to believe that ranking and designating winners (or losers) *must be* two sides of one coin. But are they, in the traditional model?

Consider a ranking rule, that is, any method f that amalgamates a profile—an arbitrary set of preference-orders of voters—into one or several preference-orders of society, where the first-place candidate is declared the winner and the last-place candidate the loser. To be practical it should possess each of the following three properties. (1) It must be *winner-loser-unanimous*: whenever all voters rank a candidate first (last), she is the winner (loser). (2) It must be *choice-compatible*: if all the voters rank a candidate first (last) and a Condorcet-component is added to the profile, then that candidate must be the winner (loser). (3) It must be *rank-compatible*: if a winner is removed from the set of candidates, then the new ranking of the remaining candidates agrees with the original ranking.[4] Notice that Borda's method (indeed, any scoring method)

4. Rank-compatibility is a weaker form of "local stability" defined by Young (1988).

satisfies conditions (1) and (2) but not (3), whereas Condorcet's satisfies (1) and (3) but not (2).

Theorem 4.3 (Incompatibility) *There is no ranking rule that is winner-loser-unanimous, choice compatible, and rank-compatible for all preference-profiles (when there are at least three alternatives).*

Proof The proof is disarmingly simple. Assume there is such a function f and consider the $(3p+q)$-voter profile with $p > q$

$$p : A_1 \succ \cdots \succ A_k \succ A \succ B \succ C \qquad p : A_1 \succ \cdots \succ A_k \succ B \succ C \succ A$$
$$p : A_1 \succ \cdots \succ A_k \succ C \succ A \succ B \qquad q : A_1 \succ \cdots \succ A_k \succ A \succ C \succ B,$$

and suppose the method f yields a ranking $\mathcal{R}$. Candidate A_1 is placed first by all voters, so by winner-loser unanimity, A_1 must be the winner and hence the first-place candidate of $\mathcal{R}$. Rank-compatibility implies that when A_1 is removed, f applied to the profile that remains must yield a ranking that is the same as $\mathcal{R}$ (except that A_1 is absent). Repeating the same reasoning, successively removing $A_2, \ldots,$ down to A_k shows that the first k candidates of the ranking $\mathcal{R}$ are $A_1 \succ_S \cdots \succ_S A_k$ and its last three candidates are ranked by f applied to the reduced profile

$$p : A \succ B \succ C \qquad p : B \succ C \succ A \qquad p : C \succ A \succ B \qquad q : A \succ C \succ B.$$

The reduced profile contains a $3p$-voter Condorcet-component, and the remaining voters unanimously place A first and B last. By winner-loser-unanimity and choice-compatibility this implies that the $(3p+q)$-voter profile does the same, so f yields the ranking $A \succ_S C \succ_S B$. Therefore $\mathcal{R} = A_1 \succ_S \cdots \succ_S A_k \succ_S A \succ_S C \succ_S B$.

Rank-compatibility now implies that when A_1 is dropped from the profile, f yields the remaining part of $\mathcal{R}$; when the winner of that part A_2 is dropped from the profile, f yields the remaining part of $\mathcal{R}$; $\ldots$, down to dropping A from the profile. But without $A_1, \ldots, A_k, A$, the profile is

$$p - q : B \succ C \qquad p + q : B \succ C \qquad p + q : C \succ B.$$

The last $2p + 2q$ voters constitute a Condorcet-component, so by choice-compatibility and unanimity, f yields the ranking $B \succ_S C$, a contradiction that completes the proof. ∎

A similar argument shows that losers instead of winners could be invoked in the definition of rank-compatibility.

The three conditions of this theorem do not demand much and are unassailable in any practical application. The first merely requires that complete agreement

on a winner or loser must be confirmed in the final standings. The second asks considerably less than Young's and Saari's cancellation for two reasons: (1) only one Condorcet-component is canceled; and (2) a Condorcet-component is canceled and a decision is assumed only if the remaining voters unanimously agree on a winner or loser (which is much less demanding than to assume a decision whatever the remaining voters' orders). The third is weaker than Arrow's independence of irrelevant alternatives property, for it asks that the ranking stay the same among the remaining candidates only when the first-place candidate is withdrawn, not when any candidate is withdrawn. The theorem tells us that runoffs—dropping a winner or a loser and then ranking the remaining candidates—are doomed to failure. As Saari realized, reranking dismisses information that is essential. More fundamentally, the theorem says that *in the context of the traditional model it is impossible to assert that the first-place candidate of the ranking of an electorate or a jury is necessarily the winner*. This is damning testimony against the very validity of that model.

More evidence may be given to show that there is a fundamental difference between the problem of choice and the problem of ranking. To begin, consider a situation where the profile of a society is k-*Cond*$(A \succ B \succ C)$:

$$k : A \succ B \succ C \qquad k : B \succ C \succ A \qquad k : C \succ A \succ B.$$

The $3k$ voters constitute a Condorcet-component. It is perfectly reasonable to conclude that *every* candidate is tied for first. But what is society's *ranking*? Borda's and every other scoring method affirm that *every* ranking is tied for first and every other place (in the ranking of rankings), but is that reasonable? The number of Condorcet-points given to a ranking $\mathcal{R}$ by a voter is a measure of how much he agrees with it: it is the number of times the voter's preference-order between pairs of candidates agrees with the order of pairs in $\mathcal{R}$'s.

In table 4.2 the rankings provide a measure of how much society likes $A \succ B \succ C$ and $A \succ C \succ B$. In the first case one-third of society agrees on all three comparisons and two-thirds agree on one comparison, whereas in the second case two-thirds agree on two comparisons and one-third agree on none. Any impartial method must view the three actual rankings of the voters as equally good as the first among them, $\mathcal{R}_1$, and the same is true for the other three rankings, $\mathcal{R}_2 = A \succ C \succ B$ and $C \succ B \succ A$ and $B \succ A \succ C$. Table 4.2 gives the Condorcet-points and Condorcet-scores for all six possible rankings: there is a three-way tie for first, and a three-way tie for fourth.

The Condorcet-score singles out $\mathcal{R}_1$ and the other two rankings with which it forms a Condorcet-cyle as the best, but that is disputable. In the face of $\mathcal{R}_1 = A \succ B \succ C$, two-thirds of the voters (those having the other two preference-orders), a solid majority, would prefer a preference-order of the opposite

Table 4.2
Condorcet's Solution for the Profile $k\text{-}Cond(A \succ B \succ C)$

	Condorcet-Points				Condorcet-Score	Ranks of Rankings
	3	*2*	*1*	*0*		
$\mathcal{R}_1 : A \succ_S B \succ_S C$	k	0	$2k$	0	$5k$	1st
$\mathcal{R}_2 : A \succ_S C \succ_S B$	0	$2k$	0	k	$4k$	4th

Note: Three rankings ($A \succ_S B \succ_S C$, $B \succ_S C \succ_S A$, and $C \succ_S A \succ_S B$) are tied for 1st; the other three are tied, too.

Table 4.3
Condorcet's Solution for the Profile $k\text{-}Cond(A \succ B \succ C \succ D)$ of $4k$ Voters

	Condorcet-Points							Condorcet-Score	Ranks of Rankings
	6	*5*	*4*	*3*	*2*	*1*	*0*		
$\mathcal{R}_1 : A \succ_S B \succ_S C \succ_S D$	k	0	0	$2k$	k	0	0	$14k$	1st
$\mathcal{R}_2 : A \succ_S B \succ_S D \succ_S C$	0	k	k	0	k	k	0	$12k$	5th
$\mathcal{R}_3 : A \succ_S C \succ_S D \succ_S B$	0	0	$2k$	k	0	k	0	$12k$	5th
$\mathcal{R}_4 : A \succ_S D \succ_S B \succ_S C$	0	k	k	0	k	k	0	$12k$	5th
$\mathcal{R}_5 : A \succ_S C \succ_S B \succ_S D$	0	k	0	k	$2k$	0	0	$12k$	5th
$\mathcal{R}_6 : A \succ_S D \succ_S C \succ_S B$	0	0	k	$2k$	0	0	k	$10k$	21st

Note: Four are tied for 1st; sixteen for 5th; and four for 21st.

set, namely, $C \succ B \succ A$. It is not clear which of the two triplets—$\mathcal{R}_1$ and its Condorcet-cycle or $\mathcal{R}_2$ and its Condorcet-cycle—is better for society. Condorcet opts for the three that are the voters' preference-orders; another rule might well opt for the "compromise" solution, the other three.

Take now the preference-profile $k\text{-}Cond(A \succ B \succ C \succ D)$ to be the Condorcet-component of $4k$ voters with four candidates:

$$k : A \succ B \succ C \succ D \qquad k : B \succ C \succ D \succ A$$
$$k : C \succ D \succ A \succ B \qquad k : D \succ A \succ B \succ C.$$

It shows even more dramatically how very different ranking is from choice.

Each $\mathcal{R}_i$ in table 4.3 represents four preference-orders that must have identical Condorcet-points (and Condorcet-scores). By impartiality, the three preference-orders that define a Condorcet-component with $\mathcal{R}_1$ have the same Condorcet-points as $\mathcal{R}_1$; the same is true for $\mathcal{R}_6$. $BCAD$ (dropping the $\succ$ for brevity), $CDBA$, and $DACB$ have the same Condorcet-points as $\mathcal{R}_2$; $BDAC$, $CABD$, and $DBCA$ have the same as $\mathcal{R}_3$; $BACD$, $CBDA$, and $DCAB$ have the same as $\mathcal{R}_4$; and $BDCA$, $CADB$, and $DBAC$ have the same as $\mathcal{R}_5$.

Some rankings are clearly better for society than others. $\mathcal{R}_1$ *dominates* $\mathcal{R}_5$ and $\mathcal{R}_6$ because for any $h = 0, \ldots, 6$, the number of voters who agree on at least h orders between pairs of candidates is always at least as great, and sometimes greater, for $\mathcal{R}_1$ than for $\mathcal{R}_5$ or $\mathcal{R}_6$ (in mathematical terminology, this is stochastic dominance). At least as many voters are in at least as great agreement with $\mathcal{R}_1$ (and either some "at least" is "more" or some "as great" is "greater") than with either $\mathcal{R}_5$ or $\mathcal{R}_6$. This is confirmed by their respective Condorcet-scores (as it must be by any reasonable criterion of comparison). The comparison between, say, $\mathcal{R}_1$ and $\mathcal{R}_2$ is not evident and depends on the criterion invoked. On the other hand, every candidate must be tied for first by any reasonable method. Moreover, all twenty-four possible rank-orderings have exactly the same Borda-score, showing how inadequate Borda's method is for ranking.

The order given by the Condorcet-scores to the various rankings of candidates is very curious. First, note that the Condorcet-scores of $A \succ B \succ D \succ C$ and its opposite $C \succ D \succ B \succ A$ are exactly the same. Next, the Condorcet-scores rank the rankings not in accord with choice-monotonicity. For consider the ranking $A \succ B \succ D \succ C$ with score $12k$, which is ahead of the ranking $D \succ C \succ B \succ A$ with score $10k$. Suppose that the preference-profile k-*Cond* $(A \succ B \succ C \succ D)$ changed, with $2k$ of the voters who place A ahead of B inverting them in their rankings. Then after the change $A \succ B \succ D \succ C$ has the score $10k$ and so is behind $D \succ C \succ B \succ A$, whose score is $12k$. But B was ahead of D and C in the preferred ranking before and was moved strictly higher by $2k$ voters, so choice-monotonicity suggests B should be ahead of C and D after the change in the preferred ranking, but it is not.

Now add one voter to k-*Cond* $(A \succ B \succ C \succ D)$ with preference-order $A \succ D \succ C \succ B$ to obtain the $(4k + 1)$-voter profile

$$\{A \succ D \succ C \succ B\} + k\text{-}Cond(A \succ B \succ C \succ D).$$

$\mathcal{R}_1$ and the other three rank-orders with which it defines a Condorcet-cycle each have Condorcet-score $14k$. All other rank-orders have Condorcet-scores at most $12k$. Since at most six Condorcet-points can be added to any ranking, one of those first four must be the Condorcet-ranking when $k > 3$. Among the first four, the rank-order $D \succ A \succ B \succ C$ has the highest Condorcet-score, so it is the Condorcet-ranking, as may be seen in table 4.4. On the other hand, it is clear that A should be the winner and B the loser by any reasonable choice function. Once again the traditional model leads to an incompatibility between winners and rankings.

Table 4.4
The Four Rankings with Highest Condorcet-Scores for the $(4k+1)$ Voter Profile $\{A \succ D \succ C \succ B\} + k\text{-}Cond\,(A \succ B \succ C \succ D)$ when $k > 3$

	Condorcet-Points							Condorcet-Score	Ranks of Rankings
	6	*5*	*4*	*3*	*2*	*1*	*0*		
$\mathcal{R}_1^4 : D \succ_S A \succ_S B \succ_S C$	k	0	1	$2k$	k	0	0	$14k+4$	1st
$\mathcal{R}_1^3 : C \succ_S D \succ_S A \succ_S B$	k	0	0	$2k+1$	k	0	0	$14k+3$	2d
$\mathcal{R}_1^1 : A \succ_S B \succ_S C \succ_S D$	k	0	0	$2k+1$	k	0	0	$14k+3$	2d
$\mathcal{R}_1^2 : B \succ_S C \succ_S D \succ_S A$	k	0	0	$2k$	k	0	1	$14k$	4th

Theorem 4.4 (Monotonic Incompatibility) *There is no ranking function that is impartial, unanimous, rank-monotonic and choice-monotonic (when there are at least three candidates).*

Proof To see the truth of this statement, assume there is such a ranking function and consider the profile

$$p : A \succ B \succ C \succ A_1 \succ \cdots \succ A_k \qquad p : B \succ C \succ A \succ A_1 \succ \cdots \succ A_k$$
$$p : C \succ A \succ B \succ A_1 \succ \cdots \succ A_k \qquad q : A \succ C \succ B \succ A_1 \succ \cdots \succ A_k$$
$$q : C \succ B \succ A \succ A_1 \succ \cdots \succ A_k \qquad q : B \succ A \succ C \succ A_1 \succ \cdots \succ A_k.$$

When A is replaced by B, B by C, and C by A, the result is exactly the same profile. Therefore, impartiality and unanimity imply that society's order is $A \approx_S B \approx_S C \succ_S A_1 \succ_S \cdots \succ_S A_k$.

Suppose that one voter with ranking $C \succ A \succ B \succ A_1 \succ \cdots \succ A_k$ exchanges C and A. Choice-monotonicity (and unanimity) then implies that society's ranking must be $A \succ_S B \succ_S C \succ_S A_1 \succ_S \cdots \succ_S A_k$.

Move A up to the first position in every voters' list to obtain the profile $\mathcal{P}$:

$$2p+q : A \succ B \succ C \succ A_1 \succ \cdots \succ A_k \qquad p+2q : A \succ C \succ B \succ A_1 \succ \cdots \succ A_k.$$

Rank-monotonicity implies society's ranking remains the same, $A \succ_S B \succ_S C \succ_S A_1 \succ_S \cdots \succ_S A_k$. If $p = q$, impartiality implies that $A \succ_S C \approx_S B \succ_S A_1 \succ_S \cdots \succ_S A_k$, a contradiction (which suffices to prove the theorem). ■

But more is true. If $p > q$ and $p+q$ is even (or if the model is continuous, as is the Dasgupta-Maskin model), one may consider the profile

$$\frac{3p+3q}{2} : A \succ B \succ C \succ A_1 \succ \cdots \succ A_k$$
$$\frac{3p+3q}{2} : A \succ C \succ B \succ A_1 \succ \cdots \succ A_k,$$

for which society's ranking is $A \succ_S C \approx_S B \succ_S A_1 \succ_S \cdots \succ_S A_k$ by impartiality. When $\frac{p-q}{2}$ voters with the second ranking exchange B and C to obtain the profile $\mathcal{P}$, choice-monotonicity implies that society's decision for $\mathcal{P}$ is $A \succ_S B \succ_S C \succ_S A_1 \succ_S \cdots \succ_S A_k$, again a contradiction.

It should be noticed that impartiality is a weaker property than the cancellation property used by Young and Saari: impartiality alone does not permit one to deduce from the profile

$$2\text{-}Cond(A \succ B \succ C \succ D) + 1\text{-}Cond(A \succ C \succ B \succ D)$$

that all the candidates are equivalent.

Suppose Condorcet's method was to be used in order to select one ranking of the candidates in the spirit of Arrow, including the possibility of ties between pairs of successive candidates—that is, to define a ranking function. Then a rule would have to be given to select one ranking whenever there was a tie among several Condorcet-rankings. The theorem in this case says that either the ranking function is not impartial or it is not choice-monotonic (since by definition the ranking function is rank-monotonic).

When there are fewer than five candidates, the Condorcet-ranking is choice-monotonic (when unique). When there are five or more candidates, it is not. The following six-candidate example is due to Andrew Jennings. For the twenty-four-voter profile ($\succ$ is omitted): 10: $ABCDEF$, 8: $BFDEAC$, 1: $ECFABD$, 5: $EFCABD$, the unique Condorcet-ranking is $ABCDEF$. If the eight voters with the second rank-order change by moving C up one place and dropping A to last place, the unique Condorcet-ranking is $BEFCAD$. Before the change $C \succ_S E$ and $C \succ_S F$; after C moves up, $E \succ_S C$ and $F \succ_S C$. For a five-candidate example see Balinski, Jennings, and Laraki (2008).

4.6 Preferences over Rank-Orders

The traditional theory considers inputs and outputs of different types.

In what might be called the Arrow approach, voters' inputs are rankings of candidates, the output is a winner (sometimes obtained by ranking all the candidates and designating the first-place candidate the winner). First-past-the-post, the alternative vote, Borda's, and most of the methods that have been discussed are of this type.

In what might be called the Condorcet approach, voters' inputs are rankings of candidates, and the output is a ranking of the candidates (obtained by ranking all the possible rankings of candidates).

The natural Arrow approach to the problem posed by Condorcet—find the best ranking of candidates—would be for voters' inputs to be rankings of the

rankings of candidates, the output to be a ranking of the candidates. Might such a formulation escape Arrow's conundrum?

Arrow imposed an inner consistency on the preferences of each voter—namely, that the voter's preference-order be transitive and strict—and wished to find a social welfare function obeying several desirable properties. Of course, the more stringent the constraints on the preferences of voters, the more difficult it is to deduce an impossibility.

Suppose there are n competitors and that every voter has a strict, transitive preference-order over the $n!$ rank-orders of the n competitors; the $(n-1)!$ rank-orders of every subset of $n-1$ competitors (since a competitor could withdraw); ... ; the $(n-k)!$ rank-orders of every subset of $n-k$ competitors for $k = 0, 1, \ldots, n-2$ (since k competitors could withdraw). Thus $\sum_{k=0}^{n-2} C_k^n (n-k)! = n! \sum_{k=0}^{n-2} \frac{1}{k!} \approx en!$ preferences are expressed by each voter. When $\mathcal{R} = \{A_1 \succ A_2 \succ \cdots \succ A_s\}$ and $\mathcal{R}' = \{A_{\sigma(1)} \succ A_{\sigma(2)} \succ \cdots \succ A_{\sigma(s)}\}$ are two rank-orders on the same subset of competitors (so that σ is a permutation of $1, 2, \ldots, s$), $\mathcal{R} \gg \mathcal{R}'$ means $\mathcal{R}$ is preferred to $\mathcal{R}'$.

The voter's preferences should satisfy some further minimal conditions. A voter's preferences are *inner-consistent* if they satisfy two properties:

- If

$$\{A_1 \succ \cdots \succ A_s\} \gg \{A_{\sigma(1)} \succ \cdots \succ A_{\sigma(s)}\}$$

and A is not in their common set of competitors, then

$$\{A_1 \succ \cdots \succ \overbrace{A}^{k^{th}} \succ \cdots \succ A_s\} \gg \{A_{\sigma(1)} \succ \cdots \succ \overbrace{A}^{k^{th}} \succ \cdots \succ A_{\sigma(s)}\},$$

that is, if A is inserted in the kth place (for $k = 1, \ldots, s+1$) in both rank-orders, the order between the two new rank-orders conforms with that of the first two.

- There is no Condorcet-cycle among the rank-orders of pairs of competitors:

$$\{A \succ B\} \gg \{B \succ A\} \quad \text{and} \quad \{B \succ C\} \gg \{C \succ B\} \quad \text{implies}$$
$$\{A \succ C\} \gg \{C \succ A\},$$

for every three competitors. It will be seen anon that these conditions can be satisfied.

The input in this model is a profile that gives the voters' preferences over the rank-orders on every subset of competitors; the output is a rank-order of society. The natural counterparts to Arrow's conditions (see chapter 3) are as follows:

(1) There must exist a solution for every possible set of inner-consistent preferences over rank-orders. ("unrestricted domain") (2) When every voter has the same preferences, the common preferred rank-order on all n competitors is society's rank-order ("unanimity"). (3) If a competitor withdraws, society's rank-order among the remaining competitors remains the same ("IIA"). (4) No one voter's preferences over rank-orders can always determine society's rank-order, whatever the preferences of all the other voters ("non-dictatorial").

Theorem 4.5 (Impossibility) *There is no method of amalgamating the preferences over rank-orders into a rank-order of society that satisfies the four conditions (when there are at least three candidates).*

Proof It is shown that there is no method for a particular profile of inner-consistent preferences because of Arrow's theorem; hence there cannot be a method for an arbitrary profile of inner-consistent preferences.

Name the n competitors $A_1, A_2, \ldots, A_n$ so that voter v's preferred rank-order over all n competitors is $A_n \succ A_{n-1} \succ \cdots \succ A_1$. Voter v's *code* is defined by $v(A_k) = k$ for $k = 1, 2, \ldots, n$. An order is defined recursively on the codes, which will in turn determine v's preference between any two rank-orders with a common set of competitors.

First, $(j_1, j_2) \gg_c (j_2, j_1)$ if $j_1 > j_2$. This simply says that the order $\{A_{j_1} \succ A_{j_2}\}$ is preferred to the order $\{A_{j_2} \succ A_{j_1}\}$ if $j_1 > j_2$.

Next, $\left(j_1, j_2, \ldots, j_s\right) \gg_c \left(i_1, i_2, \ldots, i_s\right)$, where the i_k's are a permutation of the j_k's, if either

$$j_k = s \quad \text{and} \quad \max_{1 \le l \le k} i_l < s,$$

or

$$j_k = s = i_k \quad \text{and} \quad (j_1, \ldots, j_{k-1}, j_{k+1}, \ldots, j_s) \gg_c (i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_s).$$

The idea is that a voter's top priority is for his favorite among any subset of candidates to be the highest possible in the ranking. Thus, for example, $(2, 1) \gg_c (1, 2)$ and

$$(3, 2, 1) \gg_c (3, 1, 2) \gg_c (2, 3, 1) \gg_c (1, 3, 2) \gg_c (2, 1, 3) \gg_c (1, 2, 3).$$

The code defines v's preference between any two rank-orders with a common set of competitors by

$$\{A_{j_1} \succ A_{j_2} \succ \cdots \succ A_{j_s}\} \gg \{A_{\sigma(j_1)} \succ A_{\sigma(j_2)} \succ \cdots \succ A_{\sigma(j_s)}\}$$

if

$$\left(j_1, j_2, \ldots, j_s\right) \gg_c \left(\sigma(j_1), (j_2), \ldots, \sigma(j_s)\right),$$

where σ is a permutation of $j_1, j_2, \ldots, j_s$. Call this order the *top-preferred order*. For example, when $n = 4$ and v's preferred rank-order over all four competitors is $A_4 \succ A_3 \succ A_2 \succ A_1$, the top-preferred order determines that

$$\{A_3 \succ A_1 \succ A_4 \succ A_2\} \gg \{A_3 \succ A_2 \succ A_1 \succ A_4\}$$

because $(3, 1, 4, 2) \gg_c (3, 2, 1, 4)$.

Every voter has preferences over rank-orders based in this manner on the voter's preferred rank-order over all n candidates. This is a perfectly rational set of preferences that is clearly inner-consistent.[5] These preferences are transitive over competitors (as demanded by Arrow). Suppose that there was a method for amalgamating these preferences over rank-orders into a rank-order of society that satisfied the four properties. Then there would also be one for Arrow's hypotheses, which is impossible. ■

The idea of asking for voters' preferences over all possible rank-orders is, of course, frivolous. With twelve competitors (as in the French presidential election of 2007), this would ask voters to specify over 479 million preferences (which is frivolous unless, as in the profile defined in the proof, they are generated by a simple input). And yet, this is but the choice problem over rank-orders; the welfare problem over rank-orders would be to specify as an output a preference over all rank-orders. But, in that case, should not the voters be asked to input their preferences over all possible rank-orders of the rank-orders of competitors? And so one can imagine an infinite regress in an endless search to express true preferences. As Alice would undoubtedly have remarked, "Curiouser and curiouser" (Carroll 1916, 6). What this fanciful philosophical flight shows is that the inputs of voters in the traditional model can in no way be regarded as preferences. This is why it is imperative to view the inputs of voters as mere messages.

Whenever a society or jury chooses a rank-ordering, it *inevitably* designates as winner the first-place competitor. Theorem 4.3, the several examples that follow it, and theorem 4.4 prove that the traditional model *cannot reconcile winners and first places*. Conclusion? The traditional model's inputs are inadequate messages and must be reformulated.

5. Inner-consistent preferences are more general: nothing is specified in the first part of their definition about how to compare rank-orders when the competitor A is inserted in different places.

5 Strategy in the Traditional Model

Strategem implies a concealed intention, and therefore is opposed to straightforward dealing, in the same way as wit is the opposite of direct proof.
—Carl von Clausewitz

Pierre-Simon, Marquis de Laplace was a mathematician and astronomer, a founder of the theory of probability ("perhaps the greatest and certainly the most famous physicist of his day" [Kuhn 1961, 196]). For him, "[T]he most important questions of life, are in effect, for the most part, problems of probability. One can even say, in all rigor, that almost all our knowledge is probable; and in the small number of things that we can know with certitude, in the mathematical sciences themselves, the principal means to arrive at the truth, induction and analogy, are based on probabilities, so that the whole system of human knowledge depends on the theory explained in this essay . . . It is remarkable that a science that originated in the study of games has been elevated to the most important subjects of human knowledge" (1820, v, clii–cliii).

Laplace observed "It is difficult to know or even define the will of an assembly in the midst of the variety of opinions of its members," and he imagined an entirely new model and approach. In *Laplace's model* a voter assigns to each candidate a real number score, between a minimum of 0 and some (arbitrary) maximum R, that represents (in the voter's opinion) the candidate's merit, but he only reports the order among their magnitudes, from worst to best. Laplace assumed that the scores given the candidates by a voter are uniformly distributed on the interval $[0, R]$, and he asked, what is the average or expected value of all the voters' lowest scores, of all the next to lowest, . . . , of all the highest? He found that they are proportional to $1, 2, \ldots,$ up to n, where n is the number

of candidates, and thus he gave a justification for the Borda-points (though the uniformity assumption is doubtful).[1]

Laplace went on to conclude that Borda's is the method for finding the candidate of greatest merit: "Such is the method of election indicated by the Theory of Probability" (xcii). In doing so, however, he simply assumed (as have many others) that summing the Borda-points is the evident or only reasonable way of aggregating the evaluations of many voters.

But he realized that Borda's method can only find the candidate of greatest merit if the voters *honestly* report their preference-orders: "This method of election would be without a doubt the best if considerations alien to the merit [of a candidate] did not influence the choice of the electors, even the most honest ones, and did not determine them to rank last the most dangerous opponents to their favorites, which would give a big advantage to candidates of mediocre merit. Moreover, the experience of institutions which adopted it has led them to abandon it" (277; the analysis, 275–279).

Laplace seems to be the first person to have clearly seen the importance of strategic manipulation in voting.[2] Beginning with a new and promising idea that builds a probabilistic argument to justify the Borda-points (and, he thought, the Borda-winner), he ended up supporting a Condorcet-winner, suggesting that an assembly that continues to vote often enough will eventually give to one candidate an absolute majority through sheer exhaustion of the participants (much like the procedure for electing a Pope). However, an election that requires many rounds is suspicious, for the result depends on the information transmitted from round to round, and the ultimate winner may have little to do with the original preferences of the voters (yet everything to do with one or two strong personalities among the voters).

Laplace saw a difference between electing a candidate and voting on a provision concerning some common end such as a budget. In the latter case, he assumed, there is a common search for the one correct decision, and each voter honestly assigns to alternative motions the probabilities he believes are those with which they should be chosen. But, again, the voters only report their preferences. Laplace assumed that the probabilities assigned to the alternative motions by a voter are uniformly distributed. He then used the same analysis as he did in voting for candidates, calculating the average of all the lowest

1. Grades that are very near perfection are scarce, as are often though not always grades that are extremely low. The Orsay experiment and recent experiences in grading wines confirm this (see chapters 14 and 21).

2. Farquharson (1969) traces the study of strategic manipulation to Pliny the Younger. He also cites Dodgson (1876).

probabilities, of all the next-to-lowest probabilities, . . . , up to the average of all the highest probabilities. The calculation is different because a voter's probabilities necessarily sum to 1. When there are n motions, the average of the lowest probabilities is $\frac{1}{n}\left(\frac{1}{n}\right)$; the average of the next-to-lowest is $\frac{1}{n}\left(\frac{1}{n}+\frac{1}{n-1}\right)$; . . . ; and the average of the highest is $\frac{1}{n}\left(\frac{1}{n}+\frac{1}{n-1}+\ldots+\frac{1}{1}\right)$. For example, when $n = 3$, these numbers are $\frac{1}{3}\left(\frac{1}{3}\right) = \frac{2}{18}$, $\frac{1}{3}\left(\frac{1}{3}+\frac{1}{2}\right) = \frac{5}{18}$ and $\frac{1}{3}\left(\frac{1}{3}+\frac{1}{2}+1\right) = \frac{11}{18}$, so the scheme is the same as assigning 2 points to the last-place candidate in a voter's list, 5 to the next, and 11 to the highest place, which is equivalent to assigning 0, 2, and 6 (as compared with what is equivalent to Borda's: 0, 2, and 4). When there are four candidates, the scheme is equivalent to assigning 0, 2, 5, and 11 (as compared with what is equivalent to Borda's: 0, 2, 4, and 6). Given any preference-profile and a number of candidates, a scoring scheme is defined. Laplace advocated—again without further justification—a sum-scoring method with these points instead of Borda's, assuming that dishonest expressions of the estimates of the voters would play no role. These are very different schemes.

Borda was elected to the Académie des sciences in 1756, Condorcet in 1769, and Laplace in 1773. They knew each other. The model for voting for candidates championed by the first two, by Llull and Cusanus before them and Dodgson after them—where the voters' messages are preference-orders that are meant to represent their true preferences—has been the accepted dogma down to the present day. Laplace's idea—where voters evaluate the merits of candidates with numerical scores, given short shrift by Black and ignored by others—was discarded. Perhaps this is because Laplace implicitly did so himself, having thought that for practical reasons a winner would eventually emerge. With it disappeared his proposal for how to vote on what he as well as some others believed to be a different problem: voting on motions having a common end such as a budget.

Laplace insisted that Borda's method could be used only if voters honestly expressed their true preferences, for otherwise the outcomes could be perversions of the public will. He observed that this was the reason that institutions which had used Borda's method subsequently dropped it. He was presumably referring to the fact that Borda's scheme had been adopted for electing the members of the Académie des sciences in 1784 (or shortly after) but was discarded in 1800 because of the insistence of a newly elected member, Napoléon Bonaparte.

Borda's example shows the ease with which voters can manipulate the outcome:

$$1 : A \succ B \succ C \quad 7 : A \succ C \succ B \quad 7 : B \succ C \succ A \quad 6 : C \succ B \succ A.$$

The first-past-the-post system elects A, but if the six whose true preference-order is $C \succ B \succ A$ vote for B, then B is elected, which for them is better. The two-past-the-post system elects B, but if the seven whose true preference-order is $A \succ C \succ B$ vote for C in the first round, then A is eliminated and C is elected in the second round, which for them is better. The Borda-winner is C, but if the seven whose true preference-order is $B \succ C \succ A$ switched C and A, then B is elected, which they prefer. In the French, Anglo-American, and Borda systems it may pay voters *not* to vote for their favorites or to *not* honestly report their orders of preference.

In practice, voters can, and do, send messages that misrepresent their real preferences to abet the chances of the outcomes they seek (see chapter 2). If no candidate in France's presidential elections wins an absolute majority of the votes, there is a runoff between the top two candidates. If an elector is sure his favorite has no chance to survive the first round, he may vote for the candidate he prefers among those who do have a chance; or if his favorite is sure to be one of the top two, he may well do best by voting for the weakest realistic opponent to his favorite. Neither vote is in accord with the elector's true preferences. And, as mentioned earlier, Dodgson concluded that the road to foiling Condorcet-winners is to provoke cyclic-majorities.

More subtle misrepresentations may affect the winner of an election when the alternative vote or Nanson's system is used because placing a losing candidate lower in some electors' rankings can turn him into a winner (as illustrated in a previous example). When there are only two candidates (and an arbitrary number of electors) the simple majority-rule clearly elicits honest responses, and it is the only impartial method for which every elector's optimal strategy is clearly and unambiguously to vote honestly for one's preferred candidate (May 1952).

5.1 Gibbard-Satterthwaite's Impossibility Theorem

A method of voting is said to be *strategy-proof* when honesty is the best policy for every voter—the method induces the voter to express her true preference-order because it is the best strategy—and in the contrary case the method is *manipulable*. Examples have shown that Borda's, Saari's instant-Borda-runoff, and Nanson's methods are manipulable, for by altering their true preference-orders, coalitions of voters are able to improve the outcome for themselves. Strategy-proof methods are certainly very desirable. Regrettably, they do not exist.

When Φ is a profile and ϕ_i is the preference order of a voter i, the profile consisting of all other voters is Φ_{-i}, so $\Phi = (\phi_i, \Phi_{-i})$. A choice function f is *manipulable* if there exist Φ and ϕ_i' such that

$$f(\phi_i', \Phi_{-i}) \succ_{\phi_i} f(\phi_i, \Phi_{-i}),$$

meaning that there are situations when voter i with preferences ϕ_i can elect a candidate she prefers by sending a message ϕ_i' different from ϕ_i.

Gibbard-Satterthwaite's Impossibility Theorem *There is no choice function that is unanimous, nondictatorial, and strategy-proof for all preference-profiles (when there are at least three candidates).* (Gibbard 1973; Satterthwaite 1973)

Proof Here "unanimous" means that when every voter places a candidate A first, so does society. The theorem is established in two steps. First, "strong monotonicity" is shown to be a consequence of strategy-proofness. Second, assuming such a choice rule exists contradicts Arrow's theorem (third version).

A choice function f is *strongly monotonic* when $f(\Phi) = A$ and a candidate $B \neq A$ is lowered by one or more voters in Φ' (but the order among the other candidates remains the same), A remains the winner, i.e., $f(\Phi') = A$. A nonmanipulable f is necessarily strongly monotonic (Muller and Satterthwaite 1977).[3] For suppose the contrary; then there must be a Φ such that $f(\phi_i, \Phi_{-i}) = A$ and $f(\phi_i', \Phi_{-i}) = C$, where some B is strictly lower in ϕ_i' than in ϕ_i. Either $B = C$ or not.

First, suppose $B = C$. f nonmanipulable implies $A \succ_{\phi_i} B$.[4] But since $A \succ_{\phi_i'} B$, taking (ϕ_i', Φ_{-i}) as the original profile shows that i (when his preferences are ϕ_i') can manipulate to obtain A instead of B, a contradiction.

Next, suppose $B \neq C$. Take (ϕ_i', Φ_{-i}) as the original profile, and note that i by raising B obtains A instead of C, so the nonmanipulability of f implies $C \succ_{\phi_i'} A$. But since their relative positions have not changed, $C \succ_{\phi_i} A$, so that i with the original profile (ϕ_i, Φ_{-i}) can manipulate to obtain C instead of A, a contradiction.

A choice function f that is unanimous and strongly monotonic may be extended to a choice rule g (i.e., where g is defined over any subset $\mathcal{D}$ of the candidates $\mathcal{C}$). When $\mathcal{D} \subset \mathcal{C}$, let $\mathcal{D}' = \mathcal{C} - \mathcal{D}$. $\Phi^{\mathcal{D}}$ is a preference-profile over

3. Strongly monotonic is easily shown to be equivalent to "Maskin monotonicity" in the traditional model (Maskin 1999).

4. If the inputs were weak preferences, $\succeq$, this inference would conclude $A \succeq_{\phi_i} B$.

the subset of candidates $\mathcal{D}$. $\Phi^{\mathcal{D}/\mathcal{D}'}$ is a profile over all the candidates defined as follows: for every voter the profile in the top $|\mathcal{D}|$ places (if K is a set, $|K|$ is its cardinality) coincides with $\Phi^{\mathcal{D}}$, and in the bottom $|\mathcal{D}'|$ places each voter has the candidates of $\mathcal{D}'$ in any order whatsoever. Define

$$g(\Phi^{\mathcal{D}}) = f(\Phi^{\mathcal{D}/\mathcal{D}'}) = A.$$

This is an unambiguous definition for two reasons. First, $A \notin \mathcal{D}'$, for suppose otherwise and choose some candidate $C \in \mathcal{D}$. Lower every candidate in $\mathcal{D}$ other than C below A. Strong monotonicity implies A remains the winner; but C is in the first place of every voter, so unanimity implies C is the winner, a contradiction. Second, since no candidate of $\mathcal{D}'$ can be a winner, strong monotonicity implies they can be rearranged in any order without changing the outcome. Thus g is a choice rule.

Three of the properties of the choice rule version of Arrow's theorem (see chapter 3) are clearly satisfied: unrestricted domain (1), unanimity (2″), and nondictatorship (4). It remains to show (3″), namely, that if $g(\Phi^{\mathcal{D}}) = A$ and some nonwinner C is dropped, then A remains the winner. But that is obvious because

$$A = g(\Phi^{\mathcal{D}}) = f(\Phi^{\mathcal{D}/\mathcal{D}'}) = f(\Phi^{\mathcal{D}-C/\mathcal{D}'+C}) = g(\Phi^{\mathcal{D}-C}),$$

the second and last equations by definition, the third by the strong monotonicity of f the choice function. Thus g the choice rule satisfies all the properties of Arrow's theorem: the contradiction completes the proof. ■

Corollary *There is no choice function that is unanimous, nondictatorial and strongly monotonic for all preference-profiles (when there are at least three candidates).* (Muller and Satterthwaite 1977)

Observe that the results concern choice functions—winners—not ranking functions, but that is unavoidable because in the traditional model voters are unable to express preferences among rankings.

However, when the domain of preference-profiles is restricted to single-peaked preferences with respect to a fixed and known alignment of the candidates on the real line *and* voters' inputs are the names of exactly one candidate, Gibbard-Satterthwaite's impossibility may, in theory, be avoided, just as Arrow's impossibility was avoided. Hervé Moulin (1980) addressed the problem of selecting one candidate (or an alternative) among many, concluding that "as long as the alternatives can be ordered along the real line with the preferences of the agents being single-peaked, it makes little sense to object against

the Condorcet procedure, or one of its variants." He characterized the anonymous, efficient (Pareto optimal) choice rules that are strategy-proof. In practice they may be described as n different mechanisms, where n is the number of voters.[5] Identify each voter's preferred candidate on the real line. Some candidates may be the most preferred of many voters, some of no voters. For instance, suppose there are three candidates A, B, and C that go from left to right on the real line, and fifteen voters, where seven most prefer A, three most prefer B, and five most prefer C. The *kth* mechanism elects the *kth* of the most preferred candidates in going from left to right on the real line. Thus, in the example, the first through seventh mechanisms elect A; the eighth through tenth elect B; and the last five elect C. The median mechanism opts for the candidate of the median voter—the eighth in the example—and elects B, the Condorcet-winner because the preferences are single-peaked. Each of the n mechanisms is clearly strategy-proof: a voter who reports other than his most preferred candidate either changes nothing or changes the outcome to a less preferred candidate. It has also been shown that the same result holds if each voter reports a complete, single-peaked rank-order, though the social choice function will depend only on the peaks of the reported preferences (Barberà and Jackson 1994). But as *practical* schemes these methods make no sense at all: how are candidates for public office, competing wines, or Olympic skaters to be aligned on the real line so that the preferences of all voters or all judges are single-peaked relative to that alignment?

The attempts to escape Arrow's theorem by restricting the domain of preference-profiles have their counterparts in attempting to escape Gibbard-Satterthwaite's theorem. For example, if the domain is restricted to single-peaked preferences, any Condorcet-consistent method is strategy-proof (H. Moulin 1988, 263). But how can a voter who wishes to manipulate be restricted in any way? For example, if there were three candidates L (left), C (center), and R (right), a voter would be denied the inputs $L \succ R \succ C$ and $R \succ L \succ C$. Even if it were true that preferences are single-peaked in voting, it makes absolutely no sense to imagine that voters could be restricted to single-peaked inputs (indeed, how would the law be formulated?).

The Gibbard-Satterthwaite impossibility theorem adds one more reason for rejecting the traditional model. And yet, the majority judgment shows that there is a way of combating manipulation though not of eliminating it entirely. The essence of the idea must be attributed to Galton.

5. Moulin gives a purely theoretical description that is somewhat more complicated. His description introduces fictitious alternatives, and his characterization includes more mechanisms.

5.2 Galton's Middlemost

Some one hundred years after Laplace had raised the problem, Sir Francis Galton—distinguished statistician, pioneer of correlation and regression, inventor of fingerprint identification, geographer and explorer, meteorologist, founder of differential psychology, geneticist (and eugenicist), cousin of Charles Darwin, and best-selling author—proposed a different, considerably more convincing solution to the now well-identified *budget problem*:

> A certain class of problems do not as yet appear to be solved according to scientific rules, though they are of much importance and of frequent recurrence. Two examples will suffice. (1) A jury has to assess damages. (2) The council of a society has to fix on a sum of money, suitable for some purpose. Each voter, whether of the jury or the council, has equal authority with each of his colleagues. How can the right conclusion be reached, considering that there may be as many different estimates as there are members? That conclusion is clearly *not* the *average* of all the estimates, which would give a voting power to "cranks" in proportion to their crankiness. One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount, and the more an estimate diverges from the bulk of the rest, the more influence would it exert. I wish to point out that the estimate to which least objection can be raised is the *middlemost* estimate, the number of votes that it is too high being exactly balanced by the number of votes that it is too low. *Every other estimate is condemned by a majority of voters as being either too high or too low, the middlemost alone escaping this condemnation*. (Galton 1907a; our emphasis).

The budget problem has a very distinctive property: when the alternatives are quantities, they have a natural order.

Not a man to be satisfied by mere theory alone, Sir Francis applied the idea a week later. In doing so he displayed not only his practical spirit but also his wisdom and his wit:

> In these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. . .
>
> A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6*d*. each, on which to inscribe their names, addresses, and estimates of what the ox would weigh after it had been slaughtered and "dressed." Those who guessed most successfully received prizes . . . The judgments were unbiased by passion and uninfluenced by oratory and the like. The sixpence fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best . . . The average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case. (Galton 1907b)

He went on to analyze the 787 legible cards and determined that

> [a]ccording to the democratic principle of "one vote one value," the middlemost estimate expresses the *vox populi* . . . Now the middlemost estimate is 1207 lb., and the weight of the dressed ox proved to be 1198 lb.; so the *vox populi* was in this case 9 lb. or 0.8 per cent. of the whole weight too high . . .
>
> This result is, I think, more creditable to the trustworthiness of a democratic judgment than might have been expected. (Galton 1907b)

Galton's proposal was a procedure to decide on what budget to allocate for damages or for a project, for which there is no correct answer, so that no experiment is capable of verifying the validity of the procedure. Nevertheless, he had the flair to recognize an experiment that could test the value of the procedure: there *is* a correct answer to the weight of the dressed ox. It bolstered his faith in majority decision. His idea is, of course, the good one. It is curious that no one has pursued it in more than half a century of research on ranking and electing. Of course, Black, Downs, and other social choice theorists drew attention to the median but as a normative criterion or a prediction rather than as part of an actual decision-making procedure.

Most surprisingly the idea was actually used in the state of Florida in the years 1939–1968. "Millage" is a tax rate on property expressed in mills or thousandths of the value. School boards in the state were authorized to impose millage rates up to ten mills, and could increase them up to twenty mills if approved by referendum. A typical ballot addressed the voters in the following terms, where the eleven mills rate is the recommendation of the school board:

> **Instructions to voters**: Indicate by marking an "X" in the box after line one whether you favor the proposed millage levy which is necessary for the approved school term. If you favor a different millage levy, write the levy you favor in the box after line two.
>
> 1. Estimated millage levy required for regular term (11 mills) . . . . . □
> 2. Other millage levy . . . . . □

The rate which together with all higher rates just represented a majority of all the ballots cast became the millage rate; in our terminology, the majority-rate was the choice, that is, the middlemost when the number of ballots was odd and the lower middlemost when the number of ballots was even. The system was abandoned in 1968 when Florida adopted a new constitution (Holcombe and Kenny 2007). Why and at whose urging the method was chosen remains a mystery.

It seems a strange quirk in the history of ideas that Galton's idea has lain dormant. This may be due to Duncan Black, who called Galton's a "small contribution" that had no doubt been "made independently by many other people" (1958, 188), explaining that he mentioned it only because of Galton's stature

as a statistician. Yet Galton's idea is far from alien to Black's main result. It is natural to order the different money amounts to be budgeted from lowest to highest; and it is not unreasonable to suppose that each member of the council or the jury has a preference for some one ideal budget and that her liking for the others decreases the more they differ from her ideal. Galton had realized that relative to the natural order every voter's or judge's preference-order in the budgeting problem—or the "weight of the dressed ox" problem—is single-peaked. And so he argued for the middlemost estimate: "Every other estimate is condemned by a majority of voters as being either too high or too low, the middlemost alone escaping this condemnation."

5.3 Majority Judgment Methods

The majority-grade and the majority-ranking (or majority-value or majority-gauge)—the principal concepts of the theory developed in this book—may be seen as generalizations of Galton's astute observation (though implicitly he seems to have assumed single-peaked preferences).

Two principal methods for finding solutions in the context of the traditional model survive close analysis: Borda's for winners, Condorcet's for rankings. Their major drawback is their manipulability. That is primarily due to the fact that both methods sum points to obtain scores. The majority judgment, applied respectively to Borda-points and Condorcet-points, aggregate them not by summing but by taking the middlemost points.

Borda-majority judgment method A voter's input is a rank-order of the candidates. It determines the Borda-points assigned by the voter to each candidate. A candidate's majority-grade and majority-value (or majority-gauge) are computed on the basis of his Borda-points. A candidate with a highest majority-value (or majority-gauge) is elected.[6]

To understand how the method works, consider again the real example of the Social Choice and Welfare (SCW) Society's presidential election (Brams and Fishburn 2001; Saari 2001a):

$13 : A \succ B \succ C \quad 11 : A \succ C \succ B \quad 9 : B \succ C \succ A$
$11 : C \succ A \succ B \quad 8 : C \succ B \succ A.$

C is the Condorcet-winner, and the majority-rule-ranking, $C \succ_S A \succ_S B$, is transitive (so it is the Condorcet-ranking). The Borda-points of candidates (in

6. The median of the Borda points was proposed by Basset and Persky (1999). However, no tie-breaking rule was given, which in practice is essential (see chapter 7).

Table 5.1a
Borda-Majority Judgment Method, SCW Society Election

	Borda-Points			Majority-Value	Majority-Gauge (p, α, q)	Majority-Ranking	Borda-Score
	2	*1*	*0*				
A	24	11	17	$1.\overbrace{1\ldots1}^{4}2$	(24, *1*+, 17)	1st	59
C	19	20	13	$1.\overbrace{1\ldots1}^{7}1$	(19, *1*+, 13)	2d	58
B	9	21	22	$1.\overbrace{1\ldots1}^{7}0$	(9, *1*−, 22)	3d	39

Note: Majority-values are truncated.

italics, to distinguish them as grades) are given in table 5.1a. *A*, for example, is ranked first 24 times (so is assigned that number of *2*'s), ranked second 11 times (so is assigned that number of *1*'s), and ranked last 17 times (so is assigned that number of *0*'s). All three candidates have the same majority-grade of *1*. The majority-gauges are sufficient to determine the order (here p and q are the numbers of grades higher and lower than the majority-grade rather than their percentages). The majority-values are written only with the precision needed to rank the candidates. The Borda-majority method makes *A* the winner and ranks the candidates in the order $A \succ_S C \succ_S B$ (in agreement with the Borda-ranking).

The Borda-points are *not* summed, so the method resists manipulation. For example, if two of the eleven voters with preference-order $C \succ A \succ B$ manipulated by reporting instead $C \succ B \succ A$, the Borda-ranking would change to $C \succ_S A \succ_S B$, but the majority-ranking would remain the same: it would take at least seven of those eleven voters to manipulate in the same way to change the outcome of the majority-ranking.

The method is impartial, unanimous, and nondictatorial. However, independence of irrelevant alternatives (IIA) may be violated (necessarily, by Arrow's theorem), and it is: if *B* withdraws, the profile becomes $24 : A \succ C$ and $28 : C \succ A$, so *C* is ranked ahead or wins against *A*, as shown in table 5.1b.

A similar approach defines a method for ranking:

Condorcet-majority judgment method A voter's input is a rank-order of the candidates. It determines the Condorcet-points assigned by a voter to each ranking. A ranking's majority-grade and majority-value (or majority-gauge) are computed on the basis of its Condorcet-points. A ranking with a highest majority-value (or majority-gauge) is chosen.

Table 5.1b
Borda-Majority Judgment Method, SCW Society Election, Restricted to Candidates A, C

	Borda-Points		Majority-Grade	Majority-Ranking
	1	*0*		
C	28	24	*1*	1st
A	24	28	*0*	2d

Table 5.1c
Condorcet-Majority Judgment Method, SCW Society Election

	Condorcet-Points				Majority-Value	Majority-Gauge	Majority-Ranking	Condorcet-Score
	3	*2*	*1*	*0*				
$A \succ_S C \succ_S B$	11	24	8	9	$2.\overbrace{2\ldots2}^{7}2$	(11, 2−, 17)	1st	89
$C \succ_S A \succ_S B$	11	19	22	0	$2.\overbrace{2\ldots2}^{7}1$	(11, 2−, 22)	2d	93
$C \succ_S B \succ_S A$	8	20	11	13	$2.\overbrace{2\ldots2}^{3}1$	(8, 2−, 24)	3d	75
$A \succ_S B \succ_S C$	13	11	20	8	$1.\overbrace{1\ldots1}^{4}2$	(24, 1+, 8)	4th	81
$B \succ_S A \succ_S C$	0	22	19	11	$1.\overbrace{1\ldots1}^{8}2$	(22, 1+, 11)	5th	63
$B \succ_S C \succ_S A$	9	8	24	11	$1.\overbrace{1\ldots1}^{8}1$	(17, 1+, 11)	6th	67

Note: Majority-values are truncated.

This is applied to the same example in table 5.1c. "Grades" are given to each ranking. For example, $A \succ B \succ C$ is given the grade of *3* Condorcet-points 13 times (because 13 preference-orders agree in all three comparisons of candidates, namely, $A \succ B \succ C$), *2* Condorcet-points 11 times (because 11 preference-orders agree in two comparisons of candidates, namely, $A \succ C \succ B$), *1* Condorcet-point 20 times (because 20 preference-orders agree in one comparison, namely, $B \succ C \succ A$ and $C \succ A \succ B$), and *0* Condorcet-points 8 times (because 8 preference-orders disagree completely, namely, $C \succ A \succ B$).

The Condorcet-majority method's solution to the SCW Society problem is $A \succ_S C \succ_S B$. It agrees with the Borda-majority method but not with the solution of Condorcet's method, $C \succ_S A \succ_S B$ (compatibility with Borda is not to be expected in general). But the Condorcet-points are *not* summed, as they are in Condorcet's method, and manipulation is resisted with this method, too.

Consider again the $3k$-voter profile $k\text{-}Cond(A \succ B \succ C)$:

Table 5.2
Condorcet-Majority Judgment Method, $3k$ Voters with Profile $k\text{-}Cond(A \succ B \succ C)$

	Condorcet-Points				Majority-Grade	Majority-Ranking	Condorcet-Score
	3	*2*	*1*	*0*			
$\mathcal{R}_2 : A \succ_S C \succ_S B$	0	$2k$	0	k	*2*	1st	$4k$
$\mathcal{R}_1 : A \succ_S B \succ_S C$	k	0	$2k$	0	*1*	4th	$5k$

Note: Three rankings are tied for 1st, three for 4th.

$$k : A \succ B \succ C \quad k : B \succ C \succ A \quad k : C \succ A \succ B.$$

As table 5.2 shows, Condorcet's method selects the three rankings that with $\mathcal{R}_1$ form a Condorcet-cycle; the Condorcet-majority method selects the compromise solution, namely, the other three rankings that with $\mathcal{R}_2$ form a Condorcet-cycle. But the Condorcet-majority method is not choice-compatible. For suppose there is one voter whose preference-order is $A \succ B \succ C$; the (unanimous) winner is A and the (unanimous) loser is C. When the $3k$-voter profile that is a Condorcet-component is adjoined, the method selects $A \succ_S C \succ_S B$ and $B \succ_S A \succ_S C$; choice-compatibility would have it select $A \succ_S B \succ_S C$.

Now take the profile to be $k\text{-}Cond(A \succ B \succ C \succ D)$ over $4k$ voters:

$$k : A \succ B \succ C \succ D \quad k : B \succ C \succ D \succ A$$
$$k : C \succ D \succ A \succ B \quad k : D \succ A \succ B \succ C.$$

In this case (see table 5.3) the Condorcet-majority method and the Condorcet-score yield different winning rankings and different rankings of the rankings.

The point of departure of the Condorcet-majority method is that each voter's input is a rank-order of the candidates that gives a grade to each possible rank-order, its Condorcet-points. The grades are then used to determine society's rank-order (the Condorcet method adds them, the Condorcet-majority method applies the majority-value). The grade given to a rank-order is the number of times the order between two candidates agrees with the voter's rank-order. Thus, for example, when there are three candidates A, B, and C and a voter's input is $A \succ B \succ C$, the voter gives a grade of *3* to $A \succ B \succ C$; a grade of *2* to $A \succ C \succ B$ and $B \succ A \succ C$; a grade of *1* to $C \succ A \succ B$ and $B \succ C \succ A$; and a grade of *0* to $C \succ B \succ A$. A higher number grade is better.

The construction given in the proof of theorem 4.5 suggests a more refined idea for assigning grades. If a voter's input is $A_n \succ A_{n-1} \succ \cdots \succ A_1$, then for $(j_1, j_2, \ldots, j_n)$ a permutation of $(1, 2, \ldots, n)$,

$\{A_{j_1} \succ A_{j_2} \succ \cdots \succ A_{j_n}\}$ is assigned the grade $(j_1, j_2, \ldots, j_n)$.

Table 5.3
Condorcet-Majority Judgment Method, $4k$ Voters with Profile k-*Cond*$(A \succ B \succ C \succ D)$

	Condorcet-Points						
	6	*5*	*4*	*3*	*2*	*1*	*0*
$\mathcal{R}_3 : A \succ_S C \succ_S D \succ_S B$	0	0	$2k$	k	0	k	0
$\mathcal{R}_1 : A \succ_S B \succ_S C \succ_S D$	k	0	0	$2k$	k	0	0
$\mathcal{R}_6 : A \succ_S D \succ_S C \succ_S B$	0	0	k	$2k$	0	0	k
$\mathcal{R}_2 : A \succ_S B \succ_S D \succ_S C$	0	k	k	0	k	k	0
$\mathcal{R}_4 : A \succ_S D \succ_S B \succ_S C$	0	k	k	0	k	k	0
$\mathcal{R}_5 : A \succ_S C \succ_S B \succ_S D$	0	k	0	k	$2k$	0	0

Note: Majority-values are truncated.
Four rankings are tied for 1st; four for 5th; four for 9th; eight for 13th; and four for 21st.
See table 4.3 for details concerning the rank-orders having the same Condorcet-points as each $\mathcal{R}_i$.

There are in all $n!$ *top-preferred-grades* with $(j_1, j_2, \ldots, j_n)$ a better grade than $(k_1, k_2, \ldots, k_n)$ if $(j_1, j_2, \ldots, j_n) \gg_c (k_1, k_2, \ldots, k_n)$. Thus, when there are three candidates A, B, and C and a voter's input is $A \succ B \succ C$, the voter gives the following grades (where for brevity, ABC stands for $A \succ B \succ C$):

Rank-order:	ABC	ACB	BAC	CAB	BCA	CBA
Grade:	(3, 2, 1)	(3, 1, 2)	(2, 3, 1)	(1, 3, 2)	(2, 1, 3)	(1, 2, 3)

The order among the top-preferred-grades agrees with, but is more refined than, the Condorcet-points. The Condorcet-points are the same for $A \succ C \succ B$ and $B \succ A \succ C$, whereas the first is better than the second according to the top-preferred-grades. This is not true when there are more than three candidates.

Top-preferred-majority judgment method A voter's input is a rank-order of the candidates. It determines the top-preferred-grades assigned by a voter to each ranking. A ranking's majority-grade and majority-value (or majority-gauge) are computed on the basis of its top-preferred-grades. A ranking with a highest majority-value (or majority-gauge) is a chosen.

The top-preferred-majority judgment method is applied to the SCW Society election problem in table 5.4. The ranking $A \succ C \succ B$ has 13 grades of (3, 1, 2), for example, because the input of 13 voters is $A \succ B \succ C$. The solution happens to be exactly the same rank-ordering of the rank-orders as the Condorcet-majority method, though this clearly will not always be the

Table 5.3
(cont.)

	Majority-Value	Majority-Rank	Condorcet-Score
$\mathcal{R}_3 : A \succ_S C \succ_S D \succ_S B$	*3.4*	1st	$12k$
$\mathcal{R}_1 : A \succ_S B \succ_S C \succ_S D$	$3.\overbrace{3\ldots3}^{2k-1}2$	5th	$14k$
$\mathcal{R}_6 : A \succ_S D \succ_S C \succ_S B$	$3.\overbrace{3\ldots3}^{2k-1}0$	9th	$10k$
$\mathcal{R}_2 : A \succ_S B \succ_S D \succ_S C$	*2.4*	13th	$12k$
$\mathcal{R}_4 : A \succ_S D \succ_S B \succ_S C$	*2.4*	13th	$12k$
$\mathcal{R}_5 : A \succ_S C \succ_S B \succ_S D$	*2.3*	21st	$12k$

case. Notice that not one of the three "majority judgment methods" places the Condorcet-winner C first. The top-preferred-majority method resists manipulation. The very nature of the grades it uses shows that it is nonsense to assign them numerical values and sum them to determine society's preferences.

5.4 The Majority Judgment for the Traditional Model

The two most famous impossibility theorems, Arrow's and Gibbard-Satterthwaite's, together with the incompatibility between choosing and ranking combine to prove one basic truth: within the age-old model there is no satisfactory scheme for determining a winner or an order of merit among candidates

Table 5.4
Top-Preferred-Majority Judgment Method, SCW Society Election

	Top-Preferred-Grades								
	$\alpha =$ (321)	$\beta =$ (312)	$\gamma =$ (231)	$\delta =$ (132)	$\epsilon =$ (213)	$\zeta =$ (123)	Majority–Gauge	Majority-Rank	Condorcet-Score
ACB	11	13	11	0	8	9	$(24, \gamma^+, 17)$	1st	89
CAB	11	8	11	9	13	0	$(19, \gamma^-, 22)$	2d	93
CBA	8	11	9	11	0	13	$(19, \gamma^-, 24)$	3d	75
ABC	13	11	0	11	9	8	$(24, \delta^+, 17)$	4th	81
BAC	0	9	13	8	11	11	$(22, \delta^-, 22)$	5th	63
BCA	9	0	8	13	11	11	$(17, \delta^-, 22)$	6th	67

Note: ABC stands for $A \succ B \succ C$, and (321) for (3, 2, 1).

Notice that the sum of the columns corresponding to the grades β and γ equals that of the column of the Condorcet-grade *2* in table 5.1c; the same is true for the grades δ and ϵ and the Condorcet-grade *1*.

unless there are only two candidates. But suppose a practitioner faces a situation where, owing to lack of time or lack of understanding or some other cause, it is impossible to define a common language with which to evaluate candidates; all that can be obtained from voters or judges are rank-orderings of candidates. What then? *In the context of the traditional model, use the Borda-majority judgment method.*

Why? Inputs are messages. Ideally, they would be grades, permitting voters total freedom of expression (within the bounds of the language of grades). But why use Borda-points instead of some other strictly monotonic scoring scheme $s_1 > \cdots > s_n$? Laplace justified their use assuming an underlying uniform distribution of the competitors' merits on an arbitrary interval $[0, R]$ of the real line. But that assumption seems unreasonable: very high grades and very low grades are rare. Laplace derived a completely different set of scores when he assumed voters assign probabilities to alternative motions. The Denmark school scale used the ten marks $\{0, 3, 5, 6, \ldots, 11, 13\}$, omitting *1, 2, 4*, *12* (see chapter 1), for evaluating students because of the observed distribution of their merit. Different underlying distributions may be used to justify the use of different scoring schemes (see chapter 8). When different scoring schemes are used to grade and to rank by adding or averaging, they yield very different solutions. The results of the Borda-majority method are one and the same for *every* distribution and thus for *every* strictly monotonic scoring scheme.

Furthermore, the Borda-majority judgment method enjoys many of the good properties of the majority judgment. It is *strategy-proof-in-grading*. Since a candidate's final grade is the majority-grade, any voter who ranks the candidate above it cannot increase the final grade by placing the candidate higher in his input, and similarly, any voter who ranks the candidate below it cannot decrease the final grade by placing the candidate lower in his input. The Borda-majority method is perhaps even more resistant to manipulation because if a voter places some candidate higher (or lower) in his preference-order, then he necessarily places one or more others lower (or higher), thus complicating the effort to manipulate. This method is also *group strategy-proof-in-grading*: the same holds for groups of voters. Moreover, a candidate's Borda-majority-grade is sure to increase by 1 only if a majority of the voters increase the candidate's ranking by 1. It is only in unusual circumstances that one or a very few voters can change the grade. These arguments concern grades. Other strategic considerations come into play if the aim is not grades but the rank-ordering induced by grades. The method is *partially strategy-proof-in-ranking*: if society's rank-order placed one candidate ahead of another, $A \succ_S B$, and some voter preferred the opposite, that voter could not change her input and both lower A's majority-grade and raise B's majority-grade; at most, the voter could either raise B's or lower

A's. The method resists manipulation in still other ways, as illustrated by the example of table 5.1a. Finally, it almost reconciles Borda and Condorcet: the top-preferred-majority-winner always has the highest Borda-majority-grade.

The Borda-majority judgment method is a new and more satisfactory solution to the problem of choosing a winner and a ranking in the context of the traditional model. It happens that (almost) this method was used in practice for many years by the International Skating Union to rank figure skaters. Specifically, an odd number of judges ranked the skaters, and the majority-grades determined the jury's ranking together with an ad hoc assortment of rules to break ties (see chapter 7).

The Borda-majority judgment method is impartial, resists manipulation, and satisfies winner-loser unanimity, choice-compatibility, and choice-monotonicity, but it (necessarily) fails rank-compatibility and rank-monotonicity (and so also fails independence of irrelevant alternatives). The reason for this failure is that the scale of grades changes when the number of candidates changes, and this can induce changes in the majority-ranking. That is its unavoidable weakness. When there is a common language, that weakness is overcome.

6 Fallacies of the Traditional Model in Voting

During the Middle Ages there were all kinds of crazy ideas, such as that a piece of rhinoceros horn would increase potency. Then a method was discovered for separating the ideas—which was to try one to see if it worked, and if it didn't work, to eliminate it. This method became organized, of course, into science.
—Richard P. Feynman

Several centuries of work on the theory of social choice have produced very substantial contributions, notably, in identifying a host of important properties or criteria that should (or should not) be satisfied by a mechanism that amalgamates the beliefs, desires, or wills of individuals into a decision of society.

Arrow's paradox must be avoided: a method should satisfy independence of irrelevant alternatives, that is, the presence or absence of some candidate should not cause a change of winner between two others. Condorcet's paradox must be avoided: a method should yield a transitive order of finish among the competitors. A method should be monotonic: a winning candidate who receives more votes or rises in the rank-orders of candidates must remain the winning candidate. A unanimous decision among individual voters must be the decision of society. Mechanisms should make voters' optimal strategies be those messages that honestly express their beliefs; or, if no such mechanism can be found, then one that best resists strategic manipulation and best incites the electorate to express themselves honestly must be found.

Regrettably, the theory shows that even when voters eschew strategic voting and honestly express their convictions, there exists no method that satisfies the essential criteria, unless it is assumed that voters have very restricted types of unrealistic views. The impossibility and incompatibility theorems prove that the traditional model harbors internal inconsistencies. The reason for this conundrum is the basic paradigm of social choice: voting depends on comparisons between pairs of candidates—one is better than another—so voters have

lists of preferences in their minds. Instead of inputs that evaluate the *absolute merits* of candidates, the inputs compare the *relative standings* of candidates. But even the idea of comparing is questionable: if the decision or output is to be a rank-order of the candidates, should not the voters be asked to compare the relative merits of the various possible rank-orders rather than only the relative merits of candidates?

6.1 Unrealistic Inputs

Every bit as damning as the logical inconsistencies of the theory is the fact that the traditional paradigm of voting theory—that voters, when confronted by a set of candidates, compare them or rank-order them—is simply wrong. Voters do not go to the polls with rank-orders of the candidates in their minds. The French presidential elections of 2002 and 2007 had, respectively, sixteen and twelve candidates. Instead of effecting a rank-ordering a voter ignored most candidates as unacceptable and looked upon a few with varying intensities of approval or disapproval. The model that underlies the theory simply does not correspond to reality. Experimental evidence proves this conclusively. Information drawn from three electoral experiments refute the traditional view as well as several other preconceived ideas.

The Orsay experiment (see chapters 1 and 15) tested the majority judgment, so voters' inputs were expressed in a common language of grades—*Excellent, Very Good, Good, Acceptable, Poor*, and *To Reject* [1]—evaluating the merits of candidates. Of the 2,360 who voted, 1,752 officially participated; 1,733 ballots were valid. Contrary to the predictions of some, the voters had no difficulty in filling out the ballots, usually doing so in about one minute. In fact, every member of the team conducting the experiment had the impression that the participants were very glad to have the means of expressing their opinions concerning *all* the candidates, and were delighted with the idea that candidates would be assigned final grades.[2] An effective argument to persuade reluctant voters to participate was that the majority judgment allows a much fuller expression of a voter's opinions. The actual system offered voters only thirteen possible *messages*: to vote for one of the twelve candidates or to vote for none. Several participants actually stated that the experiment had induced them to vote for the first time: finally, a method that permitted them to express

1. *Très Bien, Bien, Assez Bien, Passable, Insuffisant*, and *A Rejeter*.

2. A collection of television interviews of participants prepared by Raphaël Hitier, a journalist of *I-Télé*, confirms these impressions. Also, a questionnaire used in the ILC experiment (see chapter 17) shows that voters prefer using three number grades rather than two.

themselves. The majority judgment offered voters more than 2 billion possible messages with which to express themselves (with twelve candidates and six grades, there are $6^{12} = 2,176,782,336$ possible messages). The voters' relative ease of expression in the face of so vast a choice shows that assigning grades is cognitively simple, certainly much simpler than ranking candidates (as any teacher or professor faced with ranking students will attest). Of the 1,733 valid majority judgment ballots, 1,705 were different. It is surprising that they were not all different. Had all those who voted in France in 2007 (some 36 million) cast different majority judgment ballots, fewer than 1.7% of the possible messages would have been used. Those that were the same among the 1,733 valid messages of the experiment contained only *To Reject*s or accorded *Excellent* for one or several candidates and *To Reject* for all the others.

Voters were particularly happy with the grade *To Reject* and used it the most. There was an average of 4.1 of *To Reject* per ballot and an average of 0.5 of no grade (which, in conformity with the stated rules, was counted as a *To Reject*). Voters were parsimonious with high grades and generous with low ones (see table 6.1). Only 52% of voters used a grade of *Excellent*; 37% used *Very Good* but no *Excellent*; 9% used *Good* but no *Excellent* and no *Very Good*; 2% gave none of the three highest grades. The opinions of voters are richer, more varied and complex by many orders of magnitude than they are allowed to express with any current system.

The highest grades were often multiple (see table 6.2). In all, *more than 33% of the ballots gave the highest grade to at least two candidates*. Thus one of every three voters did not designate a single best candidate. This shows that many voters either saw nothing (or very little) to prefer among several candidates, or at the least, they were very hesitant to make a choice among two, three, or more candidates. Moreover, many voters did not distinguish between the leading candidates: 17.9% gave the same grade to Bayrou and Sarkozy (10.6% their highest grade to both), 23.3% the same grade to Bayrou and Royal (11.7% their highest grade to both), and 14.3% the same grade to Sarkozy and Royal (4.1% their highest grade to both). Indeed, 4.8% gave the same grade to all three (4.1% their highest grade to all three: all who gave their highest grade to Sarkozy and Royal also gave it to Bayrou). These are significant percentages: many elections are decided by smaller margins. These are valid, significant inputs of opinion that are completely ignored by the traditional model.

This finding is reinforced by a poll conducted on election day (by TNS Sofres–Unilog, Groupe LogicaCMG, April 22, 2007) that asked at what moment voters had decided to vote for a particular candidate. Their hesitancy in making a choice is reflected in the answers: 33% decided in the last week, one-third of

Table 6.1
Average Number of Grades per Majority Judgment Ballot, Three Precincts of Orsay, April 22, 2007

	Average No. of Grades per Ballot
Excellent	0.69
Very Good	1.25
Good	1.50
Acceptable	1.74
Poor	2.27
To Reject	4.55
Total	12.00

Note: Of the 4.55 *To Reject*, 0.5 corresponded to no grade.

Table 6.2
Multiple Highest Grades, Three Precincts of Orsay, April 22, 2007

Two or more *Excellent*	11%
Two or more *Very Good,* none higher	16%
Two or more *Good,* none higher	6%

whom (11%) decided on election day itself. For Bayrou voters 43% decided in the last week and 12% on election day; for Sarkozy voters the numbers were 20% and 6%; for Royal voters, 28% and 9%; for Le Pen voters, 43% and 18%. In contrast, the system *forced* them to make a choice of one (or to vote for no one).

Moreover, inputs that are rank-orders—or that simply show preferences between pairs of candidates—ignore how voters *evaluate* the respective candidates (just as the 2002 runoff ignored the respective evaluations of Chirac and Le Pen) except, of course, that one is evaluated higher than the other. Over one-half of the highest grades are less than *Excellent*. Two-thirds of the second highest grades are merely *Good* or worse (see table 6.3). To be first, second, or third in a ranking of at least three candidates carries very different meanings to different voters that are completely ignored by the inputs to the traditional model. This is still another reason that aggregating rank-orders (as do the methods of Condorcet and of Borda, and their combinations) is not meaningful.

The Faches-Thumesnil experiment tested two versions of the alternative vote, so voters' inputs were rank-orders of the candidates (Farvaque, Jayet, and Ragot 2007). The experiment was conducted in two of the eleven voting precincts of Faches-Thumesnil, a small town in France's northernmost department, Nord. Voters were not obliged to rank-order all candidates (as in Australia): a candidate

Table 6.3
Distributions: Highest Grade, Second Highest Grade, Third Highest Grade, Three Precincts of Orsay, April 22, 2007

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
Highest	52%	37%	9%	2%	0%	1%
Second highest	–	35%	41%	16%	5%	3%
Third highest	–	–	26%	40%	22%	13%

Table 6.4
Number of Candidates Rank-Ordered, Faches-Thumesnil Experiment, April 22, 2007

	Number of Candidates Rank-Ordered			
	1–3	4–6	7–11	12
No. of ballots	260	210	53	370
Percent of ballots	29.1%	23.5%	5.9%	41.4%

not on the list of a voter's ballot was considered off the list and thus could never be placed first on the voter's list after elimination of other candidates. Of those who voted officially, 960 (or 60%) participated in the experiment, 67 ballots were invalid, and 893 were valid. Almost 60% of the ballots did not rank-order all candidates and over 50% rank-ordered six or fewer of the twelve candidates, showing that voters are reluctant to rank-order many candidates (see table 6.4).

Admittedly, it is a difficult and time-consuming task to rank-order alternatives, and in any case, whether a voter rank-orders many or few candidates, she is unable to express any sense of how much or how little any of the candidates are appreciated. Suppose there are n candidates. To rank-order them a voter first places some one candidate on a list; then places the second in the slot above or below; then the third in one of the three slots (above, between, below); and so on. This takes $n(n+1)/2$ time units. And if ties are not allowed, ranking becomes even more difficult. In contrast, it is a much easier task to assign each candidate a grade than to rank-order candidates. In practice, with a natural well-understood language of grades, a voter quickly situates an approximate grade for each candidate (e.g., Sarkozy is *Good* or *Very Good*) and thus takes about $2n$ time units (and in any case, a maximum of mn when there are m grades). Cognitively, assigning grades seems to be a much simpler exercise than ranking candidates. But whatever the reason, ranking a large number of alternatives is clearly very difficult, as is seen by the fact that about 95% of Australian voters rely on predetermined rankings provided by their parties.

The two versions of the alternative vote tested concerned the choice of candidate to eliminate when there is no majority for any candidate among the (current) first places. The Australian system eliminates the candidate listed first the least number of times. The other version eliminates the candidate listed last the greatest number of times. The Australian version makes Sarkozy the winner; the other version makes Bayrou the winner. The Australian version is less favorable to centrists because major candidates of the right and the left are usually either high or low on voters' lists. This may explain why it is used in practice rather than the other version. The other method, sometimes called the Coombs method, guarantees the election of the Condorcet-winner when the preferences are single-peaked and the votes are sincere, which is not true of the first method (see Grofman and Feld 2004; Nagel 2007).

The official first-round results in the two voting precincts of Faches-Thumesnil were very close to the national percentages (table 6.5). The voters' rank-orders make it possible to compute the results of the face-to-face confrontations (table 6.6). They yield the same unambiguous order of finish among the four significant candidates as did the polls on March 28 and April 19 (see table 2.11). Once again the Condorcet-order agrees with the Borda-ranking: Bayrou $\succ$ Sarkozy $\succ$ Royal $\succ$ Le Pen.

Table 6.5
Official First-Round Votes, National and Two Precincts of Faches-Thumesnil, April 22, 2007

	Sarkozy	Royal	Bayrou	Le Pen	Besancenot	de Villiers
National	31.2%	25.9%	18.6%	10.4%	4.1%	2.2%
Faches-Thumesnil	29.7%	25.5%	19.7%	12.0%	3.7%	2.4%
	Buffet	Voynet	Laguiller	Bové	Nihous	Schivardi
National	1.9%	1.6%	1.3%	1.3%	1.1%	0.3%
Faches-Thumesnil	2.4%	1.4%	1.5%	0.9%	0.5%	0.3%

Table 6.6
Projected Second-Round Results, Faches-Thumesnil Experiment, April 22, 2007

	Bayrou	Sarkozy	Royal	Le Pen
Bayrou	–	52%	60%	80%
Sarkozy	48%	–	54%	83%
Royal	40%	46%	–	73%
Le Pen	20%	17%	27%	–

Note: For example, Sarkozy has 48% of the votes against Bayrou.

6.2 Statistical Left-Right Spectra

The one escape from the inner inconsistencies of the traditional model of social choice occurs when voters have single-peaked preferences relative to a common ordering of the candidates. It comes from the idea that in the political realm candidates may be listed on a line from left to right, voters place themselves somewhere along it, prefer the candidate closest to their position, and dislike candidates more, the more distant they are from their position. Were there such a line, and were it true that voters' preferences for candidates are single-peaked, inputs of rank-orders would satisfy the aims of the traditional theory: the winner would be the Condorcet-winner, and the order of finish would be transitive in conformity with the face-to-face votes. The reality, long recognized, is that there is no such left-to-right line for which preferences are single-peaked. New experimental evidence confirms it.

An electoral experiment was conducted in parallel with the French presidential election of 2002 in five of Orsay's twelve voting precincts (under the same general conditions as the 2007 Orsay experiment). Its aim was to test approval voting (see chapter 18 for a detailed description of the experiment). The experimental ballot contained a list of the sixteen candidates together with instructions saying:

Rules of approval voting: The elector votes by placing crosses [in boxes corresponding to candidates]. He may place crosses for as many candidates as he wishes, but not more than one per candidate. The winner is the candidate with the most crosses.

On average there were 3.15 crosses per ballot. The total number of different possible messages was $2^{16} = 65{,}536$. Of the 2,587 valid ballots, 813 were different. Voters had no incentive to vote other than sincerely, namely, if a cross was given to some candidate C, then a cross was given to every candidate preferred to C as well. But if there existed a left-to-right line relative to which the voters' preferences are single-peaked, then the total number of different possible sincere votes would have been 137. The crosses would have to have been consecutive with regard to the alignment along the spectrum: there are 16 sincere messages with one cross, 15 with two consecutive crosses, 14 with three consecutive crosses, . . . , 1 with sixteen consecutive crosses, and 1 with no crosses, so in all 137 sincere votes. The large discrepancy between 137 and 813 proves that the single-peaked condition was far from satisfied.

To assume single-peaked preferences is certainly not valid in elections. On the other hand, there is no denying that candidates and their political parties seeking election are commonly described in terms of a left-right spectrum and that this makes sense to political scientists, journalists, and the general public in

France, the U.K., the U.S.A., and throughout the world. The Orsay experiments of 2002 and 2007 both give solid scientific evidence that this is a valid concept.

Ballots from the 2002 experiment with several crosses yield statistical information about how voters favorable to one candidate might transfer their votes to others. For example, an estimate for, say, Bayrou may be computed as follows: among the ballots containing a cross for Bayrou and $k \geq 1$ other crosses, attribute $1/k$ to each of the other candidates with a cross, and find the sum given each candidate. The estimate of the transfer to a candidate is the percentage that the candidate's sum represents of the total sum (see table 6.7). *Statistically, the voters' transfers are almost single-peaked among the important candidates*. For instance, among those who gave a cross to Chirac, Bayrou was the most likely transfer, and the further distant from Chirac on the left-right line, the less likely the transfer. This does not hold for the unimportant candidates. The deviations are strikingly small among the important candidates and are easily explained. Chirac, the incumbent president, often exerted an appeal to voters in excess of the left-right spectrum (e.g., 16% of Chevènement voters go to Chirac, only 14% to Bayrou); crosses were sometimes given to the far right and the far left as expressions of opposition (e.g., more Gluckstein voters go to Le Pen than to Bayrou).

Table 6.7
Estimated Transfers of Votes to Important Candidates, Based on 2002 Orsay Experiment

	Left ⟵ Mamère	Jospin	Chevènement	Bayrou	Chirac	⟶ *Right* Le Pen	Ten Others
Gluckstein	15%*	9%	5%	2%	4%	5%	60%
Laguiller	14%	20%*	9%	4%	7%	5%	45%
Hue	13%	33%*	10%	3%	2%	2%	37%
Besancenot	20%	21%*	9%	5%	3%	3%	39%
Mamère	–	**38%***	**8%**	**7%**	**4%**	**1%**	42%
Taubira	15%	28%*	10%	8%	4%	0%	35%
Jospin	**26%***	–	**15%**	**8%**	**5%**	**1%**	45%
Chevènement	**8%**	**20%***	–	**14%**	**16%**	**6%**	36%
Bayrou	**6%**	**10%**	**13%**	–	**27%***	**4%**	40%
Chirac	**3%**	**5%**	**13%**	**24%***	–	**10%**	45%
Madelin	3%	4%	9%	22%	32%*	6%	24%
Lepage	7%	12%	12%	17%*	16%	2%	34%
Boutin	4%	4%	6%	23%*	17%	5%	41%
Le Pen	**3%**	**3%**	**13%**	**10%**	**26%***	–	45%
Saint-Josse	3%	6%	10%	10%	23%*	9%	39%
Mégret	1%	1%	5%	11%	22%	36%*	24%
Average transfer	9.4%	14.3%	9.9%	11.2%	13.7%	6.4%	

Note: Boldface indicates important candidates.
Percentages with asterisks are the largest in their rows.

A left-right line—among all the candidates shown in column 1 of table 6.7, it would go from top to bottom—is constructed as follows. The three candidates receiving the highest average transfers are singled out: Jospin, Bayrou, and Chirac. One is the principal candidate of the left, one of the center, and one of the right. The statistics show that Jospin voters favor Bayrou (8%) over Chirac (5%) and Chirac voters favor Bayrou (24%) over Jospin (5%), so Bayrou must be placed in the center if one seeks order along a line consistent with single-peaked transfers. The choice of Jospin on the left and Chirac on the right is arbitrary (but in keeping with the political meaning of the directions). With sixteen candidates, the average transfer is $100/15 \approx 6.7\%$; the important candidates are those with a larger average, except for Le Pen (important because he survived the first round).

A candidate is classified to the left (to the right) if her voters transfer to the principal candidate of the left more than (less than) to the principal candidate of the right. So, for example, Laguiller is to the left (20% to Jospin, 7% to Chirac), and Madelin is to the right (4% to Jospin, 32% to Chirac). The result—Gluckstein through Chevènement to the left, Chirac through Mégret to the right—is consistent with the media's and the generally accepted classification of the candidates. The precise order of all the candidates on the left-right line can be found in various ways and so relies to a certain extent on an arbitrary rule. In this case, Chevènement is closer to Bayrou than Jospin and thus is classified as center left. To the left of Jospin the candidates are listed according to increasing total transfers to the two center candidates: Gluckstein gives them 7%, Laguiller 13%, . . . , Taubira 38%. To the right of Chirac the candidates are listed according to decreasing total transfers to the two center candidates: Madelin 31%, Lepage 29%, . . . , Mégret 16%. The single peak in the rows is accompanied by a single peak in the columns among the important candidates: in Bayrou's row the percentages among the important candidates decrease the further they are from Bayrou, and the same is true for Bayrou's column. In the rows of the unimportant candidates a single-peaked property holds: Madelin's peak is Chirac, and the further away, the smaller the percentage. When it is possible to achieve single peaks in rows and columns, the order is clearly unique: it defines the *statistical left-right spectrum.*

It is amusing to note that if the left-right line were as shown in table 6.7 and the median-voter mechanism proposed in Moulin (1980) were applied to the national vote (see table 2.9), Bayrou would have missed being the winner in 2002 by a hair; Bayrou's vote plus that of the candidates to his left was 49.7%, so Chirac would have been the winner. Were Lepage (with her 1.9% of the vote) classified between Bayrou and Chirac—not unreasonable in view of the transfers of her voters to candidates of the left—the median-voter mechanism would

have elected her. In 2002 the election of Bayrou with 7% of the first-round votes or of Lepage with 2% is unacceptable: neither was a major candidate.[3]

The estimates of transfers in 2007 are given in table 6.8 and are computed for, say, Bayrou as follows. Among the ballots whose highest grade goes to Bayrou, either $k \geq 1$ other candidates are given the same grade or Bayrou is the only candidate with that grade and there are $k \geq 1$ candidates who are given the next highest grade. Attribute $1/k$ to each of the other candidates in either case, and find the sum accorded to each candidate. The estimate of the transfer to a candidate is the percentage his sum represents of the total sum. Exactly the same rules are used to determine candidates of the left and the right and the order among them. In this case, there is only one center candidate, Bayrou. The left (from Buffet to Royal) and the right (from Sarkozy to Le Pen) correspond to the usual media designations. Important candidates are Le Pen and those whose average transfer is above $100/11 = 9.1\%$. Once again, *statistically, the voters' transfers are almost single-peaked among the important candidates*. The single peak in the rows (one small exception for the important candidate Le Pen) is accompanied by a single peak in the columns (two small exceptions, Besancenot and Le Pen). In two practical political situations *single-peaked transfers are real when seen in terms of probabilities.*

More strikingly than in 2002, the bulk of the transfers go to the important candidates: to Besancenot (far left), to Royal (moderate left), to Bayrou (center), and to Sarkozy (right). In fact, if the grades are used as determinants of preference among the three major candidates, 4.1% expressed the preference Royal $\succeq$ Sarkozy $\succ$ Bayrou and 5.8% Sarkozy $\succeq$ Royal $\succ$ Bayrou. So 90.1% of the ballots agree with the single-peaked preferences hypothesis on the left-right line among the three, going from Royal to Bayrou to Sarkozy (though among more candidates this is not true).

Not surprisingly, Bayrou is the choice of the median-voter nationally with respect to the left-right line of table 6.8: Bayrou's vote plus that of the candidates to his left was 57.7%; his vote plus that of the candidates to his right was 60.9%. These numbers are close to the estimates that are available of face-to-face confrontations with Royal (thus against the left) and with Sarkozy (thus against the right). A poll taken two days before the election shows the same

3. The 2002 Orsay experiment allows estimates to be made of the face-to-face races. To compute the estimate between two candidates, a vote is given to one whenever he is given a cross and the other is not. Jospin (19.5%) and Bayrou (9.9%) did better in the Orsay official vote than nationally, and Le Pen (10.0%) did worse. The estimates show Jospin winning against Chirac (with 53%), Bayrou (with 56%) and Le Pen (with 75%); Chirac winning against Bayrou (with 54%) and Le Pen (with 80%); Bayrou winning against Le Pen (with 74%). Jospin is at once the Condorcet-winner and the Borda-winner, and the Condorcet- and Borda-rankings are the same as well: Jospin $\succ_S$ Chirac $\succ_S$ Bayrou $\succ_S$ Le Pen.

Table 6.8
Estimated Transfers of Votes to Important Candidates, Based on 2007 Orsay Experiment

	Left ⟵ Besancenot	Royal	Bayrou	Sarkozy	⟶ *Right* Le Pen	Seven Others
Buffet	28%*	24%	5%	2%	2%	39%
Laguiller	32%*	17%	14%	11%	3%	23%
Bové	17%*	13%	15%	9%	2%	44%
Schivardi	29%*	11%	17%	5%	8%	30%
Besancenot	–	**26%***	**18%**	**3%**	**2%**	51%
Voynet	13%	34%*	24%	9%	1%	19%
Royal	**11%**	–	**44%***	**10%**	**1%**	35%
Bayrou	**6%**	**34%**	–	**36%***	**2%**	22%
Sarkozy	**2%**	**15%**	**43%***	–	**12%**	28%
Nihous	14%	11%	18%	19%*	7%	31%
de Villiers	2%	4%	9%	60%*	11%	14%
Le Pen	**4%**	**8%**	**6%**	**38%***	–	44%
Average transfer	14.5%	18.1%	19.5%	18.4%	4.5%	

Note: Boldface indicates important candidates.
Percentages with asterisks are the largest in their rows.

Table 6.9
Transfers of Votes to Important Candidates, Polling Results, April 20, 2007

	Left ⟵ Besancenot	Royal	Bayrou	Sarkozy	⟶ *Right* Le Pen	Eight Others	Not Counted
Royal	12%	–	34%*	5%	4%	27%	18%
Bayrou	7%	28%*	–	25%	3%	15%	22%
Sarkozy	3%	10%	37%*	–	7%	19%	24%
Le Pen	1%	12%	8%	31%*	–	23%	25%

Source: Polling Results by BVA.
Note: Percentages with asterisks are the largest in their rows. To compare them with the percentages in table 6.8, the percentages in this table must be normalized.

qualitative results (though they are national estimates, not Orsay estimates; see table 6.9).

6.3 Borda's and Condorcet's Bias for the Center

It is striking that in the 2007 election (for which there is so much polling and experimental evidence), the Condorcet-winner and the Borda-winner—those centuries-old opposing concepts—are consistently one and the same candidate (Bayrou). Why? The evidence of tables 6.7 and 6.8 suggests that when there is a statistical left-right spectrum, a voter's second choice is most likely to be a major candidate (protest voters are an exception). So if there are two major candidates,

each beats a minor candidate by a large margin, the Condorcet-winner is also the Borda-winner and must be one of the two major candidates. When there are three major candidates, the Condorcet- and Borda-winners are again limited to those three candidates, and the analysis may be restricted to them.

Suppose three candidates on the left-right spectrum, A the major candidate of the left, B the major candidate of the center, and C the major candidate of the right, are respectively the favorites of $x_A\%$, $x_B\%$, and $x_C\%$ of the voters. Let γ be the probability that a voter who prefers C votes for B when B opposes A; α be the probability that a voter who prefers A votes for B when B opposes C; and β be the probability that a voter who prefers B votes for A when A opposes C. The expected pair-by-pair votes are given in table 6.10.

There is a *statistical left-right spectrum* if the matrix of transfers is single-peaked in rows and in columns. With three candidates this occurs when $\alpha > \frac{1}{2}$, $\gamma > \frac{1}{2}$, and $\max\{\beta, 1-\beta\} < \min\{\alpha, \gamma\}$, which holds when β is close to $\frac{1}{2}$. When, in addition, α and γ are sufficiently large, implying $v_{BA} > v_{CA}$ (B against A wins more votes than C against A), and symmetrically, $v_{BC} > v_{AC}$, there is a *strong statistical left-right spectrum*, meaning that the matrix of pairwise votes is single-peaked in rows and in columns. In this case, it is easy to prove that there always exists a Condorcet-winner. The experimental evidence from both the 2002 and 2007 Orsay experiments shows that a strong statistical left-right spectrum existed among the three principal candidates (table 6.11).

When the centrist candidate B is the Condorcet-winner, then B is necessarily the Borda-winner. For $v_{BC} > v_{AC}$, and $v_{BA} > 50\%$ implies $v_{AB} < 50\%$, so $v_{BA} + v_{BC} > v_{AB} + v_{AC}$, and the symmetric argument gives the same

Table 6.10
Pair-by-Pair Votes among Three Candidates

vs.	*Left* ⟵ *A*	*B*	⟶ *Right* *C*
A	–	$v_{AB} = \big(x_A + (1-\gamma)x_C\big)\%$	$v_{AC} = (x_A + \beta x_B)\%$
B	$v_{BA} = (x_B + \gamma x_C)\%$	–	$v_{BC} = (x_B + \alpha x_A)\%$
C	$v_{CA} = \big(x_C + (1-\beta)x_B\big)\%$	$v_{CB} = \big(x_C + (1-\alpha)x_A\big)\%$	–

Table 6.11
Strong Statistical Left-Right Spectrum, Pairwise Votes, 2002 and 2007 Orsay Experiments

2002	Jospin	Chirac	Le Pen	2007	Royal	Bayrou	Sarkozy
Jospin	–	56%	75%	Royal	–	44%	52%
Chirac	47%	–	80%	Bayrou	56%	–	60%
Le Pen	25%	20%	–	Sarkozy	48%	40%	–

Table 6.12a
Number of Wins among Royal, Bayrou, and Sarkozy Only, 2007 Orsay Experiment

	Left ⟵ Royal	Bayrou	⟶ *Right* Sarkozy	Tie	Cycle
First-past-the-post winner	4,274	1,772	3,574	380	–
Two-past-the-post winner	3,410	4,671	1,225	694	–
Majority judgment-winner	1,462	7,573	956	9	–
Condorcet-winner	772	8,894	65	246	23
Borda-winner	369	9,526	67	38	–

Note: Ten thousand samples of 101 ballots, which were drawn from 1,733 ballots. "Cycle" indicates a Condorcet paradox.

Table 6.12b
Number of Wins among All Candidates, Winner Always Royal, Bayrou, or Sarkozy, 2007 Orsay Experiment

	Left ⟵ Royal	Bayrou	⟶ *Right* Sarkozy	Tie	Cycle
First-past-the-post winner	2,324	2,260	5,379	37	–
Two-past-the-post winner	3,175	5,830	801	194	–
Majority judgment-winner	1,290	7,756	943	11	–
Condorcet-winner	623	9,152	5	184	36
Borda-winner	348	9,639	0	13	–

Note: Ten thousand samples of 101 ballots, which were drawn from 1,733 ballots. "Cycle" indicates a Condorcet paradox.

conclusion when comparing B with C. On the other hand, if A is the Condorcet-winner, then the Borda-winner is either A or B. For $v_{BC} > v_{AC} > 50\% > v_{CB}$ and $v_{BA} > v_{CA}$ imply that C cannot be the Borda-winner; and symmetrically, if C is the Condorcet-winner, then the Borda-winner is either C or B. So the Borda-winner favors the centrist candidate more than does the Condorcet-winner. However, with only a statistical left-right spectrum it is entirely possible for the Condorcet paradox to occur in theory (by varying the data in table 6.10) and in practice, as the experimental evidence shows (see tables 6.12a, 6.12b, 6.14a, 6.14b).

Evidence from the 2007 Orsay experiment supports these arguments and observations. Two sets of independent random drawings were made. In one, 10,000 samples from 101 ballots were drawn from the 1,733 valid ballots in order to compare the behavior of the principal methods applied only to the three major candidates, Bayrou, Royal, and Sarkozy (table 6.12a). In the other, conducted separately, 10,000 random samples from 101 ballots were drawn to compare the behavior of the principal methods applied to all the candidates

(table 6.12b). In every case one of the three major candidates was the winner. To compute the winners by one or another of the methods a candidate was accorded the vote of a ballot if she had the highest grade; when there was a tie among k candidates for the highest grade on a ballot, each was attributed $\frac{1}{k}$.

The results in tables 6.12a and 6.12b clearly show that as one passes from one method to another down the list—from first-past-the-post to Borda—the centrist candidate is more and more favored. Borda's method favors the centrist candidate Bayrou slightly more than Condorcet's, and Condorcet's much more than the majority judgment. At the opposite end of the spectrum, the first- and two-past-the-post methods disfavor the centrist candidate in comparison with the majority judgment. The nine and eleven ties in the majority judgment mean ties in the majority-gauge (not the majority-value): a 0.001 probability of a tie with only 101 voters is sufficiently small. The twenty-three and thirty-six occurrences of the Condorcet paradox show that the preferences among the candidates is not single-peaked and that though there is a statistical left-right spectrum, it is not strong. One of the twenty-three Condorcet paradoxes of table 6.12a showed Bayrou with 59% against Sarkozy, Sarkozy with 52.5% against Royal, and Royal with 52% against Bayrou. The striking contrast between tables 6.12a and 6.12b is the large increase in Sarkozy first-past-the-post wins when there are twelve candidates rather than three: it reflects a large number of occurrences of Arrow's paradox coming from the dispersion of votes among candidates of the left. And, of course, the more candidates there are, the more Borda favors the centrist. The majority judgment is unaffected by the number of candidates: the small differences are due to the independently drawn samples.

The official first-round votes in the three precincts of the 2007 Orsay experiment were quite different from the official first-round votes nationally (table 6.13). In particular, Royal's 29.9% in Orsay was above her 25.5% nationally, Bayrou's 25.5% in Orsay was much above his 18.6% nationally, Le Pen's 5.9% in Orsay much below his 10.4% nationally. So it is no surprise to see Bayrou—the choice of the median-voter in the official first-round vote in the

Table 6.13
Official First-Round Votes, National and Three Precincts of Orsay, April 22, 2007

	Sarkozy	Royal	Bayrou	Le Pen	Besancenot	de Villiers
National	31.2%	25.9%	18.6%	10.4%	4.1%	2.2%
Orsay	29.0%	29.9%	25.5%	5.9%	2.5%	1.9%
	Buffet	**Voynet**	**Laguiller**	**Bové**	**Nihous**	**Schivardi**
National	1.9%	1.6%	1.3%	1.3%	1.1%	0.3%
Orsay	1.4%	1.7%	0.8%	0.9%	0.3%	0.2%

Table 6.14a
Number of Wins among Royal, Bayrou, and Sarkozy Only, 2007 Orsay Experiment

	Left ⟵ Royal	Bayrou	⟶ *Right* Sarkozy	Tie	Cycle
First-past-the-post winner	1,678	42	8,089	191	–
Two-past-the-post winner	2,145	820	6,470	565	–
Majority judgment-winner	1,288	4,001	4,701	10	–
Condorcet-winner	671	6,462	1,993	669	205
Borda-winner	484	7,109	2,225	182	–

Note: Ten thousand samples of 101 ballots, which were drawn from a sample of 501 ballots representative of the national vote. The same approach was used as in estimating first-round results on the basis of majority judgment ballots; the percentage of votes of each candidate in the sample of 501 ballots came close to that of the candidate's national vote. In this sample, Sarkozy had 30.7%, Royal 25.5%, and Bayrou had 18.7%. (Le Pen 9.3%.)
"Cycle" indicates a Condorcet paradox.

Table 6.14b
Number of Wins among All Candidates, Winner Always Royal, Bayrou, or Sarkozy, 2007 Orsay Experiment

	Left ⟵ Royal	Bayrou	⟶ *Right* Sarkozy	Tie	Cycle
First-past-the-post winner	2,061	50	7,874	15	–
Two-past-the-post winner	2,174	716	6,731	379	–
Majority judgment-winner	1,309	4,034	4,649	8	–
Condorcet-winner	616	6,538	2,002	630	214
Borda-winner	354	9,608	26	12	–

Note: Ten thousand samples of 101 ballots, which were drawn from a sample of 501 ballots representative of the national vote.
"Cycle" indicates a Condorcet paradox.

voting precincts of the Orsay experiment and in the nation—so often the winner (see tables 6.12a and 6.12b). Accordingly, parallel sets of independent random drawings were made from a subset of 501 ballots (of the 1,733 valid ballots) whose estimated first-round votes were representative of the national vote. In one, 10,000 samples from 101 ballots were drawn from the 501 to compare the methods applied to the three major candidates (table 6.14a); in the other, conducted separately, 10,000 samples from 101 ballots were drawn from the 501 to compare the methods applied to all the candidates (table 6.14b).

The results show more dramatically how Borda's method and to a lesser extent Condorcet's method favor the centrist candidate and how the first- and two-past-the-post methods penalize him, while in contrast the majority judgment appears to be more evenhanded. Note, in particular, the chaotic

behavior in the centrist's Borda-wins when there are twelve rather than only three candidates.

This has practical significance: most thoughtful commentators reject election mechanisms that *systematically* elect the centrist candidate. As the well-known popularizer of science William Poundstone wrote, "We want a system that doesn't *automatically* exclude [moderate] candidates from winning. We also want a system that doesn't make it easy for any goof who calls himself a moderate to win" (2008, 211). On the other hand, the fact that Bayrou had merely forty-two and fifty wins with first-past-the-post when by all reasonable estimates Bayrou was the Condorcet- and Borda-winner seems derisory. *A good election mechanism should eliminate extremes and give all major poles—left, center, and right—a fighting chance to win.*

To this day, the Condorcet-winner and the Borda-ranking dominate the thinking in the theory of social choice: they continue to be proposed and reproposed, alone and in combinations. Agreement between them would therefore seem to be a happy concurrence giving a particularly valid result. But both of these mechanisms are heavily biased in favor of moderate candidates. Major candidates of the right and the left, such as Sarkozy and Royal, often elicit strong support and strong opposition, so they are given high or low evaluations. A moderate candidate, on the other hand, is often placed second or third. Face-to-face confrontations and rank-orders ignore how voters evaluate the respective candidates (just as the 2002 French presidential runoff merely compared Chirac and Le Pen but did not evaluate them).

The ballots of the 2007 Orsay experiment show that these evaluations are significant: two-thirds of the second highest grades are merely *Good* or worse, three-quarters of the third highest grades are *Acceptable* or worse (see table 6.15). Both Condorcet and Borda ignore evaluations; they rely only on comparisons. When there are twelve candidates, a voter's list gives 11 points to the first candidate, 10 to the second, 9 to the third, and so on. The difference between being first, second, or third on the list is marginal, especially in the presence of many candidates. Perhaps this exaggerated bias in favor of moderate candidates explains why these mechanisms are hardly ever used in practice.

Table 6.15
Distributions: Highest Grade, Second Highest Grade, Third Highest Grade, Three Precincts of Orsay, April 22, 2007

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
Highest	52%	37%	9%	2%	0%	1%
Second highest	–	35%	41%	16%	5%	3%
Third highest	–	–	26%	40%	22%	13%

Chapter 19 compares the qualitative properties of methods in more detail. The majority judgment results are given here only as a point of comparison. The first- and two-past-the-post systems both systematically eliminate centrist candidates (when they exist). France and the U.K. are excellent examples. The centrist candidate Bayrou did not survive the first round in 2002 or 2007, either in the precincts of the Orsay experiments or in the entire nation. In the British general elections of 2005 the Liberal Democrats won only 9.6% of the seats for 22% of the votes, in those of 2010 they won but 8.8% of the seats for 23% of the votes.

6.4 Conclusion

More experimentation is called for. Nevertheless, the experimental evidence already shows the following:

· The inputs imputed to voters in the traditional theory of social choice—relative comparisons of candidates or rank-orders of candidates—are completely unrealistic. Voters do not think in those terms and do not wish to express themselves in those terms.

· Most voters—three-fifths of them—refuse to rank-order all candidates when they are asked to do so.

· Many voters—one-third of them—refuse to single out one preferred candidate when they have the opportunity to give the same evaluations to more than one candidate.

· Many voters—one-half of them—refuse to declare their preferred candidate *Excellent*. To be first in a rank-order has very different meanings, so aggregating rank-orders is meaningless.

· The fact that there is a statistical left-right spectrum according to which the preferences of many voters are single-peaked shows why the Borda- and Condorcet-winners are often the same and why both of these mechanisms (more so Borda's) strongly favor centrist candidates.

· "The free communication of thoughts and opinions is one of the most precious rights of man." (*Déclaration des droits* 1789, article 11). Not one of the electoral systems used in practice—whether it be the Australian rank-order or the one vote allowed in England, France, and the United States—gives voters anywhere near the freedom of expression they wish to have.

The traditional model of social choice has been tried in theory and in practice and does not work. By Richard Feynman's definition of science, it must be eliminated.

7 Judging in Practice

A thing may look specious in theory, and yet be ruinous in practice; a thing may look evil in theory, and yet be in practice excellent.
—Edmund Burke

Mais là où les uns voyaient l'abstraction, d'autres voyaient la verité. (But where some saw abstraction, others saw truth.)
—Albert Camus

"We're ranking everybody," said the playwright Arthur Miller, "every minute of the day": economists and peace-makers, mathematicians and physicists, novelists and journalists, students and professors, divers and skaters, beauty queens and muscle-men, cities and countries, hotels and restaurants, movies and theatrical performances, hospitals and universities, wines and cheeses. To rank these and many other competitors, accomplishments, endowments, performances, goods, or services is fraught with differences of opinion among the judges—or conflicting appreciations of their characteristic attributes—that must be reconciled into the verdict of a jury.

Athletes compete for glory (and money) at Olympic games; chess and go players do, too, though elsewhere. Wines and cheeses compete for prizes and other accolades of excellence in trade fairs. Flautists, pianists, and violinists compete for international, national, and regional prizes. Ranking and designating the winner among runners, high-jumpers, chess champions, and go players is simple enough: time distinguishes among runners, height among jumpers, and the winners and losers between pairs of chess and go players are obvious (though how to rank many players is not).

A heated argument between two (wonderful but fictional) late-eighteenth-century zoologists over emotion and its expression in animals was described, in which one says to the other: "How can [emotion] be measured? It cannot be measured. It is a notion; a most valuable notion, I am sure; but, my dear sir,

where is your measurement? It cannot be measured. Science is measurement—no knowledge without measurement" (O'Brian 1969, 380). Similarly, there is no obvious scale by which to measure most goods, services, performances, or accomplishments. But *in practice* measures are routinely invented and defined by means of which judges decide which are to be the winners and losers, and what are to be the rankings.

7.1 Students

Students are regularly ranked at all levels. Today, more often than not, their examinations, essays, and class performances are graded, and the grades determine their ranking. Candidates for positions—in the civil services, as qualified medical doctors, as lawyers admitted to the bar, as students wishing to pursue medicine, the law, the sciences, or any discipline—have from time immemorial been examined, and ranked or graded by groups of individuals. Important examinations—the baccalauréat in France, college aptitude tests in the United States, A-levels in Great Britain, bar examinations, doctoral qualifying examinations—occupy an important place in the psyche of those who have suffered through them. This must explain why systems that are devised to find rankings so often adopt the "language" with which their inventors were familiar in their youth. These are, among many others, important examples of a method of social choice—familiar to all from their very first memories of classroom grades—that is neither voting nor a market mechanism.

"In China since medieval times, imperial dynasties, gentry-literati elites, and classical studies were tightly intertwined in the operation of the civil service examinations. All three dimensions were perpetuated during the late empire (1368–1911), and they stabilized for five hundred years because of their interdependence . . . [B]oth local elites and the imperial court continually influenced the government to reexamine and adjust the classical curriculum and to entertain new ways to improve the institutional system for selecting those candidates who were eligible to become officials" (Elman 2000, xvii, xxiii–xxiv). So begins a magisterial study of an extremely elaborate system for examining and ranking candidates seeking positions in the Chinese civil service and, more generally, both social and economic importance in society.

The system originated in the Sui Dynasty (581–618) and evolved over time. By 1065 examinations were held every three years, in three levels, at fixed dates: provincial Autumn examinations with results given in the Laurel list, followed first by state Spring examinations with results given in the Apricot list, then a month later by palace examinations with results given in the Golden list. Their importance is measured by an incident of 1397. The first emperor of

the Ming Dynasty, angered because no candidates from the north had survived to the palace examinations, demanded a change in the results. Refused by the chief examiner on the grounds that the evaluations had been strictly anonymous, the emperor ordered new evaluations by new officials. They reported the same results and were promptly put to death.

The system had to be elaborate in view of the numbers of candidates in local, provincial, and metropolitan examinations. In the metropolitan examinations of 1742, four chief and eighteen associate examiners had to evaluate 5,993 candidates, of whom they retained 319 (Elman 2000, 680). To begin, associate examiners accepted or rejected—using such phrases as "deep thoughts, rich in force, sufficient life, and perspicacious," or when there was little time, "studies that have a base," "lack of subtlety," or "correct but ordinary" (Elman 2000, 426; Zi 1894, 152)—then ranked the surviving candidates. "Exactly how the rankings were reached is unclear, but grading was linked to examiner comments, not scores . . . The rankings served as the scores. It seems to be more like a process of elimination within a fixed quota."[1] An earlier scholar hints at the possibility of a rudimentary system where each of eight examiners placed a circle on excellent exam copies, points on less than good ones, or combinations of them, so that "eight circles" was the highest accolade (Zi 1894, 198). Another account tells of up to six circles being given for each answer, the total number of circles determining the ranking of the candidates.[2] There is often a dearth of precise information concerning the mechanics of grading and ranking in the historical records.

Gaspard Monge—mathematician; founder of the École Polytechnique; inventor of descriptive geometry; a favorite of Napoléon, whom he accompanied to Egypt; president of General Bonaparte's Institut d'Egypte; Minister of the Navy for eight months (1792–1793), and in that capacity the man who witnessed and signed the official *procès verbal* of Louis XVI's decapitation (Balinski 1991)—appears to be one of the first to have developed and used a system for grading. Monge, who had been elected to the Royal Academy of Sciences in 1772, was appointed examiner of the navy in October 1783, a position which he continued to occupy until 1790. He had to examine the students at two academies in the provinces—to which later were added others in the ports of Brest, Rochefort, and Toulon—to determine "those students who will have

1. B. A. Elman, private communication, January 11, 2006.

2. We are indebted to Wanyan Shaoyuan for some of this information, cited from two recent Chinese publications, Huang Mingguang, *Studies of the Imperial Examination System in the Ming Dynasty*, Guangxi Normal University Press, 2006, and Wang Kaixuan, *Research and Discourses on the Imperial Examination System in the Ming Dynasty*, Shengyang Press, 2005.

satisfactorily mastered the required parts of the program, who will immediately be sent to the ports after their examination and admitted as cadets third class of the navy with a salary of 300 pounds per year" (Julia 1990). In 1789 he orally examined thirty-eight candidates at the academies; by 1790 the number had grown to seventy. Monge recorded letter grades in his private notebook—going from *a* to *g*, though no *a* may be found in it—for each of the courses of study, sometimes mixing them by writing "*c–d*" and "*g–h*." To these he added descriptions of the quality of presentations. He also sought to assess the intelligence and character of the candidates, using words such as "promising," "very promising," "fairly intelligent," "very intelligent," "ordinary," and "slow witted" for the first; and for the second, "reasonable," "thoughtful," "lively," "bold," and "light." He was required by the ministry to rank-order all thirty-eight candidates, but there is no record as to how he integrated the letter grades and the verbal descriptions of intelligence and character into a single list.

Grades come in a myriad of scales and have changed over time. United States universities are a case in point. "As for grading, while some system of grading was implied in the ranking of seniors for commencement parts in the colonial colleges, 'the initiative in attempting to formulate a scale for grading students' was not taken until 1813 at Yale, which adopted a numerical scale of four for evaluating course work. The numerical scale took the place of four terms that had been used as early as 1783 . . . : *optimi, second optimi, inferiores (boni),* and *pejores*" (Rudolph 1977, 147). Harvard adopted a scale of 100 in 1879, replaced it with five letter grades from *A* to *E* in 1883, and chose a scale of three verbal scores, "passed with distinction," "passed," and "failed," in 1895. The pioneering historian of U.S. colleges and universities, Frederick Rudolph, goes on to remark, "All this thrashing around in search of the perfect grading system was a response to a changing curriculum and a changing climate of academic life. Examining and grading systems were barometers of curricular health and style and purpose. In adopting the numerical scale, Harvard stressed competition as an inducement to student effort. Letter grades reduced competitive pressures and deemphasized class rank, which now could not be calculated" (147). He adds elsewhere that changes in grades also reflected the changing moods of the country: the election of Andrew Jackson—"that unschooled orphaned soldier" (Rudolph 1991, 201)[3]—and the spirit of Jacksonian democracy, hostile to privilege in all its forms, had discouraged the practice of ranking students at all.

3. John Quincy Adams, Jackson's immediate predecessor, decried Harvard's award in 1833 of an honorary degree to the "barbarian" Jackson as a "disgrace." This opinion was so widely shared that Harvard abstained from so honoring another president for almost forty years.

Today grades are ubiquitous. In some countries a uniform standard prevails. In France it goes from a bottom of 0 to a top of 20; 10 is the threshold for passing, 12 is *Assez Bien* (quite good), 14 is *Bien* (good), and 16 is *Très Bien* (very good). In Denmark ten numbers were used from 1963 to 2003: 13 the best, 0 the worst, 6 the lowest passing grade (the numbers 1, 2, 4, and 12 were omitted). In the Czech Republic excellent is 1, unsatisfactory 5; in Poland, 6 is excellent, 1 unsatisfactory in schools, and 5 excellent, 2 unsatisfactory in universities. In the United States the scale varies, but letter grades going from *A*, the best, to *E* or *F*, the worst, dominate the older 100 to 0 scale, with the following rough description: *A* exceptional, over 90; *B* good, 80 to 90; *C* fair, 70 to 80; *D* poor, 60 to 70; and *E* or *F*, fail, below 60. On the other hand, averages of letter grades are often computed by assigning 4 to *A*, 3 to *B*, 2 to *C*, 1 to *D*, and 0 to fail, with a + or − adding or subtracting 0.3 or one-third. Canadian universities offer an astonishing variety, including, in order of worst score to best, 0 to 4, to 4.3, to 4.33, to 4.5, to 9, to 10, to 12, and to 100. German universities use words (with numerical counterparts going in the opposite direction and gradations of ± 0.3): *sehr gut* (excellent, 1), *gut* (good, 2), *befriedigend* (satisfactory, 3), *ausreichend* (sufficient, 4), and *mangelhaft* (deficient, 5 or 6).

In every case where numbers or symbols such as letters come to be used, they are given careful definitions *in words*; they are given absolute meanings. A particularly good example is the set of definitions that were given for the marks used in Denmark from 1963 to 2003:

13: exceptionally independent and excellent,

11: independent and excellent,

10: excellent, not particularly independent,

9: good, a little above average,

8: average,

7: mediocre, slightly below average,

6: just acceptable (lowest passing grade),

5: hesitant, not satisfactory,

3: very hesitant, very insufficient, unsatisfactory,

0: completely unsatisfactory.

The moral seems to be that any scale will do, for it is defined and in turn learned and better understood through use, just as any language is learned and better understood through use, so in practice the numbers take on meanings of their own. However, as will be seen anon, *any scale will not do* if its sums and averages are to be considered meaningful.

What mechanism of social choice should be used to fuse the grades assigned to a candidate or a student by different examiners or professors of different disciplines into one grade has had a well nigh unanimous answer: calculate the average values of the grades, or calculate the averages of the grades weighted by a factor of their importance. For example, students at France's prestigious École Polytechnique are ranked at graduation on the basis of the averages of all their weighted grades, and U.S. universities elect students to Phi Beta Kappa and bestow the distinctions of *magna* or *summa cum laude* on the basis of the students' grade averages (sometimes with supplementary evidence such as letters of professors or difficulty of programs).

Two well-known French psychometricians, Henri Piéron and Henri Laugier, are generally credited with having founded *la docimologie*, the science of examinations, *doci* coming from the Greek *dokimé,* meaning test, which enjoyed a certain vogue in the 1920s but did not survive as a focused discipline (Martin 2002). The docimologists analyzed grades given in examinations and showed that significant variations in the results came from the varying "subjective coefficients" of the different examiners: some are severe and others generous, some assign widely varying marks, others not. Thus the grades of a student sitting an examination together with many others (such as the national baccalauréat in France) depend on who does the grading. Piéron and Laugier contested the idea of using average scores. They were right: the clear solution is to have a jury of several examiners assign grades and to use the majority-grade (though having more examiners obviously implies greater expense).

Piéron noted that electors called on to examine candidacies for public office, committees of clubs charged with designating new members, and university professors deciding whom to appoint as new faculty face exactly the same problems as do examiners (Piéron 1963, 53–54). He clearly recognized that the problem of grading and ranking students is a problem of social choice.

7.2 Employees

Employees are evaluated, sometimes they are ranked, and many firms distribute year-end bonuses on the basis of their performances. "Forced ranking" became a hotly debated innovation with the publication of General Electric's annual report in 2000. In it their corporate executive office, led by chief executive officer Jack Welch, declared, "In every evaluation and reward system, we break our population down into three categories: the top 20%, the high-performance middle 70%, and the bottom 10%. The top 20% must be loved, nurtured and rewarded in the soul and wallet because they are the ones who make magic happen ... The top 20% and middle 70% are not permanent labels. People

move between them all the time. However, the bottom 10%, in our experience, tend to remain there. A Company that bets its future on its people must remove the lower 10%, and keep removing it every year—always raising the bar of performance and increasing the quality of its leadership."

Forced ranking or *forced* distribution rating systems—"rank and yank"—refer specifically to the idea that the bottom 10% must either be let go with a severance package or be put on notice to do better within three months and face severance with no parting gift. Many companies are said to use it (or to have used it), including Lucent, the infamous Enron, Ford, General Motors, Goodyear, and Microsoft, though instead of a 20%–70%–10% formula some practice(d) 10%–80%–10%, others 25%–25%–25%–25%.

Class-action suits were filed against each of the last four firms, claiming their systems discriminated against blacks, women, older workers, or noncitizens. In the suit against Ford, 500 employees accepted a $10.5 million settlement. Dick Grote, a strong proponent of the system, claimed (in 2003) that forced ranking is "probably the most controversial issue in management today." It is generally acknowledged that about one-quarter of the Fortune 500 companies use a performance management system based on forced ranking, though increasingly companies declare they do not or refuse to speak about it; it does not have the image of a loving, nurturing system, and it has led to costly suits and unflattering publicity.

A recent article investigates the extent to which introducing such systems can be expected to improve the quality of the workforce (Scullen, Bergey, and Aiman-Smith 2005). It leaves aside how morale, profitability, and productivity may be affected, concentrating on "performance potential" as a function of the percentage fired, the managers' ability to judge performance, the quality of personnel selection procedures, and the usual levels of turnover. It concludes that such systems definitely "hold promise" but that after impressive gains in the first few years, the expected improvement goes to zero.

The important point is that organizations rank and classify employees in terms of their past and expected future performances. They may force-rank or simply wish to determine bonuses and raises. So methods are needed for ranking employees, deciding on raises, and distributing bonuses among them from a presumably fixed pool of money. On the other hand, nothing of the sort may be found in the main book on the subject, issued by the Harvard Business School Press (Grote 2005). A publication clearly designed to sell the consulting services of its author—"It's about jump-starting a leadership development process . . . It's about understanding the depth of your talent pool and seeing where the true leadership potential lies in your company . . . It's about talent management . . . it's about grooming great leaders. This book will help you raise the bar

and lift the boat" (xi–xii)—it repeats checklists of the obvious questions that should be asked, and recommends that managers should be trained to the task by an outside facilitator who should also chair the committees that establish the rankings. Nothing is said about *how* the opinions of many managers are to be reconciled to arrive at a company decision. By implication the decisions are made by some sort of collective consensus (which more often than not must mean that one or two individuals impose their views).

French jurisprudence upheld the right of companies to rank (but clearly not, in a country known for its social support system, to yank) in 2002, stating (in full): "Ranking systems permit fixing salary increases as a function of the relative performance of employees and positioning them according to pre-established, known, objective and controllable criteria; the individual performance of employees is appreciated in comparison with the performances of employees occupying comparable positions. The classifications, brought to the knowledge of the employees beforehand, are neither subjective nor discriminatory. Therefore, including a [ranking system] in the interior regulations [of a firm] to classify an employee at the lowest performance level, in conformity with rules established in a manner known to all and that constitutes a licit individualization of salary increases, is to be allowed" (*Répertoire de jurisprudence* 2002).[4] How the French court decided that *any* ranking system is objective and not subjective, known to all, controllable, and not discriminatory, is a total mystery.

7.3 Musicians

The University Interscholastic League was created by the University of Texas at Austin in 1909 to provide educational extracurricular academic, athletic, and music contests to high schools. Their Web site states, "The initials UIL have come to represent quality educational competition administered by school people on an equitable basis" (University Interscholastic League 2006). Written rules govern all their contests. Original essays are written for the "ready writing" contest: the regulations state that judges are to rank-order them by consensus. On the other hand, a detailed procedure is given for scoring and ranking competing marching bands in preliminaries and finals. There are three music judges, and two marching judges. Each ranks every competing band in order; the top-ranked band is assigned 1 point, the next 2 points, and so on; a band's score is the sum of its points. The bands are ranked according to their scores, the one with the

4. The original French text is everywhere as stilted as the English translation.

lowest score the winner, the one with the highest score last (except that in the finals a band ranked first by two music judges and one marching judge is immediately declared the winner—to be certain that if a band is judged best by a qualified majority then it must win—and the others are ranked as described). This is nothing but an equivalent, mirror image of Borda's method (and one of the very rare instances where Borda's method is actually used), with low scores good and high scores bad, together with a provision to assure that if a (qualified) Condorcet-winning band exists, it wins. In addition, the UIL demands a strict ordering with no ties. If two or more bands have the same score, the procedure says the same rule is to be applied to them alone to decide how they should be ranked among themselves (and if ties occur again, then they are to be broken with still another application of the rule). Regrettably, this recipe does not (always) work. When an even number of bands are tied, the rule cannot give the same score to all of them, so either the bands are strictly ordered or some are still tied. However, when an odd number of bands are tied, they can have the same score. For example, suppose that in a competition among twenty bands A to T every one of the five judges J_i consistently placed three of them, A, B and C, in one of the first three places, as follows:

$$J_1 : A \succ B \succ C \succ \cdots \qquad J_2 : C \succ A \succ B \succ \cdots \qquad J_3 : B \succ C \succ A \succ \cdots$$
$$J_4 : A \succ C \succ B \succ \cdots \qquad J_5 : B \succ C \succ A \cdots .$$

Each has a UIL score of 10, the scores of all other bands are higher than 10, and considering the three alone changes nothing, so the rule fails. It is surprising that such an outcome has never been confronted in practice. The UIL could correct this flaw by accepting such situations as unavoidable, true ties. Theirs would then be a genuinely new method.[5]

The Frederick Chopin International Piano Competition began in 1927. One of the oldest and most distinguished piano competitions, it is held in Warsaw every five years (except for the disruptions caused by World War II). The methods for ranking candidates have changed over the years and are sketched rather than defined precisely (Frederick Chopin International Piano Competition 2006). The system has gradually evolved into one where competitors are successively eliminated at each of three stages, and (usually) six of the finalists are ranked. Before the 2000 competition, judges awarded points to competitors, and the

5. The Texas State Director of Music, Richard Floyd, informed us that the UIL was aware of this "rare circumstance." They would place C last because it was first least often. Between two bands the procedure is the same as simple majority voting, so there must be a clear outcome with five judges, giving the result $A \succ_S B \succ_S C$. But their rule is arbitrary: they might have said eliminate the bands that are last most often, A and B, yielding the outcome $C \succ_S A \succ_S B$.

sums (or averages) determined who survived, who did not, and how the finalists were ranked. The scale of points assigned by judges has always gone from a low of 0 or 1 up to some upper but varying limit: to 12 in 1927, to 15 in 1932, to 20 in 1937, to 25 between 1955 and 1985, then in 1990 and 1995 to 10 at the first stage and to 25 at the later stages. New regulations have been in use since 2000. They state that "in principle" the first stage of the competition should reduce the number of competitors to 80, the second to 30, the third to 12, and the final stage should rank 6 of them. In 2005, beginning with 257 participants, 80 survived the first stage, 32 the second, and 12 the third. Of the six finalists who were ranked, one was awarded first place, two were tied for third, two were tied for fourth, and one was sixth (in 1990 and 1995 no first place was awarded). Two systems are prescribed. The first, used in each of the three stages, asks that each judge say yes, the competitor should be admitted to the next stage, or no, he should not. The second, a "supporting" system, also used at each stage, asks that each judge accompany his yes-no decision with a grade ranging from 1 to 100. At the end of each stage the average scores of the competitors (without their names) are displayed in descending order together with the yes-no counts to enable the judges to admit the "correct" number of competitors to the next stage (by consensus or by voting in some way that is not specified). The six finalists, evaluated on a 1–12 scale, are ranked according to their average scores.

A quick overview of musical competitions reveals that the vast majority first eliminate in one or more stages, then rank a very limited number of finalists, such as six. It seems that usually average scores determine the ranking, though often the regulations of a competition only say "internal regulations" determine how a jury reaches decisions (including cutoffs, the scale of points, the calculation to merge judges' points into the jury's scores). Sometimes the top and bottom scores are dropped from the calculations to discard the influence of extremes on the average score and to try to eliminate the effects of cheating or favoritism when, for example, a member of a jury evaluates the performance of her student. In at least one piano competition the final result was a shock to all: the jury contained two sets of judges representing distinctly differing sensibilities, causing the piano performances most highly admired by one set of judges to earn mediocre averages from the other.[6] If neither of the two sets constitutes a majority of the jury, the majority-grade and majority-ranking avoid this problem.

A very different procedure has been used to judge flute competitions at the Conservatoire National Supérieur de Musique in Paris. Annual auditions award

6. We are obliged to Thérèse Dussault, an experienced jury member of international piano competitions, for some of this information.

first, second, and third prizes. After all the contestants have been heard, the judges are asked to vote on the question, Should a first prize be awarded at all? If there is not a majority of yeas, no first prize is awarded. Otherwise, a new question is put to vote: Should two first prizes be awarded? If the vote is against, only one is awarded. Otherwise, again a new question: Should three be awarded? The votes continue until the number of first prizes is determined, call it n. Each judge is now asked to write the names of at most n contestants on a ballot; all those who receive a majority win a first prize, and they are listed in order of the number of votes received. The identical procedure is used for second and third prizes.[7] This may be thought of as a kind of approval voting to elect either one or more candidates.

The Bösendorfer and Van Cliburn International Piano Competitions both use what they refer to as a "sophisticated computer program that calculates results based on numerical scores." Little is revealed about it,[8] beyond suggesting that "to balance the scores of a consistently high-scoring juror with a consistently low-scoring juror, the scores of all jurors are processed by computer software to the same statistical distribution" (Herberger College 2006). It thus recognizes a difficulty, pointed out by the docimologists, that is inherent when numbers are used without giving an explicit and absolute sense of what they mean. In any case, a black box whose insides remain hidden should never be accepted for electing or ranking: to be fair—and to be considered fair by all, competitors, judges and the public—a procedure must be known to all and understood by all.

7.4 Skaters and Gymnasts

Judging figure skating has a particularly rich recent history. The big controversy of the 2002 winter Olympic games held in Salt Lake City, Utah, concerned the first two finishers in the pairs figure skating competition and focused on the method of scoring that gave the gold medal 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. Though many admitted that the Russians' final performance was more challenging, it had flaws, whereas the Canadians' did not. Described by many as skating's worst scandal, it provoked a heated debate and deep divisions

7. We are indebted to Michel Debost, an experienced jury member of international flute competitions, for this information.

8. J. A. MacBain, who conceived the system, answered the request for a description with the message, "I have taken the position that my methods are proprietary."

within the skating world. The outcry was so strident that the International Skating Union (ISU), recognized by the International Olympic Committee (IOC) as the international governing body for the sport, ended up changing the verdict and giving both pairs a gold medal. It then went on to institute a complete and fundamental change in the system for judging all figure skating competitions, abandoning the new One-by-One (OBO) method it had just adopted after the last Olympic games. In a furor a dissident camp formed the World Skating Federation with the avowed intent of keeping the old ordinal system and becoming the new IOC-recognized governing body for the sport. It brought suit against the ISU, ultimately lost, then disbanded (*World Skating Federation* 2005). Judging athletic competitions is not, it seems, a benign activity.

The old ordinal system had survived for years despite its foibles. Vote trading was not unknown, one judge exaggerating the score of a skater in return for another judge's doing the same for another skater. Accumulated evidence shows that judges had strong national biases. Sonja Henie of Norway won the 1927 World Championships with the votes of three Norweigian judges, the German and Austrian judges having voted for a German skater. A statistical analysis concludes, "[Judges] appear to engage in bloc judging or vote trading. A skater whose country is not represented on the judging panel is at a serious disadvantage. The data suggests that countries are divided into two blocs, with the United States, Canada, Germany and Italy on one side and Russia, the Ukraine, France and Poland on the other" (Zitzewitz 2006). During the 2002 scandal, a French judge first confessed having favored the Russian over the Canadian pair, saying she had yielded to pressure from her hierarchy, then denied it. Another factor abetted national biases: judges at international competitions were appointed by *national* skating unions, contrary to the practice in ski jumping, for example, where judges are appointed by the Fédération internationale du ski (FIS). By the ISU's new rules, the judges are no longer appointed by the national committees.

The ordinal system also bewildered the public. The rankings among skaters were naturally updated after each individual performance. Often the order between two skaters would be reversed solely as the result of a third skater's performance. The new OBO system (ISU 1998), introduced in 1998, cured neither problem. Moreover, each exercise performed by a skater would be followed by a posting of the judges' technical and presentation marks—between 0 and 6 inclusive—and their averages, along with the new relative standings of all the skaters. Although the language of a score between 0 and 6 seems to have been understood and accepted by all, including the public, the role of the scores was not well understood. That one judge gave a skater a technical mark of 5.3 and a presentation mark of 4.8 (a total of 10.1) on a performance, and another judge gave him a 5.6 and a 5.9 (a total of 11.5) had *no* direct effect

Table 7.1a
1997 European Championships, Ordinals of Men's Free Skating among Top Five before Performance of Vlascenko

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	Mark / Maj	Place
Candeloro	3	2	5	2	3	2	5	5	5	3 / 5	3d
Kulik	2	4	2	3	5	4	3	4	4	4 / 8	4th
Urmanov	1	1	1	1	1	1	1	1	1	1 / 9	1st
Yagudin	4	3	3	5	4	5	4	3	2	4 / 7	5th
Zagorodniuk	5	5	4	4	2	3	2	2	3	3 / 5	2d

Note: "Mark" is the middlemost (or median) of the ordinals; "Maj" is the number of judges in favor of at least the mark.

whatsoever on the ultimate order among the skaters. The sum of the technical and presentation marks given by a judge in a particular exercise to each of the skaters served *only* to determine how that *one* judge ranked all the skaters in that exercise. Thus, in particular, the average total mark earned by a skater in an exercise—made known to the spectators almost instantaneously—meant nothing whatsoever. Since the marks given by a judge served only to determine that judge's ranking of the skaters, why not simply ask judges to rank them straightaway? Because *that* is much harder to do. When there are, say, some fifteen competitors, placing them in order is a very complicated task. It is far more practical to use an absolute measure than to try to compare relative merits. But why, then, did they not use the scores themselves? Perhaps, like Laplace before them, they believed the language was not common at the outset and thus decided to rely only on the orders determined by the evaluations of individual judges.

The 1997 men's competition in the European Championships shows how the old ordinal system worked (here restricted to the top six finishers) (Loosemore 1997). The competitors finished the short program in the following order: first I. Kulik, second V. Zagorodniuk; third A. Vlascenko; fourth P. Candeloro; fifth A. Yagudin, sixth A. Urmanov. With only Vlascenko yet to perform in free skating, the marks of the nine judges resulted in the ordinals—meaning the order of finish according to each of the judges—shown in table 7.1a, where, for example, judge J_3 ranked Candeloro fifth.

A majority principle deduces the jury's decision from the ordinal rankings of the judges. We give a different, simpler, but equivalent description of the system than the usual one. A competitor's *mark* is the middlemost (or median) of the ranks ascribed to him by the (odd number of) judges. This is—except for how to resolve ties—the Borda-majority method described in chapter 5, or Galton's idea applied to ranks rather than scores or grades. A majority of the judges

believe that a competitor's rank should be at least his mark, and a majority of them believe that his rank should be at most his mark. Thus, in fact, the method effectively resists manipulation, unless a majority of the judges collude, which is precisely what was happening. Each of the two blocs presumably leaned on the one other judge to manipulate the final outcome. This is wholesale cheating.

By this method Urmanov's mark is obviously 1. Candeloro's mark is 3: a majority of at least five place him third or better, and a majority of at least five place him third or worse. If there are ties among some marks, a skater with the greater majority in favor of at least his mark takes precedence: thus Kulik takes precedence over Yagudin. If there are ties among some marks that have the same majority in favor (Maj), a skater for whom the sum of the ordinals at least equal to his mark (called the total ordinals of majority, or TOM) is smaller takes precedence: the TOMs of Candeloro and Zagorodniuk are both 12. If, as in this case, ties remain, a skater whose sum of all ordinals (called TO) is smaller takes precedence: Candeloro's TO is 32, Zagorodniuk's is 30, so Zagorodniuk is ahead of Candeloro. If the TOs were also equal, then the two competitors would be considered tied. These complicated rules, advanced with no justification, show the great importance of ending with a complete, *strict* order among the competitors. It happens that the Borda-majority method gives the same results in this case. The majority-values given to the needed precision are (low numbers are better): Urmanov 1., Zagorodniuk 3.4, Candeloro 3.5, Kulik 4.43434, and Yagudin 4.43435.

The final standings are determined by adding each skater's place number in the short program to twice his place number in the free skating program to obtain his index: a skater with a lower index takes precedence, and any ties are resolved by the standings in the free skating program. Thus before Vlascenko's free skating appearance the current standings were, in order (indices in parentheses): Zagorodniuk $(2+2\times 2=6)$, Urmanov $(6+1\times 2=8)$, Kulik $(1+4\times 2=9)$, Candeloro $(4+3\times 2=10)$, Yagudin $(5+5\times 2=15)$.

Then Vlascenko performed. The final ordinals in free skating are given in table 7.1b, along with the final free skating outcome. *Alone*, Vlascenko's performance caused Yagudin to move ahead of Kulik, and Candeloro to move ahead of Zagorodniuk. The final standings were then, in order (ties settled by the order of finish in free skating): first Urmanov $(6+1\times 2=8)$, second Candeloro $(4+2\times 2=8)$, third Zagorodniuk $(2+2\times 3=8)$, fourth Kulik $(1+5\times 2=11)$, fifth Yagudin $(5+4\times 2=13)$, sixth Vlascenko $(3+6\times 2=15)$.

Alone also, Vlascenko's performance in the free skating program pushed Urmanov into first place and Candeloro into second place, relegating the prior provisional leader, Zagorodniuk, to third place. These reversals, or flip-flops,

Table 7.1b
1997 European Championships, Ordinals of Men's Free Skating among Top Six

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	Mark / Maj	Place
Candeloro	3	2	5	2	3	3	5	6	6	3 / 5	2d
Kulik	2	4	2	3	6	5	3	4	5	4 / 6	5th
Urmanov	1	1	1	1	1	2	1	1	1	1 / 8	1st
Yagudin	4	3	3	6	4	6	4	3	2	4 / 7	4th
Zagorodniuk	5	5	4	4	2	4	2	2	3	4 / 7	3d
Vlascenko	6	6	6	5	5	1	6	5	4	5 / 5	6th

Note: "Mark" is the middlemost of the ordinals; "Maj" is the number of judges in favor of at least the mark.

Table 7.1c
1997 European Championships, Ordinals of Men's Free Skating among Top Six Borda-Majority Method

										Majority-Value	Place
Candeloro	2	2	3	3	**3**	5	5	6	6	3....	2d
Kulik	2	2	3	3	**4**	4	5	5	6	4.435...	5th
Urmanov	1	1	1	1	**1**	1	1	1	2	1....	1st
Yagudin	2	3	3	3	**4**	4	4	6	6	4.4343...	4th
Zagorodniuk	2	2	2	3	**4**	4	4	5	5	4.4342...	3rd
Vlascenko	1	4	5	5	**5**	6	6	6	6	5....	6th

Note: Ordinals are ordered for clarity.

were ubiquitous in the old ordinal system and should long ago have encouraged the ISU to hunt for a new system. There happens to be agreement, once again, with the Borda-majority method (table 7.1c).

The ISU then opted for the OBO system (used only once in Olympic competition, 2002). In explaining it, the ISU wrote, "Nothing has been changed in the work of the individual Judge, so that he/she judges every event in the same manner as with the previous Result System. The difference is how the opinion of the majority of the Judges is taken into account" (1998, 1). Thus, in particular, the "0 to 6" language was maintained. The innovation was to use the Dasgupta-Maskin method to obtain a ranking, that is, Llull's (or Copeland's) method, together with Cusanus's (or Borda's) in order to break ties, so it is simple to describe. As before, judges rank-order competitors in the short and free skating programs. In each, the number of times a competitor is preferred by a majority of judges to other competitors is counted (her number of wins): a skater with a higher number of wins is ranked ahead of one with a lower number, with ties among them resolved by their Cusanus- or Borda-scores (when computed for all contestants). The final standings are calculated as before, by

adding a skater's rank in the short program to twice her rank in the free skating program, and breaking ties with the results of the free skating (ISU 2002).[9] All kinds of reversals and discontinuities can occur, as was noted earlier in the context of voting.

The importance of a method's obeying the independence of irrelevant alternatives (IIA) condition is crystal clear when the method is used in a dynamic environment such as sports or cultural competitions. A method that fails the condition provokes immediate suspicion. How can one competitor's performance be relevant to the jury's decision concerning the order among the others? That is why *in practice* procedures for judging competitors have turned more and more to scoring or grading as a means for ranking.

The ISU ended up adopting a completely different and very complex scoring system patterned on the one used for judging gymnastic competitions (ISU 2004). To be specific, consider the men's competition. As in the past, there are two performances, the short program and the free skating program. An "executed element" of both is a part of a program, such as a "layback spin level 3," a "double axel," a "triple-flip," a "death spiral," or combinations thereof. Each executed element has a "base value" of points that is predetermined by a technical committee. A skater's program is formally announced as a collection of executed elements: at the 2006 European Championships in Lyon, every man's short program had eight such elements, and their free skating programs usually had fourteen, though several had thirteen and one fifteen. A judge gives to each executed element a score of 0, ± 1, ± 2, or ± 3—call them merits (if negative, demerits)—that modifies the base value by that amount, up if he considers the execution good, down if bad. He also grades each of five "program components" on a scale of 0.25 to 10 in increments of 0.25: skating skills, transition/linking footwork, performance/execution, choreography/composition, and interpretation. Each program component is multiplied by a factor of 1 in the short program and a factor of 2 in the free skating program. For women's and pairs skating, however, the factor is 0.8 for the short and 1.6 for the free skating program. The order of the contestants is exactly the same as when the factors are respectively 1 and 2. Why, then, the distinction? To ensure that the men's totals dominate any that may be earned by women? At Lyon, the two top men had final scores of 245.33 and 228.87; the top two women, 193.24 and 177.81, which would become 241.55 and 222.26 if the factors were the same as the men's.

9. The ISU data is incomplete in one respect. The Russian and Canadian scores in free skating of the judge accused of deceit are not given.

In any case, for every competitor there are two matrices or tables of numbers: (1) the element table, one line for each executed element and one column for each judge; and (2) the program component table, one line for each component and one column for each judge. The twelve judges are anonymous; it is not revealed which judge announced what merits, demerits, or grades. The system selects three judges at random and ignores their evaluations; which judges and what scores is unknown to all (before the competition has ended). For each skater, the trimmed average of the nine judges' merits or demerits for each element modifies the base value to obtain the score of that element, and the trimmed average of their grades for each program component, multiplied by its factor, gives the score of that component. The *trimmed average* is the average value after the highest and lowest values have been deleted. The sum of the element score plus the program component score gives the total score in each of the two programs. The two are added together to give the final complete ranking of the contestants.

Since scores are involved, no flip-flops are possible. Manipulation is supposed to be discouraged, first, by randomly eliminating some judges, next, by dropping the highest and lowest grades. On the other hand, the anonymity of the judges gives them added impunity (though their scores are known to the ISU)[10] and would seem to damage the credibility of the grades in front of the general public. Moreover, randomly selecting a panel of nine of twelve judges whose grades count invites a question: Suppose some other panel had perchance been chosen?

In all there are 220 different possible panels that can be chosen. Were they to be used to rank the top three women figure skaters in the short program of the 2006 Olympics, 67 panels would agree with the official outcome, 153 panels would not; 92 would agree on first place, 128 would not. The rankings that result from using all 220 panels are given in table 7.2. The official point totals in the short program were very close: Shizuka Arakawa 66.02, Irina Slutskaya 66.70, and Sasha Cohen 66.73. Had the grades of all twelve judges counted, the outcome in the short program would have been Slutskaya first and Cohen second (by 0.28 points). The official final result, based on the short and the free skating programs, was Arakawa first, Cohen second, and Slutskaya third. It was determined by one of the 48,400 ($= 220 \times 220$) different possible combinations of panels; 16,295 of those combinations would have placed Slutskaya ahead of Cohen, and 132 of them would have declared them tied. There is something

10. One judge was barred by the ISU several days before the 2006 Olympics. The ISU had admonished her four times for errors in judgment.

Table 7.2
Rankings of Top Three Finishers in Women's Figure Skating Short Program, 2006 Olympic Games, Turin, Italy, According to 220 Different Possible Panels

No. of Panels	92	3	33	**67**	25
First	Slutskaya	Slutskaya	Slutskaya	**Cohen**	Cohen
Second	Cohen	Arakawa / Cohen	Arakawa	**Slutskaya**	Arakawa
Third	Arakawa	Arakawa / Cohen	Cohen	**Arakawa**	Slutskaya

Note: Boldface indicates official outcome (analyses due to Emerson 2006).
Three panels had Slutskaya first, Arakawa and Cohen tied for second.

fundamentally unfair about a procedure that completely discards the opinions of three of twelve judges (and does so for reasons that are not at all clear).

This system seems overly complex. It forces judges to assign separate scores for each of the executed elements—some thirteen in the women's free skating program of 4 minutes, some fourteen in the men's free skating program of $4\frac{1}{2}$ minutes—in addition to the five program components. How can the judges have the time to look at and appreciate the entire performance given that they must evaluate individually each of fourteen elements in $4\frac{1}{2}$ minutes? Moreover, this means that a judge's evaluation of a whole performance is merely a sum of evaluations of separate components rather than an integrated view of its technical difficulty and of its artistry. Why should the sum be a decent measure of the whole? As the Asian proverb says, "Knowing in part may make a fine tale, but wisdom comes from seeing the whole."[11] Will this system, once skaters and their coaches have worked out the strategies that maximize the number of points they can win, kill the creativity and overall artistry of programs with a quest for perfection in high-valued executed elements? The system's demand on computer processing and data manipulation turns it into a black box that mysteriously returns outputs (the final scores and the rankings) for inputs (each judge's scores). It fails the most important basic principle: to be fair, a system must be transparent and understood by all. The remarks of skaters, coaches, and expert commentators during and after the Olympics of 2006 cast doubts on its validity.

The very year in which the ISU formally introduced the new system, the source of its inspiration, the system of the International Gymnastics Federation (FIG), provoked the major scandal at the 2004 summer Olympic games in Athens. The U.S. gymnast Paul Hamm was awarded the men's all-around

11. In describing his personal approach to evaluating a flautist's performance, Michel Debost explained that he takes into account three distinctly separate elements: the musical quality of tone, the style of playing, and what he called the "cote d'amour," this last clearly an appreciation for the *entire* performance.

gold medal, but if the correct base value (called the start value by the gymnasts) had been given to the Korean Yang Tae Young for his routine on the parallel bars, then *ceteris paribus* Yang would have won the gold instead of the bronze medal. There is no dispute about the error: a move called a belle, with a start value of 9.9, was recorded instead of a move known as a morisue, with a start value of 10. The FIG suspended three judges for the error, declared Yang to be the true winner, and suggested that Hamm should return his medal "as the ultimate demonstration of fair play." Hamm refused. The Korean Olympic Committee demanded a hearing at the Court of Arbitration for Sport. In the words of the court, "It is beyond argument that judges operate under conditions of great pressure when a routine compresses so many elements into so short a time frame . . . Immediately after a gymnast's routine . . . the marks . . . are published on three-sided electronic score boards near the respective apparatus for a period of approximately 15 seconds . . . The arena was noisy as the gymnasts proceeded to their last apparatus—the climax of the event . . . A conversation between a Colombian and a Korean [a Korean judge of the performances had noticed the error and tried to communicate it to a Colombian judge of the start values] was fraught with potential for linguistic misunderstanding" (2004, 6–9). The complexity of the system was unable to withstand the brouhaha and the press of time. Yet, the court ended up refusing to change the awards, for three reasons: "Courts may interfere only if an official's field of play decision is tainted by fraud or arbitrariness or corruption . . . [Any] protest to be effective . . . had to be made before the end of the competition . . . We have no means of knowing how Yang would have reacted had he concluded the competition in this apparatus as points leader . . . So it needs to be clearly stated that while the error *may* have cost Yang a gold medal, it did not necessarily do so" (38–42).

7.5 Divers

The rules of the Fédération Internationale de Natation that apply to diving competitions are clear and straightforward (FINA 2005). Each dive has a degree of difficulty computed by adding five factors, each of which has its own variations of difficulty: somersaults ($0–2\frac{1}{2}$), flight position (e.g., forward, backward, reverse, inward), twists ($\frac{1}{2}–2\frac{1}{2}$), approach (forward, back, reverse, inward groups, armstand), and unnatural entry. Judges grade on a scale of 0, completely failed; $\frac{1}{2}–2$, unsatisfactory; $2\frac{1}{2}–4\frac{1}{2}$, deficient; 5–6, satisfactory; $6\frac{1}{2}–8$, good; and $8\frac{1}{2}–10$, very good. The written rules do not specify that judges are to limit the grades to multiples of one-half, but that may be inferred from detailed FINA results. Clear-cut, absolute meanings are ascribed to each of the numbers.

In diving, there are either five or seven judges. If five, the highest and lowest scores of a dive are eliminated, leaving three scores; if seven, the two highest and two lowest scores are eliminated, leaving again three scores. The sum of the three remaining scores is multiplied by the degree of difficulty to obtain the score of the dive. The diver's final score is the sum of those of his individual dives. In the case of synchronized diving, when a pair of divers perform together, there are four execution judges (two for each diver) and five synchronization judges. The same procedure is followed except that the top and bottom scores of each type of judge are eliminated, leaving five scores.

FINA's rules are transparent and easy to understand. Scores determine rankings, so there can be no flip-flops. Top and bottom, or top two and bottom two, scores are dropped, so excessive cheating is eliminated. The simplicity, in contrast with the rules for skaters and gymnasts, is striking.

7.6 Countries

The Economist Intelligence Unit (EIU) issues a quality-of-life index that compares countries throughout the world. According to the 2005 index, the first ten countries and several others, each with an indication of its place, are Ireland, Switzerland, Norway, Luxembourg, Sweden, Australia, Iceland, Italy, Denmark, Spain, . . . , United States (13), . . . , France (25), Germany (26), . . . , United Kingdom (29). How is this done? First, subjective "life-satisfaction" surveys are made that ask how satisfied people are with their lives on a four-point scale. Second, a statistical regression analysis is done to explain the responses as a function of nine "measurable" indicators of the quality-of-life: material well being (GDP per person), health (life expectancy at birth), stability and security (using another EIU index), family life (divorce rate translated into an index on a scale of 1 to 5), community life (1 if high church attendance or elevated union membership, 0 otherwise), climate and geography (latitude), job security (unemployment rate), political freedom (average of indices of political and civil liberties, on a scale of 1 to 7), and gender equality (ratio of average male and female earnings). When the values of the parameters have been estimated, a country's index is found by multiplying its indicators by the corresponding parameters and summing them (EIU 2005). The idea is that the weight or relative importance of the various attributes of satisfaction are estimated in one way or another, and then satisfaction in a country is a weighted sum of supposedly measurable quantities. This general approach has also been used to calculate indices for ranking universities and hospitals, though it has absolutely no theoretical justification. It is difficult to pretend that adding weighted

latitudes, life expectancies, and unemployment rates means anything at all. Note, nevertheless, the approach: scores are calculated, and they provide the ranking.

7.7 Wines

Wines have been ranked since the first century. "It is the property of wine, when drunk, to cause a Feeling of warmth in the interior of the viscera, and, when poured upon the exterior of the body, to be cool and refreshing. . . . Who can entertain a doubt that some kinds of wine are more agreeable to the palate than others, or that even out of the very same vat there are occasionally produced wines that are by no means of equal goodness, the one being much superior to the other . . . ? Let each person, therefore, constitute himself his own judge as to which kind it is that occupies the pre-eminence" (Pliny the Elder). In his treatise, Pliny the Elder (23–79 C.E.) goes on to catalog the known wines, placing them into four ranks according to their qualities, using such phrases as, "there is not a wine that is deemed superior . . . the Cæcubum enjoyed the reputation of being the most generous of wines . . . there is now no wine known that ranks higher."

Philippe le Bel, King of France (1285–1314), established an official society of Agents-Gourmets-Tasters of wine[12] (Peynaud and Blouin 1999) in 1312, to be responsible for tasting, regulating, and classifying wines. A later decree relative to the sale and distribution of wine, adopted on December 14, 1813, clarifies a member's role:

> Napoléon, Emperor of the French, King of Italy, Protector of the Confederation of the Rhine, Mediator of the Swiss Confederation, etc. . . . We have decreed and so decree what follows:
>
> 14. There shall be named agents-gourmets-tasters of wine. Their number may not exceed fifty.
>
> 15. Their functions are: (1) Exclusive of all others, to store, and when necessary to serve as intermediaries between sellers and buyers of spirits. (2) To taste, to that effect, the said spirits, and to faithfully indicate the vintage and the quality. (3) To serve also, exclusive of all others, as experts when there are disputes as to the quality of wines, and allegations against carriers and boatmen arriving at ports or warehouses claiming wines have been altered or falsified . . .
>
> 17. They shall be named by our minister of commerce . . .
>
> 19. They cannot buy or sell on their own account or on commission, under penalty of destitution. (Décret 1813)

12. *Courtiers-Gourmets-Piqueurs de Vins.*

The famous and still important Bordeaux classification of 1855 was carried out by them.

How are wines classified and ranked today? Robert M. Parker, Jr., the most famous and most powerful critic alive—no less a person than Jacques Chirac anointed him with the presidential accolade, "Robert Parker is the most followed and influential critic of French wines in the entire world" (Parker 2002)—explains how he does it. He judges alone. Seven attributes should bless the critic, he affirms: independence, courage, experience, individual accountability, an emphasis on pleasure and value, a focus on qualitative issues, and candor. He uses a 50–100 scale, undoubtedly modeled on what he experienced as a student in the United States. The grade 90–100 is equivalent to an A, the very best of wines, and therefore rare; 80–89 is equivalent to a B, very good, especially if 85 or above (many such wines are in Parker's personal cellar).[13] The grade 70–79 is a C, average, lacking complexity, though above 75 may be pleasant and, when cheap, "ideal for uncritical quaffing." Below 70 is a "D or an F, depending on where you went to school," flawed, dull, unbalanced wines. Parker begins by giving a wine 50 points; its general color and appearance add at most 5 points; its aroma and bouquet contribute up to 15 points depending on the intensities and its cleanliness; flavor and finish, depending on balance, depth, and length on the palate, can garner as many as 20 points; and 10 additional points are available to rate the overall impression and the potential for improving with age. The grades, published in his bimonthly newsletter, *Wine Advocate*, are awaited with anguish, and they hugely influence sales. It is said that one point more or one point less can cause prices to skyrocket or sink. A measure of the importance of Parker's opinions is his starring role in an excellent documentary movie on wines and their production (*Mondovino* 2004).

A great many national and international wine competitions are held every year throughout the world that are judged by juries of several members. Generally speaking, the anonymity of each wine being tasted is assured by strict rules. The International Wine and Spirit Competition (IWSC) has been organized in Great Britain since 1969. At the first step all judging is done by region, variety or type, and vintage. Samples are presented in numbered glasses, and judges record their scores and give them to a panel chairman, who may decide to discuss them. The highest possible score is 100, broken down into clarity up to 20 and taste and bouquet up to 40 each. The differing grades of the judges are amalgamated into a panel's grade by consensus. "Where the judges are unable to reach a majority decision, flights will be referred to another panel" (IWSC

13. His book of 1,635 pages, just cited, contains very, very few wines with grades under 85.

2006). A grade of 90 to 100 earns a gold medal, 80 to less than 90 a silver medal, and 75 to less than 80 a bronze medal.[14]

Most Australian competitions judge wines with a 20-point scale: 3 points for appearance, 7 for bouquet, and 10 for palate. Scores are given in multiples of 0.5. An average score of 18.5 or above earns a gold medal, 17 to below 18.5 a silver medal, and 15.5 to below 17 a bronze medal. The Regional Agricultural Societies organize most shows with the objective of improving the quality of the products and the efficiency of their production. These are mainly Australian shows, run by them for their wines, with judges looking for faults that are to be eliminated in future years. Some 60% of the entries receive a medal. The director of the Sydney Wine Competition points out "that there is an in-built negativism to this system."[15] He explains that when two of three judges give gold medal scores of 18.5 to a wine, but the third believes it deserves no medal at all and gives it a 15, then even if the first two "cheat" and give it scores of 20, they—the majority—cannot impose their opinion. This is giving power to cranks in proportion to their crankiness! The majority-grade avoids this problem completely.

Wine competitions can have different aims. "The Sydney International Wine Competition's main objective is to help consumers choose pleasing wines to complement their dining Table" (Mason 2006). Its juries are mostly foreigners, and it accords medals to only 20% of the entries. It has thirteen judges, including a chief judge, and upwards of 2,000 wines to rank. In the first phase, there are six panels of two judges, and each judge uses her preferred system of marking to select a certain prescribed percentage (e.g., 20%) of wines out of sets of no more than forty that belong to a same "varietal category." A chief judge intervenes when necessary, in particular when there are disagreements. In the second phase, the wines are arrayed "in a line of perceived palate weight, from lightest bodied to fullest bodied." The competition director admits, "So far as I am aware, there is no scientific method to accurately predict a wine's mouth-feel, its palate weight; this is a perceptual thing." The chief judge groups the wines on the light-to-heavy-bodied scale for subsequent analysis. In the third phase, the director informs members of the jury, "You will be judging the wines in each of these style categories alongside appropriate food complexing your palate." There are two panels of six judges; judges give scores to wines between 1 and 10 (here a 1 corresponds to the usual 15.5, and a 10 to the usual 20) and

14. The details concerning the components of the scores and consensus decisions were provided by Lesley Gray.

15. Warren Mason, in an email received March 16, 2006. We are indebted to him for the information he provided.

write comments (of some fifty to sixty words). A wine's score is the average of the scores given by the judges. Of the total entry, 10% win Blue-Gold awards, a further $2\frac{1}{2}$% win Highly Commended awards.

Vinitaly, an annual Italian wine competition, classed some 4,500 wines in 2005. They are judged in separate categories defined on the basis of color, age, still or sparkling, sweet or dry, and so on. Each jury panel has five members: two Italian members, one foreign technician, and two members of the international trade press. The system is well defined:

> Every wine entered in the Competition is assessed by a jury. The final score for every sample is calculated from the arithmetical average of the individual numerical assessments after eliminating the highest and the lowest scores. Wines achieving the best score (for no more than 30% of entries of each group of every category . . .), provided that they have achieved a minimum score of 80 hundredths in accordance with the "Union Internationale des Œnologues" [U.I.Œ.] evaluation method, will be awarded ex-æquo with a Special Mention Diploma. The 20 wines in every group in each of the categories . . . which achieve the highest scores, provided that they are above 80 hundredths, are then subjected to further evaluation by 3 different juries. In this stage, the score for every sample is calculated from the arithmetical average of the individual numerical assessments after eliminating the highest and lowest scores of each jury. The top 4 wines in each group achieving the best score, of no less than 80 hundredths, will be respectively awarded the Grand Gold Medal, Gold Medal, Silver Medal and Bronze Medal. (Vinitaly 2006)

The U.I.Œ. is an international group of national œnological associations. Their former method for grading wines (OIV 1994), used until 2009, is most easily explained via their standard rating sheet for each wine (table 7.3a). Fourteen different attributes, including a "global opinion," are individually given numerical scores corresponding to seven absolute, self-defined levels: *bad, mediocre, inadequate, passable, good, very good*, *excellent.* Each attribute carries a predetermined weight. A judge circles or brackets his opinion of the wine's respective attributes, and the total score represents his global opinion of the wine. As a rule, the jury's score is taken to be the average of the judges' scores (sometimes with the lowest and highest eliminated, as in the Vinitaly competition). In July 2009 this standard score sheet was slightly changed (OIV 2009), no doubt because the two lowest grades were never or almost never used. Regrettably, the word descriptions of scores were dropped, except for *Excellent* and *Inadequate* (table 7.3b). However, an additional row was adjoined below the score sheet, "Eliminated due to a major defect", and a wine with two eliminations cannot be awarded a medal. Medals are awarded as indicated in table 7.4.

Table 7.3a
Components of U.I. Œ.'s "Sensorial Analysis Tasting Sheet for Wine Judging Competitions" for Still Wines, 2006

	Excellent	Very Good	Good	Passable	Inadequate	Mediocre	Bad
Aspect							
Limpidity	6	[5]	4	3	2	1	0
Nuance	[6]	5	4	3	2	1	0
Intensity	6	5	[4]	3	2	1	0
Aroma							
Frankness	[6]	5	4	3	2	1	0
Intensity	8	[7]	6	5	4	2	0
Finesse	8	7	[6]	5	4	2	0
Harmony	[8]	7	6	5	4	2	0
Taste, flavor							
Frankness	6	[5]	4	3	2	1	0
Intensity	8	[7]	6	5	4	2	0
Body	8	7	[6]	5	4	2	0
Harmony	[8]	7	6	5	4	2	0
Persistence	8	[7]	6	5	4	2	0
After-taste	8	[7]	6	5	4	2	0
Global opinion	[8]	7	6	5	4	2	0

Note: In this example, the bracketed grades sum to 90.

Table 7.3b
Components of U.I. Œ.'s Score Sheet for Still Wines, July 2009

	Excellent +	⟶	⟶	⟶	Inadequate −
Visual					
Limpidity	5	[4]	3	2	1
Aspect other than limpidity	[10]	8	6	4	2
Nose					
Genuineness	[6]	5	4	3	2
Positive intensity	8	[7]	6	4	2
Quality	[16]	14	12	10	8
Taste					
Genuineness	6	[5]	4	3	2
Positive intensity	8	[7]	6	4	2
Harmonious persistence	8	7	[6]	5	4
Quality	[22]	19	16	13	10
Harmony–Overall judgement	[11]	10	9	8	7

Note: In this example, the bracketed grades sum to 94.

Table 7.4
Awards Conferred to Wines

Medal	OIV System	U.I.Œ. System
Grand Gold	score = 0–3	$90 \leq$ score ≤ 100
Gold	score = 4–8	$85 \leq$ score < 90
Silver	score = 9–14	$80 \leq$ score < 85
Bronze	score = 15–21	$75 \leq$ score < 80

The International Organization of Vine and Wine describes itself as "an intergovernmental organisation of a scientific and technical nature of recognized competence for its works concerning vines, wine, wine-based beverages, table grapes, raisins and other vine-based products" (OIV 1994). At its 74th General Assembly in June 1994, it adopted a standard for international wine competitions, arguing that this would guarantee procedural fairness and allow for the comparison of results. The rules stipulate that a jury should be composed of seven jurors (never fewer than five), most of whom should be œnologists (persons who because of their scientific and technical knowledge, and diplomas, are experts on the production and distribution of wine). OIV insists on absolute anonymity in the presentation of the wines to the jurors. The order of presentation is sacrosanct: beginning with whites and going on to rosés and reds, within each first sparkling wines then still wines, and ending up with sweet wines and mistelles.[16] Furthermore, within each category the wines are presented to the jurors in increasing order of "persistence of aromatic intensity." Judges should work in isolation, in well-ventilated, well-lit rooms, at temperatures between 18° and 22° centigrade, and be supplied with a carafe of water, small pieces of bread to clear the palate, and a receptacle in which to discard wine. Wines are to be tasted individually, not comparatively. Each type of wine must be served at its ideal temperature.

"The qualitative evaluations indicated on the jurors' wine tasting rating sheets [table 7.5] are translated into numbers by the secretariat in accordance with the following calculation chart" (OIV 1994). Each of six attributes is rated at one of five levels by a check mark at the appropriate place: the calculations are done by others, with 0 the best possible mark. "Each [wine] receives a rating which is the median rating based on the ratings resulting from the calculation of the evaluation of each of the jurors." A footnote reads, "When there is an odd number of jurors, the median is immediately evident. If the number happens to be even, the median is based on the average of the closest two ratings in the middle of

16. A mistelle is a must whose fermentation has been arrested by the addition of alcohol.

Table 7.5
OIV's "Calculations to Be Made by the Secretariat: Still Wines," 1994

	Excellent 0	Very Good 1	Good 2	Inadequate 4	Eliminated ∞	Weight	Result
Eye							
Aspect	√					× 1	0
Nose							
Intensity		√				× 1	1
Quality		√				× 2	2
Mouth							
Intensity	√					× 1	0
Quality		√				× 3	3
Harmony		√				× 3	3
Total							9

Note: In this example, the checked evaluations give a total grade of 9.

the ratings" (OIV 1994). The reason for taking the median or the middlemost of the jurors' scores accords with precedent (rather than with the incisive insight of Sir Francis Galton). In the days before handheld calculators, computing an average of seven numbers was a chore, but singling out the middlemost was (and is) trivial. In fact, more and more, the jury's grade is taken to be the average of the jurors' grades.[17]

"How use doth breed a habit in a man!" To change a method is very difficult. Nevertheless, the OIV has been and is continuing to work on defining a new method. This can only reflect dissatisfaction with the present system. Its antecedents go back some thirty years when wines could display serious defects in one or another aspect; accordingly, the rating sheets obliged judges to give marks on all aspects. The advances in the technology of making wines and in avoiding abusive transportation and storage have changed the problem. Today wines that enter competitions do not have serious flaws. They are all more or less good in every aspect, yet the whole may be flawed. As Parker writes, "Although technology allows wine-makers to produce wines of better and better quality, the continuing obsession with technically perfect wines is unfortunately stripping wines of their identifiable and distinctive character. Whether it is excessive filtration of wines or insufficiently critical emulation of winemaking styles, the downside of modern winemaking is that it is now increasingly difficult to tell an Italian Chardonnay from one made in France or

17. We are indebted to Jacques Blouin, œnologist and an organizer of international wine competitions, for this and other authoritative information given in the discussion of grading wines.

California or Australia" (2002, 18). Grading a wine on the basis of the sum of the scores of its individual characteristics misses the point, for it "has difficulty in detecting exceptional wines by overly favoring wines that are 'taste-wise correct' " (Peynaud and Blouin 1999, 109). Indeed, many say that professional judges work backward: they first decide what grade a wine should receive, then they score the individual characteristics so that the scores give the desired outcome. Even the great Parker seems to have hinted he may proceed in this manner.

7.8 The Paris Wine Tasting of 1976

The famous—or infamous—wine tasting of May 22, 1976, pitted vintage Cabernet Sauvignon Wines of Bordeaux against those of California. Organized by an English wine expert, Steven Spurrier, nine renowned French connoisseurs together with Spurrier and an American, Patricia Gallagher, tasted six American and four French wines. The tasting was blind, the judges graded each wine on a 0–20 scale, and the order of finish was determined by the wines' average grades. As *Time* magazine wrote on June 7, 1976, "The unthinkable happened: California defeated Gaul." But did it really, or was it simply a happenstance of the mechanism used?

The wines were

A : Stag's Leap 1973 (Californian),

B : Château Mouton Rothschild 1970 (French),

C : Château Montrose 1970 (French),

D : Château Haut-Brion 1970 (French),

E : Ridge Monte Bello 1971 (Californian),

F : Château Léoville–Las Cases 1971 (French),

G : Heitz Martha's Vineyard 1970 (Californian),

H : Clos du Val 1972 (Californian),

I : Mayacamas 1971 (Californian),

J : Freemark Abbey 1969 (Californian).

The judges' grades are given in table 7.6 together with the official ranking (determined by the averages) and several other rankings. The four rankings are different; in particular, the majority judgment places the French wines first, third, fourth, and sixth, an honorable finish that avoids "the unthinkable."

This is a much studied and discussed example of wine tasting, and many have criticized the use of the average score. Most studies concentrate on statistical analyses and deny the significance of the grades themselves except for the

Table 7.6
Judges' Grades, Official Rankings, and Borda-, Quandt, and Majority-Rankings, Cabarnet-Sauvignon Wine Tasting, Paris, May 22, 1976

	A	*B**	*C**	*D**	*E*	*F**	*G*	*H*	*I*	*J*
P. Brejoux	14	16	12	17	13	10	12	14	5	7
A. de Villaine	15	14	16	15	9	10	7	5	12	7
M. Dovaz	10	15	11	12	12	10	11	11	8	14
P. Gallagher	14	15	14	12	16	14	17	13	9	14
O. Kahn	15	12	12	12	7	12	2	2	13	5
C. Dubois-Millot	16	16	17	13.5	7	11	8	9	9.5	9
R. Oliver	14	12	14	10	12	12	10	10	14	8
S. Spurrier	14	14	14	8	14	12	13	11	9	13
P. Tari	13	11	14	14	17	12	15	13	12	14
C. Vanneque	16.5	16	11	17	15.5	8	10	16.5	3	6
J.C. Vrinat	14	14	15	15	11	12	9	7	13	7
Average	14.4	14.3	13.6	11.8	11.6	10.9	10.6	10.5	10.0	5.7
Official rank	1st	2d	3d	4th	5th	6th	7th	8th	9th	10th
Borda-rank	1st	3d	1st	4th	5th	7th	6th	10th	8th	9th
Quandt rank	1st	3d	2d	4th	5th	7th	6th	10th	9th	8th
Majority-rank	2d	1st	3d	4th	5th	6th	8th	7th	9th	10th

Note: Asterisks indicate French wines.
Official rankings are listed from left to right, in accord with averages, highest to lowest.

judges' rankings they induce. Quandt (2006) writes, for instance, "[If] one judge assigns to three wines the grades 3,4,5, while another judge assigns the grades 18,19,20, and a third judge assigns 3,12,20, they appear to be in complete harmony concerning the ranking of wines, but have serious differences of opinion with respect to the absolute quality. I am somewhat sceptical about the value of the information contained in such differences. But we always have the option of translating grades into ranks and then analyzing the ranks." Quandt advocates adding the ranks (low numbers good, high numbers bad) and translating grades into ranks by giving each of k wines having the same grade the average of the next k places on the list. Call it *Quandt's method*. It is an ad hoc kind of Borda idea. Translating grades into ranks discards important information, as shown by Quandt's example. This ignores the strategic aspects of giving grades, which clearly has importance. A very confident judge could well exaggerate her grades up or down to try to impose her will on the decision of the jury.

The number of (expected) wins of a wine X against a wine Y, given in table 7.7, is the number of judges that give a higher grade to X than to Y plus 0.5 for each tie in the grades (thus, for example, A has a higher grade than B five times and three are ties, so A's number of wins against B is 6.5 and *ipso facto* B's number of wins against A is 4.5).

Table 7.7
Number of Wins between Every Pair of Wines, and Borda-Scores, Cabernet-Sauvignon Wine Tasting, Paris, May 22, 1976

	A	*B**	*C**	*D**	*E*	*F**	*G*	*H*	*I*	*J*	Borda-Score
A	–	6.5	4.5	5.5	7.5	10	8	8.5	10.5	8.5	69.5
*B**	4.5	–	5	5.5	8	9	9	9	8	10	68
*C**	6.5	6	–	6.5	5.5	10	8	8.5	9.5	9	69.5
*D**	5.5	5.5	4.5	–	6.5	7.5	7.5	8.5	8	7.5	61
E	3.5	3	5.5	4.5	–	6.5	9	8	6	9	55
*F**	1	2	1	3.5	4.5	–	5	7	6.5	7.5	38
G	3	2	3	3.5	2	6	–	6.5	6	7	39
H	2.5	2	2.5	2.5	3	4	4.5	–	6	4	31
I	0.5	3	1.5	3	5	4.5	5	5	–	5	32.5
J	2.5	1	2	3.5	2	3.5	4	7	6	–	31.5

Note: Asterisks indicate French wines.

In Table 7.7 the number of wins 5.5 indicates a tie (represented by $\approx_S$ in the following formula) between two wines. A number of wins above 5.5 means that a majority of the judges gave a higher grade ($\succ_S$) to the wine listed in the row than to the wine listed in the column of the table. No one seems to have noticed that this is a real example of the occurrence of the Condorcet paradox:

$$E \approx_S C \succ_S D \approx_S A \succ_S B \succ_S E.$$

These five wines—*A, B, C, D,* and *E*—were preferred by a majority of judges to the remaining five wines, and all four methods agree with this (though they disagree on the rankings among the first three). The last five wines are ranked transitively according to the simple majority-rule

$$G \succ_S F \succ_S J \succ_S H \succ_S I,$$

but again there is no agreement among the methods as to the order among them.

The Paris wine tasting of 1976, sometimes called "the judgment of Paris," shows how important it is to have a reliable method of amalgamating opinions founded on sound theoretical grounds.

7.9 Conclusion

The conclusion is inescapable: scores, measures, or grades have been invented to classify and to rank in an incredibly wide variety of circumstances. The facts show that it can be done, that a language is established that permits meaningful measurement. The historical record suggests that, more and more, practical

people needing practical solutions devise mechanisms that transform judges' scores (instead of their ranked preferences) into the jury's scores to determine the final rankings. The language is often different for the same activity. Witness the different scales used for students' grades in different countries. Or look at wines: in some systems 0 is best with higher numbers worse, in others 100 is best with lower numbers worse, in still others—notably in France and Australia—the best is 20, the worst 0. Cultural heritages suggest different languages. But this does not matter, for as Émile Peynaud states in his great French classic on tasting wines, "I renew my advice: to give a global grade to a wine, one should not immediately think in terms of a numerical score, rather one should first decide on its quality; the grade, in the chosen scale, follows automatically" (Peynaud and Blouin 2006, 104).

Practice shows that the scores assigned by judges to an individual competitor (or wine or attribute) are combined into the competitor's total jury score in many different ways—most often by calculating their average or trimmed average, or by summing them, or by finding their middlemost or median score, or by using them to deduce each judge's preference-order among all the competitors and then combining them in some manner—just as candidates to public office are elected by many different systems.

How this *ought* to be done, whatever may be the scale of measurement—numbers, letters, or descriptive qualities—is the subject of this book.

8 Common Language

Language, that wonderful crystallisation of the very flow and spray of thought.
—James Martineau

For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language.
—Ludwig Wittgenstein

Everywhere, in all pursuits, scientific and societal, scales are invented to measure, to understand, to classify, to evaluate, to rank, or to make decisions. This applies to every activity, attribute, candidate, or alternative, be it an immutable concept of the universe—temperature and its degrees—or an ephemeral fancy—the value of a painting and its price. These scales or measures constitute common languages of words that have absolute meanings, clearly understood by those who use them. Many domains of the physical world have natural units of measurement, imposed as it were, by the physics of the situation: time, mass, distance, speed, pressure, energy. Others of the physical world do not, or do not in the present state of knowledge. Yet they show that words and phrases—indeed, sometimes even colors and faces—can and do define perfectly understandable absolute measures.

8.1 Examples of Common Languages

Three examples are sufficient to show how such languages may be defined. The Mohs scale of mineral hardness was proposed by Friedrich Mohs in 1812. Ten specific substances are given, from hardest, rated 10, to softest, rated 1 (American Federation of Mineralogical Societies 2008):

10 diamond (C),

9 corundum (e.g., sapphire and ruby) (Al_2O_3),

8 topaz ($Al_2SiO_4(OH\text{-},F\text{-})_2$),

7 quartz (SiO_2),

6 orthoclase ($KAlSi_3O_8$),

5 apatite ($Ca_5(PO_4)_3(OH\text{-},Cl\text{-},F\text{-})$),

4 fluorite (CaF_2),

3 calcite ($CaCO_3$),

2 gypsum ($CaSO_4 \cdot 2H_2O$),

1 talc ($Mg_3Si_4O_{10}(OH)_2$).

A substance scratched by a diamond but not by a ruby, for example, is rated between 9 and 10. A fingernail is 2.5, gold and silver between 2.5 and 3, a windowpane between 5 and 6. The idea of a scratch test has been traced to Pliny the Elder (77 C.E.), who proposed it to detect false gems.

The Richter scale M_L, whose inventor described it as desirable "for rating [earthquakes] in terms of their original energy" (Richter 1935), though widely used in the press and elsewhere, is considered less reliable than the more recent moment magnitude, MW. However, neither is intended as a measure of the destructive power of a seismic event—its effects on people, houses, buildings and other structures—which is what the modified Mercalli intensity scale aims to do. It describes twelve levels, going from I to XII. Due to Wood and Neumann (1931), it is given here in the abridged version of the United States Geological Survey (1989):

I Not felt except by a very few under especially favorable conditions.

II Felt only by a few persons at rest, especially on upper floors of buildings.

III Felt quite noticeably by persons indoors, especially on upper floors of buildings. Many people do not recognize it as an earthquake. Standing motor cars may rock slightly. Vibrations similar to the passing of a truck. Duration estimated.

IV Felt indoors by many, outdoors by few during the day. At night, some awakened. Dishes, windows, doors disturbed; walls make cracking sound. Sensation like heavy truck striking building. Standing motor cars rocked noticeably.

V Felt by nearly everyone; many awakened. Some dishes, windows broken. Unstable objects overturned. Pendulum clocks may stop.

VI Felt by all, many frightened. Some heavy furniture moved; a few instances of fallen plaster. Damage slight.

VII Damage negligible in buildings of good design and construction; slight to moderate in well-built ordinary structures; considerable damage in poorly built or badly designed structures; some chimneys broken.

VIII Damage slight in specially designed structures; considerable damage in ordinary substantial buildings with partial collapse. Damage great in poorly built structures. Fall of chimneys, factory stacks, columns, monuments, walls. Heavy furniture overturned.

IX Damage considerable in specially designed structures; well-designed frame structures thrown out of plumb. Damage great in substantial buildings, with partial collapse. Buildings shifted off foundations.

X Some well-built wooden structures destroyed; most masonry and frame structures destroyed with foundations. Rails bent.

XI Few, if any (masonry) structures remain standing. Bridges destroyed. Rails bent greatly.

XII Damage total. Lines of sight and level are distorted. Objects thrown into the air.

The last example is the measurement of pain. Those who have suffered severe pain may recall having been asked where they situated their pain on a 0–10 scale, 0 indicating no pain, 10 the most intense imaginable. With the question formulated in this way, the answer means little: it remains a vague, completely subjective suggestion having little to do with an absolute measure.

The following implicit questions together with descriptive words and colors attached to a 10–0 visual scale give meanings to the levels. The scale goes from *unbearable distress* to *no distress*: an intense red is rated 10 and described as *agonizing*; the color gradually transforms into pinks, is rated 8, and called *horrible*; 6 is *dreadful*; the color becomes yellow and is rated 5; then it turns into a very light green, is rated 4, and categorized *uncomfortable*; the green gradually deepens, is rated 2, and is said to be *annoying*; finally, 0, dark green, is *none* (Adams 2008). Here at least some clear distinctions are made as to the meanings of the numbers.

The Mankoski pain scale (Wilderness Emergency Medical Services Institute 2008) presents a much more detailed set of definitions:

0 Pain Free

1 Very minor annoyance—occasional minor twinges. No medication needed.

2 Minor Annoyance—occasional strong twinges. No medication needed.

3 Annoying enough to be distracting. Mild painkillers take care of it. (Aspirin, Ibuprofen.)

4 Can be ignored if you are really involved in your work, but still distracting. Mild painkillers remove pain for 3–4 hours.

5 Can't be ignored for more than 30 minutes. Mild painkillers ameliorate pain for 3–4 hours.

6 Can't be ignored for any length of time, but you can still go to work and participate in social activities. Stronger painkillers (Codeine, narcotics) reduce pain for 3–4 hours.

7 Makes it difficult to concentrate, interferes with sleep. You can still function with effort. Stronger painkillers are only partially effective.

8 Physical activity severely limited. You can read and converse with effort. Nausea and dizziness set in as factors of pain.

9 Unable to speak. Crying out or moaning uncontrollably—near delirium.

10 Unconscious. Pain makes you pass out.

These definitions mix criteria expressed by a person suffering from pain with others that may be used by an objective observer.

An altogether different common language of grades, the Faces Pain Scale—Revised (2010), is used as a measure of pain in pediatric medicine (Hicks et al. 2001). Six faces are aligned on a 0–10 scale. Instructions say,

> These faces show how much something can hurt. This face [*point to left-most face*] shows no pain. The faces show more and more pain [*point to each from left to right*] up to this one [*point to right-most face*]—it shows very much pain. Point to the face that shows how much you hurt [*right now*].
>
> *Score the chosen face 0, 2, 4, 6, 8, or 10, counting left to right, so "0" = "no pain" and "10" = "very much pain."* . . . This scale is intended to measure how children feel inside, not how their face looks.

Each of the three measures of pain uses a 0–10 scale, but their common languages are given different, though related, definitions. Every activity has its own individual scale even when it uses a measure in common with others. The high temperature of a patient is in no way as hot as the lowest temperature of a kiln for baking ceramics. The scale of a thermometer used to assess a patient's temperature is other than that of a thermostat that controls the ambient heat of a room: tenths of degrees (Celsius or Fahrenheit) are significant for the first, integral units suffice for the second, but a much broader interval of temperatures is required to control the warmth of a room than to measure the heat of a human body.

8.2 Measurement Theory

How exactly to construct a scale for a given activity is a science in itself and has received considerable attention. In 1946 the experimental psychologist S. S. Stevens proposed that four levels of measurement should be used to classify the different types of scales:

1. *Nominal measurement*, in which numbers, names, or labels simply assign a particular category (e.g., the number of a bus, a person's blood type, the telephone code of a country): the only meaningful comparisons are "the same" and "different."

2. *Ordinal measurement*, in which numbers are ordinals, letters, or other symbols indicating order (e.g., letter grades on students' exams, the Mohs scale of mineral hardness, the modified Mercalli intensity scale, the several scales of pain): the meaningful comparisons "equal to," "greater than," or "less than."

3. *Interval measurement*, in which numbers indicate order, and in addition, equal intervals have the same significance (e.g., the Gregorian, Hebrew, or Moslem calendars, Celsius and Fahrenheit temperatures): the comparisons of the previous level of measurement are meaningful but so, too, are addition, subtraction, and averages.

4. *Ratio measurement*, in which numbers have the significance of interval measures, and in addition, zero has an absolute meaning (e.g., price, length, mass, the Kelvin temperature scale): multiplying and dividing the numbers of the scale makes sense, in addition to what made sense before.

This classification scheme—disputed by some—imparts an idea about how scales may vary in type and significance.

F. Mosteller and J. W. Tukey are briefer in giving their classification: "We distinguish the following kinds of values: (i) amounts and counts, (ii) balances, (iii) counted fractions, (iv) ranks, (v) grades" (1977, 114). "Amounts" are non-negative real numbers; "counts" are non-negative integers; "balances" are real numbers that may be positive or negative; "counted fractions" are percentages or bounded intervals; "ranks" are orders, 1 being the smallest or largest, 2 the next smallest or largest, and so on; and "grades" are ordered labels. Statisticians disagree as to which statistical methods are appropriate to use in analyzing values coming from one or another level of measurement.

Measurement theorists face two core problems. How to assign scale values to empirical observations is the representation problem. The validity of statements about empirical observations based on analyses of the scale values is the meaningfulness problem.

A *scale* $\mathcal{S}$ is a set of functions from a set which is to be measured X, among which one or more relations hold, into the real numbers. Each element $\phi \in \mathcal{S}$ is a *representation*. For example, if distance is to be measured, one representation is in meters, another in yards. $\mathcal{S}$ is

- an *ordinal scale* if for every $\phi \in \mathcal{S}$, the range of ϕ is an interval of the reals (possibly infinite) and

$$\mathcal{S} = \{f \circ \phi : f \text{ a strictly monotonic function from the range of } \phi \text{ onto itself}\};$$

- an *interval scale* if for every $\phi \in \mathcal{S}$, $\mathcal{S} = \{r\phi + s : r > 0 \text{ and } s \text{ real}\}$;
- a *ratio scale* if for every $\phi \in \mathcal{S}$, $\mathcal{S} = \{r\phi : r > 0\}$.

Roughly speaking, if something is claimed for an attribute that is measured by a particular representation of a scale $\mathcal{S}$, then it is meaningful if the same claim is true measured by any other representation of $\mathcal{S}$. If it is true that "the ratio

of Stendhal's weight to Jane Austen's on 3 July 1814 was 1.42" in kilograms, then it is also true in pounds; on the other hand, if it is true that "the ratio of the maximum temperature today to the maximum temperature yesterday is 1.10" in degrees Celsius, then it is (almost) certainly not true in degrees Fahrenheit.[1]

Measurement theorists seek scales whose meanings can be verified experimentally: "When measuring some attribute of a class of objects or events, we associate numbers (or other familiar mathematical entities, such as vectors) with the objects in such a way that the properties of the attribute are *faithfully* represented as numerical properties" (Krantz et al. 1971, 1, our emphasis). Our search is similar: we seek definitions of levels of ordinal scales—much in the spirit of the levels in the scales of pain or of the destructive force of earthquakes—that faithfully represent the merit of candidates, the excellence of performances, and the quality of competitors. We verify experimentally (in chapters 15 and 21), via statistical analyses, that a common language exists for electing candidates and evaluating wines.

8.3 Common Languages of Grading

The focus of the present enquiry is scales and levels of measurement that have no clear-cut physical existence or validity, such as length, weight, time, or sound. It deals primarily with common languages of grading that are intellectual constructs and that have no meaning other than what is ascribed to them by their users. Obvious examples are a student's grade, the quality of a wine, the brio of a pianist's interpretation, the level of a skater's or a diver's performance, the excellence of a candidate for political office. No instrument for listening will faithfully measure a pianist's performance; no set of chemical tests (in the present state of knowledge) will faithfully measure the quality of a wine; no sets of questions will adequately measure the competence of an aspiring politician. For the most part, there is no a priori demonstration of the validity of most of the scales that are used and analyzed in this book; for the common languages of measurement in these applications, the proof of the pudding is in the eating. And proof there is, as has been seen in a variety of real, practical examples presented in this and the preceding chapter and as proven for the 2007 voting experiment carried out in Orsay.

The language of price in terms of units of money is a completely natural and unquestioned example: it is a common language expressed in arbitrary units (euros, dollars, pounds, francs) that is clearly understood. Indeed, when a

1. The quotations and definitions are taken from Narens and Luce (2008).

currency is revalued—as when the old one hundred French francs became the new one French franc in 1960—or when there is a change in denomination—as when 6.55957 new French francs became 1 euro—most people translate prices back into the older language to fully appreciate their meaning and significance when important expenditures are in the offing. Today in France there are people (undoubtedly adults in 1960) who still evaluate important expenses in old francs: they are the benchmarks!

The real, practical common languages are usually given very careful definitions. The meanings of the U.I.Œ.'s language used to evaluate wines until 2009 is evident; it has seven words: *Excellent*, *Very Good*, *Good*, *Passable*, *Inadequate*, *Mediocre*, *Bad* (see table 7.3a). The numbers associated with each word are used to determine total grades, though they may also be used as synonyms and their intermediate values express more nuances. The same is true for the language of the OIV, which contains only five words, *Excellent*, *Very Good*, *Good, Inadequate*, and *Eliminated*, and its number scheme is different and opposite (see table 7.4). If the same jury used both languages in parallel, the results—in particular, the rankings they imply—could well be different, for two reasons. First, the numbers of words are different; second, although they share certain appellations, the languages are not the same and can therefore elicit different appreciations.

The meaning of the 0–10 scale in increments of $\frac{1}{2}$ used to grade divers is clearly explained in short expressions, much as the language used for judging wines (see chapter 7). It is fair to say that in both of these cases a common language has not only been accepted but its words have become better understood over time and in use. In time, the numbers themselves come to have shared, common meanings. The same cannot be said of the new rules used for skaters and gymnasts. The regulations in those instances are so detailed and so complex—deliberately designed, presumably, to counter cheating—that there is (at present) no common language. This is regrettable, for with the old rules common languages had been established.

The scales used to give grades to students are many and vary across nations, as noted earlier. But each constitutes a well-defined common language. One example concerns Belgium, France, Morocco, Portugal, Peru, Venezuela, Senegal, Mali, Iran, and Tunisia, which all use a 0–20 scale, explained in the following terms: 10–11 is "adequate"; 12–13 is "passable"; 14–15 is "good"; 16 is "excellent"; 17 is "outstanding"; 18–19 is "nearing perfection"; 20 is "perfect" (Wikipedia 2007). Until 2006 the Danish scale consisted of "10 grades ranging from 00 to 13, with 00 being the worst. Grade 00, the completely unacceptable performance; 03 the very hesitant, very insufficient, and unsatisfactory performance; 5 the hesitant and not satisfactory performance; 6 the just

acceptable performance; 7 the mediocre performance, slightly below average; 8 the average performance; 9 the good performance, a little above average; 10 the excellent but not particularly independent performance; 11 the independent and excellent performance; 13 the exceptionally independent and excellent performance" (*Wikipedia* 2007). Note that the grades 1, 2, 4, and 12 were absent. In practice extreme grades are often shunned. This appears to be an attempt at forcing the grades to be interval measures.

The differences in these two languages for grading students result not only from the fact that they have a different number of words but also because the words themselves may well elicit different evaluations. This is an important point. Most voters and judges try to evaluate according to the instructions they receive: different questions evoke different answers. Even the most cynical strategic voter is affected by a change in the language of measurement because the behavior of the others will change. Thus languages for measuring are not merely ordinal. Nor, of course, are they cardinal because adding them makes no sense. The crucial property of a language is that it has an absolute meaning that is common to all. This is, of course, an ideal. French is the language shared by those who are residents of France, English of those who live in the United States and Great Britain, and yet *bleu* is not exactly the same in the mind's eye of all those who speak French, and *green* may evoke different shades to English speakers. Language is never completely common to any two persons. There is another significant analogy between the languages that are measures and ordinary languages such as Spanish or Arabic: they are learned and better understood over time—in fact, become well defined through use—and a combination of time and use can change the meaning of words. A typical example of this is the description of the modified Mercalli intensity scale, which has changed over time in order to enhance precision. Other examples are price inflation (deflation is rare) and grade inflation.

Judges who evaluate a number of competitors, or voters who assess the merit of a number of candidates, are indubitably influenced by the comparisons they observe among the participants of any one competition, for that is the context in which they must assign grades. Take the example of a professor evaluating a class of, say, twenty-five students. The grades assigned are meant to be absolutes yet depend in part on the relative performances of the twenty-five. However, an experienced professor who has taught for twenty years has evaluated some thousands of students—after himself having received grades as a student—and he has a well-developed set of benchmarks that together define absolute evaluations that dominate the relative comparisons. A newly appointed professor's benchmarks are much less clearly defined; he is given guidance via the distributions of grades that are usually observed. A significant deviation may

lead the department chairman to ask for modifications (which an experienced professor would almost certainly reject). The same is true in all competitions: experienced judges have well-established benchmarks with respect to which they judge. The same is true of voters: they have seen and learned about able statesmen—presidents and prime ministers of countries, mayors of cities, senators, representatives—as well as inept or corrupt officeholders. They also have clear-cut benchmarks.

8.4 On the Optimal Number of Grades

How many grades should be used? Mineral hardness uses ten, the Mercalli intensity scale defines twelve, pain is measured on eleven, divers are judged on twenty-one. The official Union Internationale des Œnologues applies seven grades to each characteristic; the International Skating Union modifies the base value of each executed element with one of seven grades ($0, \pm1, \pm2, \pm3$) and assigns one of forty grades to each of the five program components; in France students are given one of twenty-one grades; in the United States they are given one of six letter grades. An interesting but unrelated game theoretical analysis has shown that to best motivate students who are primarily concerned about their status it is preferable to assign them a grade from a coarse set of few grades than one from a fine set of many grades (Dubey and Geanakoplos 2006). A greater number of grades permits a finer distinction but demands a higher degree of expertise and discernment. When competitors (e.g., wines, skaters) are judged according to several different criteria or characteristics, a small number of grades for each may well suffice.

George A. Miller, persecuted by the integer 7 (or so he claimed), was driven to write the most quoted article (Miller 1956) of the first hundred years of the journal *Psychological Review* (Kintsch and Cacioppo 1994). He considered two problems: a human being's capacity for absolute judgment, or the capacity of people to transmit information, and the span of immediate memory (the latter problem—for him fundamentally different—does not concern us). His is the clearest explanation:

> In the experiments on absolute judgment, the observer is considered to be a communication channel . . . The experimental problem is to increase the amount of input information and to measure the amount of transmitted information. If the observer's absolute judgments are quite accurate, then nearly all of the input information will be transmitted and will be recoverable from his responses. If he makes errors, then the transmitted information may be considerably less than the input. We expect that, as we increase the amount of input information, the observer will begin to make more and more errors; we can test the limits of accuracy of his absolute judgments. If the human observer is a

reasonable kind of communication system, then when we increase the amount of input information the transmitted information will increase at first and will eventually level off at some asymptotic value. This asymptotic value we take to be the *channel capacity* of the observer: it represents the greatest amount of information that he can give us about the stimulus on the basis of an absolute judgment. The channel capacity is the upper limit on the extent to which the observer can match his responses to the stimuli we give him. (Miller 1956, 82)

Miller then analyzed the findings of others concerning absolute judgments of "unidimensional stimuli," namely, the absolute judgments of tones (frequencies covering a range from 100 to 8,000 cps in logarithmic steps), of loudness (from 15 to 110 db), of saltiness (concentrations from 0.3 to 34.7 gm), and of visual position of a point in a linear interval. Concerning tones: "The result means that we cannot pick more than six different pitches that the listener will never confuse. Or, stated slightly differently, no matter how many alternative tones we ask him to judge, the best we can expect him to do is to assign them to about six different classes without error . . . Of course, there is evidence that a musically sophisticated person with absolute pitch can identify accurately any one of 50 or 60 different pitches . . . The channel capacity for pitch seems to be about six and that is the best you can do" (85). In loudness the result is five discernible alternatives, in taste intensity it is four, and in visual positioning it is between 10 and 15 (the largest capacity that had been measured). Miller concludes, "On the basis of the present evidence it seems safe to say that we possess a finite and rather small capacity for making such unidimensional judgments and that this capacity does not vary a great deal from one simple sensory attribute to another" (87). And for him that small capacity is 7 ± 2.

Tone, loudness, saltiness, and position on a line all have clear-cut physical measures, so this conclusion concerns people's capacity to measure with their senses—hearing, taste, sight—or to appreciate directly, without recourse to some other device, intensities or levels. The conclusions Miller reaches from the diverse findings is that people in general are able to distinguish 7 ± 2 levels of intensity, whereas a specialist such as a wine connoisseur or someone with absolute pitch may well be able to accurately distinguish many more levels.

This suggests that every evaluation, whether of wines, candidates, or skaters, has a natural limit on the number of levels that evaluators can meaningfully assign to it. When the evaluators are expert judges—of diving, skating, or wines—the number of levels may be relatively high (e.g., twenty or forty). When evaluating candidates in public elections, many voters are naturally not able to make the distinctions if the language exceeds seven or so grades, so the language ceases to be common. This conclusion is confirmed by the 2007 Orsay experiment. The optimal number is the highest number of grades that

constitutes a common language, that is, that allows judges or voters to make absolute judgments.

8.5 Interval Measure Grades

In the great majority of applications—in evaluating students, wines, skaters, divers, politicians (when it is proposed to evaluate them with numbers), and so on—the procedure is to use competitors' sums or averages (or the "trimmed" sums or averages) of the grades to rank them. An implicit, but crucial, assumption is made: the grades constitute a proper interval measure. Do they? The procedure is valid only if it is truly an interval measure.

The decathlon is an athletic competition consisting of ten track and field events, including the 100-meter dash, the shot put, the long jump, and the high jump. For each event a competitor receives a number of points depending on his performance. The sum of the points across all events is the competitor's final score. A natural question presents itself: how should the points be related to the performance? According to the current formula for the 100-meter dash, a time of 12 seconds earns 651 points; 11 seconds, 861 points; 10 seconds, 1,096 points; and 9 seconds (never achieved), 1,357 points. Going from 12 seconds to 11 seconds garners 210 additional points; going from 10 seconds to 9 seconds garners only an additional 285 (although no one has even run the distance in 9 seconds or less). The value in points of reducing the time by one second is not and should not be linear but should be related to the difficulty of the improvement in the performance if the points are to constitute a valid interval measure. This difficulty is measured by the frequency with which it is realized: the distribution of the performances across all competitors must determine how the points are assigned. Thus, given a distribution for, say, the 100-meter dash, ideally each time should be mapped into points so that the same percentages of performances belong to any two intervals of points $[x, x + \epsilon]$ and $[y, y + \epsilon]$. This gives to each interval of the same length the same meaning, and so transforms the performances into points that belong to an interval measure. Similarly, any distribution of performances may be mapped into a uniform distribution in an interval scale of points.

A good example is Denmark's seven-grade number language adopted for the academic year 2006–07 (in order to conform with the new European Credit Transfer Accumulation System's ECTS grading scale). It has seven numerical grades: 12, 10, 7, 4, 2, 0, -3. For sums and averages to make any sense at all, this scale must be an interval measure. Is it? The language of grades is described as follows:

12 (*A*)— *outstanding*, no or few inconsiderable flaws, 10% of passing students,
10 (*B*)—*excellent*, few considerable flaws, 25% of passing students,
7 (*C*)—*good*, numerous flaws, 30% of passing students,
4 (*D*)—*fair*, numerous considerable flaws, 25% of passing students,
2 (*E*)—*adequate*, the minimum acceptable, 10% of passing students,
0 (*Fx*)—*inadequate*,
−3 (*F*)—*entirely inadequate*.

As with the 100-meter dash, the numbers must be related to the percentages of passing students if they are to constitute an interval measure. Imagine that all the real numbers from 2 (the minimum acceptable) up to 12 are possible passing grades in an examination. Underlying the idea of an interval measure is that in the long run, over many students, in the closed interval [2, 12], the percentages of students who obtain a grade in all intervals of length $\epsilon > 0$ are the same. Which of the five passing grades should be assigned to a 5.7? The grade whose number {2, 4, 7, 10, 12} is closest to 5.7, namely, 7 or *good*; or more generally, any number from the interval [5.5, 8.5] should be mapped into a *good*. By the same token, any grade from the interval [2, 3] is mapped into an *adequate*, from [3, 5.5] into a *fair*, from [8.5, 11] into an *excellent*, and from [11, 12] into an *outstanding*. The five numbers (2, 4, 7, 10, 12) were chosen so that the intervals occupy, respectively, the percentages of the whole equal to the percentages of passing grades specified in the definition: [2, 3] occupies 10% of the interval from 2 to 12; [3, 5.5] occupies 25%; [5.5, 8.5] occupies 30%; [8.5, 11] occupies 25%; and [11, 12] occupies 10%. Thus equal intervals have the same significance: on average the same percentages of passing students belong to all such intervals, and on average 10% are *outstanding*, 25% are *excellent*, and so on, down to 10% are *adequate*. Thus, the Danish system is an interval measure.

More formally, suppose k number grades, $x_1 < x_2 < \cdots < x_k$, are to be given, and their percentages are to be $(p_1, p_2, \ldots, p_k)$, so $\sum p_j = 100$. The grades constitute an interval measure when for all i, x_i is in the interval $[p_1 + \cdots + p_{i-1}, p_1 + \cdots + p_i]$ and $\sum_{j=1}^{i} p_j$ is the midpoint of the interval $[x_i, x_{i+1}]$. Let $q_i = \sum_1^i (-1)^{j+1} p_j$ for $i = 1, \ldots, k$.

Theorem 8.1 *There exist number grades $x = (x_1, \ldots, x_k)$ that constitute an interval measure for the percentage distribution $(p_1, \ldots, p_k)$ if and only if there exists a $\delta \geq 0$ that satisfies*

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

When such δ exist, x satisfying

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

defines a set of interval measure grades for each possible value of δ.

The theorem is proved by taking $x_1 = \delta$, noting that

$$2\sum_{j=1}^{i} p_j = x_i + x_{i+1}$$

and

$$0 \le x_1 \le p_1 \le x_2 \le p_1 + p_2 \le x_3 \le \cdots \le x_k \le p_1 + p_2 + \cdots + p_k,$$

then doing a bit of algebraic manipulation.

In the Danish case, namely, $p = (10, 25, 30, 25, 10)$, there is a unique $\delta = 0$ because $q = (10, -15, 15, -10, 0)$ and $\max\{-15, -10\} \le \min\{10, 15, 0\}$. Thus, $\delta = 0$ is unique, and $x = (0, 20, 50, 80, 100)$. Rescaling them by dividing by 10 and translating up by 2 yields the equivalent Danish grades.

On the other hand, if the Danes had observed, or asked for, $p = (8, 24, 36, 24, 8)$, then $q = (8, -16, 20, -4, 4)$, $\max\{-16, -4\} \le \delta \le \min\{8, 20, 4\}$, so any $\delta \in [0, 4]$ yields interval measure grades. For $\delta = 0$ they are $x = (0, 16, 48, 88, 96)$, the lowest but not the highest point in the interval [0, 100] is a grade; for $\delta = 2$ they are $x = (2, 14, 50, 86, 98)$, neither the lowest nor the highest point in the interval [0, 100] is a grade; and for $\delta = 4$ they are $x = (4, 12, 52, 84, 100)$, the highest but not the lowest point in the interval [0, 100] is a grade.

Suppose they had observed the percentages $p = (10, 20, 40, 20, 10)$. Then $q = (10, -10, 30, 10, 20)$, $\max\{-10, 10\} \le \min\{10, 30, 20\}$, so $\delta = 10$ is unique. But now $x = (10, 10, 50, 90, 90)$: the percentages of the five grades cannot be achieved, but it is possible to have three grades, with 30% *A/B*'s, 40% *C*'s, and 30% *D/E*'s, so the sums of the percentages for A and B and for D and E do meet the requirements.

Finally, if instead the Danes had observed or stipulated the percentages $p = (10, 19, 42, 19, 10)$, then $q = (10, -9, 33, 14, 24)$, so $\max\{-9, 14\} \not\le \min\{10, 33, 24\}$ and there is no set of interval measure grades.

Corollary *There are percentage distributions $(p_1, \ldots, p_k)$ for which no set of interval measure grades exist.*

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

8.6 The Lesson

Common languages exist. They are used to measure many things. They may seem to be completely arbitrary, each invented only to serve as a common language—of numbers, alphabets, words, or faces—in order to make common assessments or to arrive at group or collective decisions. But when defined, they have absolute meanings, even for such subjective experiences as pain. And when they are used repeatedly they acquire more and more precise absolute meanings.

"But 'glory,' doesn't mean 'a nice knock-down argument,' " Alice said. "When *I* use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean,—neither more nor less." "The question is," said Alice, "whether you *can* make words mean so many different things." "The question is," said Humpty Dumpty, "which is to be master—that's all." (Carroll 1871)

Juries of experts who class wines, committees of professors who grade students, Olympic officials of various nationalities who judge divers' performances, and earthquake victims who gauge damages use their languages of measures in exactly the same manner as Humpty Dumpty used words. This is but an echo of Wittgenstein's somewhat more somber sounding, "The meaning of a word is its use in the language."

9 New Model

Revolution is not the uprising against pre-existing order, but the setting-up of a new order contradictory to the traditional one.
—José Ortega y Gasset

When we mean to build,
We first survey the plot, then draw the model;
And when we see the figure of the house,
Then must we rate the cost of the erection;
Which if we find outweighs ability,
What do we then but draw anew the model
In fewer offices, or at last desist
To build at all?
—William Shakespeare

Over seven hundred years of effort and a host of impossibility theorems show that the "Arrovian model", where many individual rankings are to be resolved into a single collective ranking, cannot be made to work: there is no satisfactory mechanism for doing what is wanted. Experience shows, on the other hand, that it is a relatively simple matter to invent grades, scores, levels, or measures to evaluate the performances of students, figure skaters, divers, and musicians, the qualities of wines and cheeses, and the intensities of seismic events, and so by inference to determine the relative merits of competitors in any situation. With use, the grades take on meaning, so they come to constitute a common language of evaluation. Experience also shows that what to do with judges' scores—how to resolve them into a single score—is far from evident. Practical people have devised many different mechanisms.

The first step is to formulate the problem precisely. That is the aim of this chapter. It presents the basic model, which consists of a common language, a set of judges, and a set of competitors.

9.1 Inputs

A *language* Λ is a set of *grades* (words, levels, or categories) denoted by lowercase letters of the Greek alphabet, $\alpha, \beta, \ldots$. It is strictly ordered; specifically, supposing $\alpha, \beta, \gamma \in \Lambda$, (1) any two levels may be compared, $\alpha \neq \beta$ implies either $\alpha \prec \beta$ (β is the higher grade) or $\alpha \succ \beta$ (α is the higher grade); and (2) transitivity holds, $\alpha \succ \beta$ and $\beta \succ \gamma$ imply $\alpha \succ \gamma$. Otherwise, there is no restriction: a language Λ may be either finite or a subset of points of an interval of the real line. $\alpha \succeq \beta$ means that either α is a higher grade than β or $\alpha = \beta$.

There is a finite set of m *competitors* (alternatives, candidates, performances, competing goods) $\mathcal{C} = \{A, \ldots, I, \ldots, Z\}$. Individual competitors are denoted by uppercase Latin letters.

There is also a finite set of n *judges* $\mathcal{J} = \{1, \ldots, j, \ldots, n\}$. Individual judges are denoted by lowercase Latin letters, typically i, j, k.

A problem is completely specified by its *inputs*, or 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 judges $j \in \mathcal{J}$ to each of the competitors $I \in \mathcal{C}$. Thus, if $\mathcal{C}$ is a collection of wines, $\mathcal{J}$ is a jury of five œnologists, and Λ is a language of six grades or levels—say, *excellent, very good, good, mediocre, poor, bad*. Each judge gives to each wine one of the six grades, and the profile Φ is a matrix of grades with five columns and as many rows as there are wines in the collection to be tasted.

9.2 Social Grading Functions

A *method of grading* is a function F that assigns to any profile Φ—any set of grades in the language Λ assigned by judges to competitors—one single grade in the same language for every competitor:

$$F : \Lambda^{m \times n} \to \Lambda^{m}.$$

Thus $F(\Phi)$ is a vector whose Ith component is the collective or *final grade* assigned to competitor $I \in \mathcal{C}$ by the mechanism F. As an example, suppose three wines A, B, and C were to be evaluated by five judges in the language postulated earlier. A possible profile Φ is the matrix that is the argument of the function F, and a possible set of grades for the three wines A, B, and C (in that order) is on the right:

$$F\begin{pmatrix} very\ good & good & good & mediocre & good \\ excellent & good & good & poor & very\ good \\ mediocre & excellent & poor & good & bad \end{pmatrix} = \begin{pmatrix} good \\ very\ good \\ mediocre \end{pmatrix}.$$

In Arrow's model the inputs are the judges' rankings of the candidates; there is no language or measure. A ranking function—or what he calls a social welfare function—assigns to any preference-profile, one single ranking of society. In terms of wines this would mean that every judge rank-orders all of them, and the ranking function deduces one collective rank-order among them. But the primary aim of the grading model is to *classify* competitors or alternatives, to give them final grades as students are given final grades. The final grades may be used to rank competitors, but only up to a point, because several competitors appreciated differently by the judges may have a same final grade.

Many different grading methods F may be imagined. When the language is numerical, say grades range from 0 to 100, the most often encountered example is an F that assigns to each competitor the average of the grades given her by the judges. Other possibilities would be to assign each competitor the lowest of all her grades or the highest of all of them. But F should obey some minimal requirements to be deemed acceptable. What should they be?

They are directly inspired by the requirements imposed on the traditional model of social choice theory. There should be no inherent advantage or disadvantage given to any one or more competitors: all should be treated equally. So if, for example, the three wines A, B, and C were listed in a different order—say, B followed by A, then by C—then F should yield the same answer: B very good, A good, C mediocre. When the rows (or competitors) of a profile Φ are permuted, F should give the identical answer permuted in the same way. Axiom 9.1 states this formally.

Axiom 9.1 *F is neutral, $F(\rho\Phi) = \rho F(\Phi)$, for any permutation ρ of the competitors (or rows).*

Similarly, all judges are to have the same influence on the grades, meaning that when the columns (or judges) of a profile are permuted, F should give the identical answer. In some situations—to protect minority rights, for example—it may be desirable to give certain judges more weight than others. It suffices to count their inputs several times.

Axiom 9.2 *F is anonymous, $F(\Phi\tau) = F(\Phi)$, for any permutation τ of the judges (or columns).*

A method of grading is *impartial* when it is both neutral and anonymous.

Three other properties naturally impose themselves.

First, if every judge is in agreement on the grade to be given to a competitor, then he must be assigned that final grade.

Axiom 9.3 *F is unanimous: If a competitor is given an identical grade α by every judge, then F assigns him the grade α.*

That is, $F(\Phi)(I) = \alpha$ when $\Phi(I, j) = \alpha$ for every $j \in \mathcal{J}$.

Next, if in comparing two profiles, a competitor I's grades in the second are all the same or lower than in the first, then F cannot assign the competitor a higher grade in the second case than in the first. Moreover, if all the competitor's grades are strictly lower in the second profile, then F must assign him a strictly lower grade in the second case.

Axiom 9.4 *F is monotonic: If two inputs Φ and Φ' are the same except that one or more judges give higher grades to competitor I in Φ than in Φ', then $F(\Phi)(I) \succeq F(\Phi')(I)$. Moreover, if all the judges give strictly higher grades to competitor I in Φ than in Φ', then $F(\Phi)(I) \succ F(\Phi')(I)$.*

In other words, if $\Phi(I, j) \succeq \Phi'(I, j)$ for every $j \in \mathcal{J}$, then $F(\Phi)(I) \succeq F(\Phi')(I)$; and if $\Phi(I, j) \succ \Phi'(I, j)$ for every $j \in \mathcal{J}$, then $F(\Phi)(I) \succ F(\Phi')(I)$. When F satisfies only the first of the two properties, it will be said to be *weakly monotonic*; when only the second, it will be said to be *strictly monotonic* (sometimes referred to as Pareto efficiency or unanimity).

Finally, Arrow's famous independence of irrelevant alternatives is a very natural condition when translated into the context of this model. The collective grade of a competitor should depend on her grades alone: it should certainly not depend on any other competitor's grades.

Axiom 9.5 *F is independent of irrelevant alternatives in grading* (IIAG)*: If the lists of grades assigned by the judges to a competitor $I \in \mathcal{C}$ in two profiles Φ and Φ' are the same, then $F(\Phi)(I) = F(\Phi')(I)$.*

That is to say, if $\Phi(I, j) = \Phi'(I, j)$ for every $j \in \mathcal{J}$, then $F(\Phi)(I) = F(\Phi')(I)$.

These axioms are already sufficient to reduce the choice of a method of grading to a manageable, well-defined class. A function

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

that transforms grades given to one competitor into a final grade will be called an *aggregation function* if it satisfies the following three properties:

- *anonymity*: $f(\ldots, \alpha, \ldots, \beta, \ldots) = f(\ldots, \beta, \ldots, \alpha, \ldots)$;
- *unanimity*: $f(\alpha, \alpha, \ldots, \alpha) = \alpha$;
- *monotonicity*:

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

and

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

When f only satisfies the first of the two monotonicity properties, it will be said to be *weakly monotonic*.

A language Λ is often parameterized as a bounded interval of the non-negative rational or real numbers. In either case an obvious example of an aggregation function assigns the mean value of its arguments. Other examples are those that assign the geometric, the harmonic, or any other of the well-known means; those that assign the minimum or the maximum value of its arguments; or more generally, those that assign the kth largest of the arguments for $1 \leq k \leq n$ (called order statistics by probabilists).

Theorem 9.1 *A method of grading F is impartial, unanimous, monotonic, and independent of irrelevant alternatives in grading if and only if $F(\Phi)(I) = f\big(\Phi(I)\big)$ for every $I \in \mathcal{C}$, for some one aggregation function f.*

Proof If there is an aggregation function f that defines F as in the statement of the theorem, then the axioms are obviously met by F. On the other hand, suppose F satisfies the axioms. IIAG and neutrality imply that F determines the grade of a competitor $I \in \mathcal{C}$ solely on the basis of the grades assigned to I; so call the function that does this f. The other three axioms—anonymity, unanimity, and monotonicity—immediately establish the corresponding properties of f, so it must be an aggregation function. ∎

The theorem is obvious. Yet it already eliminates Arrow-type impossibilities.

In practice, grades are almost always numbers, final grades almost always averages or truncated averages. The language of a judge's grades may be discrete—in Australian wine tastings, for example, the grades of individual judges can range from 0 to 20, and they are assigned in multiples of $\frac{1}{2}$—but the language of final grades is richer: with five judges a wine's final grade can be 17.10. Or to take a more probing example, consider the seven grades used by the Union Internationale des Œnologues (U.I.Œ.) (see chapter 7) and assign them numbers: excellent 6, very good 5, good 4, passable 3, inadequate 2, mediocre 1, bad 0. The average of a five-person jury giving the grades—very good, good, good, good, bad—would be 3.40; that might reasonably be described as a passable+. Though that word is not in the judges' original language, it almost surely *becomes* a word in the language. So why not let it as well as its sisters and its cousins and its aunts enter all at once? The point is that in practice final grades are often more detailed than the grades a judge

is allowed to give, yet they cannot help but take on meanings of their own. Were they then to be used in the judges' language, a further enrichment of the final grades would ensue. Why not simply take all the possible numbers as grades to begin with?

Accordingly, in conformity with most practical applications, the common language is parameterized as a subset of real numbers and whatever aggregation is used, small changes in the parameterization or the input grades should naturally imply small changes in the outputs or the final grades. The analysis in the rest of this book could equally well have chosen any open or half-open interval, including $[0, \infty)$. Even if the original language is finite, the possibility of taking an arbitrary aggregation function implies that all possible parameterizations must be considered. Accordingly, the common language will be taken to be $[0, R]$, as did Laplace. Whereas Americans may like $R = 100$, the French $R = 20$, and the Danes $R = 13$, most mathematicians probably prefer $R = 1$. The choice is unimportant. Grades are almost always bounded. The grades used by the International Organization of Vine and Wine (OIV) are an exception: 0 is the best, ∞ the worst.

Suppose that Λ and Λ' are the number grades corresponding to two languages or parameterizations that are ϵ-*close*: $r \in \Lambda$ implies there exists an $r' \in \Lambda'$ with $|r - r'| < \epsilon$, and symmetrically, $r' \in \Lambda'$ implies there exists an $r \in \Lambda$ with $|r - r'| < \epsilon$. It is then clear that a method of grading F should be defined by an aggregation function f that satisfies $|f(r_1, \ldots, r_n) - f(r'_1, \ldots, r'_n)| < \eta(\epsilon)$ when $\max_{j \in \mathcal{J}} |r_j - r'_j| < \epsilon$ for some positive function $\eta(\epsilon)$ that converges to 0 when ϵ approaches 0. That is, f should be uniformly continuous, so, since $[0, R]$ is compact, f should be continuous.

Axiom 9.6 *F (and its aggregation function f) is continuous.*

Two lists of grades that are very similar should clearly be assigned final grades that differ by very little. Enriching a language by embedding it into a real interval opens the door to vastly more possible methods of grading, but it will turn out that the aggregation functions that emerge as those that *must* be used are directly applicable in the seemingly more restrictive discrete languages as well. *The characterizations in the next chapters sometimes require the language to be sufficiently rich and the functions to be continuous; but the properties of the functions that are characterized hold for arbitrary finite languages.* Many theorems do not require all axioms.

A *social grading function* (SGF) F is a method of grading that satisfies the six axioms of the basic model.

Thus F defines, and is defined by, a unique continuous aggregation function f.

The number of candidates or of judges may be varied to study certain properties and phenomena.

9.3 Social Ranking Functions

With Arrow's model, "one of the consequences of the assumption of rationality is that the choice to be made from any set of alternatives can be determined by the choices made between pairs of alternatives" (Arrow 1951, 20). But, as Arrow's theorem shows, this ideal cannot be realized. One of the consequences of rationality in the new model is that every single alternative has a final grade that is independent of all other alternatives: there is no ambiguity in grading. But what happens when the aim is to rank alternatives?

Given a finite language Λ, judges assign grades to any number of competitors that are the inputs or profile

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

Imagine that a competitor A is assigned the list of grades $\alpha = (\alpha_1, \ldots, \alpha_n)$ and a competitor B the list $\beta = (\beta_1, \ldots, \beta_n)$. A *method of ranking* is a nonsymmetric binary relation $\succeq_S$ that compares any two competitors, A and B, whose grades belong to some profile Φ. By definition, $A \approx_S B$ if $A \succeq_S B$ and $B \succeq_S A$; and $A \succ_S B$ if $A \succeq_S B$, but it is not true that $A \approx_S B$. Thus $\succeq_S$ is a complete binary relation.

Any reasonable method of ranking should possess certain minimal properties.

Axiom 9.7 *The method of ranking $\succeq_S$ is neutral: $A \succeq_S B$ for the profile Φ implies $A \succeq_S B$ for the profile $\sigma\Phi$, for σ any permutation of the competitors (or rows).*

Axiom 9.8 *The method of ranking $\succeq_S$ is anonymous: $A \succeq_S B$ for the profile Φ implies $A \succeq_S B$ for the profile $\Phi\sigma$, for σ any permutation of the judges (or columns).*[1]

Axiom 9.9 *The method of ranking $\succeq_S$ is transitive: $A \succeq_S B$ and $B \succeq_S C$ implies $A \succeq_S C$.*

1. So, for example, A's grades $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ are permuted to $(\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_{\sigma n})$.

Axiom 9.10 *The method of ranking $\succeq_S$ is independent of irrelevant alternatives in ranking* (IIAR): *If $A \succeq_S B$ for the profile Φ, then $A \succeq_S B$ for any profile Φ' obtained from Φ by eliminating or adjoining some other competitor (or row).*

Axiom 9.9 demands that the Condorcet paradox be avoided. Axiom 9.10 is strong independence of irrelevant alternatives, as defined in chapter 3. It demands that Arrow's paradox be avoided. These are the two important paradoxes that have been observed to occur in real competitive situations.

A method of ranking *respects grades* if the rank-order between two candidates A and B depends only on their sets of grades.

Thus, the preference lists induced by the grades must be forgotten: it matters not which judge gave which grade. In other words, if two judges or voters switch the grades they give to a candidate, then nothing changes in the jury's or the electorate's ranking of all candidates.

A method of ranking *respects ties* if when any two competitors A and B have an identical set of grades, they are tied, $A \approx_S B$.

Respecting grades together with impartiality implies respecting ties.

Theorem 9.2 *A method of ranking is neutral, anonymous, transitive, and independent of irrelevant alternatives in ranking if and only if it is transitive, and respects ties and grades.*

Proof To compare the sets of grades of two competitors A and B it suffices to compare them alone (by IIAR).

Suppose A's list of grades is $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and B's list of grades is a permutation σ of them, $(\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_{\sigma n})$.

To begin, consider the profile

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

where the grades of A are in the first row and those in the second row are called those of A′. Suppose $A \succeq_S A'$. Permuting the grades of the two judges 1 and $\sigma 1$ changes nothing by anonymity,

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

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 consider the profile

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

and permute judges 2 and $\sigma 2$ to conclude, as in the first step, together with transitivity, that $(\alpha_1, \alpha_2, \ldots, \alpha_n) \approx_S (\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_n)$, the second list agreeing with B's in the first two places. Continuing, in at most n steps, $(\alpha_1, \alpha_2, \ldots, \alpha_n) \approx_S (\alpha_{\sigma 1}, \alpha_{\sigma 2}, \ldots, \alpha_{\sigma n})$, so ties are respected.

The order between any two lists α and β respects grades because each of them is equivalent to a unique representation in which the list is written from the highest to the lowest grade and it suffices to compare them, so grades are respected. The converse is immediate. ∎

Notice that no axiom asks that the language of grades be understood in the same way by all voters or judges. The only implicit assumption is that the scale of grades be absolute for each individual judge or voter. This implies that if some competitor is added or dropped, a judge's true grade remains the same. On the other hand, as will be seen anon, for the final decisions to be meaningful, the scale of grades must be common to all judges or all voters.

A *social ranking function* (SRF) is a method of ranking that satisfies the four ranking axioms.

This simple theorem is essential; it says that if Arrow's and Condorcet's paradoxes are to be avoided, then the preferences induced by the grades must be forgotten. *Who gave what grade cannot be taken into account*: only the sets of grades themselves may be taken into account. It also says that in the new model Arrow's ideal can be realized: "The choice to be made from any set of alternatives can be determined by the choices made between pairs of alternatives."

9.4 The Role of Judges' Utilities

Nothing has been said yet concerning the behavior of the judges or the voters, their complex and often secret aims, their likes and dislikes. The tradition in the theory of social choice is to assume that judges and voters have preferences, invariably expressed as rank-orders. But the word "preference" misleads. A judge in a court of justice evaluates in conformity with the law, which has nothing to do with his preferences; a judge may dislike a wine presented in a competition yet give it a high grade because of its merits, or he may like one and without qualm give it a low grade because of its demerits; an elector may cast a vote not in accord with his personal opinions of the candidates but rather

in the hope of making the correct decision by electing the best candidate for the job (see, e.g., Goodin and Roberts 1975; but Llull, Cusanus, Condorcet, and all the early thinkers formulated the problem in these terms, as do also some philosophers today, e.g., Estlund 2008).

Thus whereas the traditional model pretends to aggregate the preferences of judges and voters, in fact it does nothing of the kind. It amalgamates individuals' rankings of the candidates—the input—into society's ranking of the candidates—the output. The possible outputs are *rankings*, yet the inputs say nothing about how a judge or a voter compares the rankings.

In the real world the deep preferences or utilities of a judge or a voter are a very complicated function that depends on a host of factors, including the decision or output, the messages of the other judges (a judge may wish to differ from the others, or on the contrary resemble the others), the social decision function that is used (a judge may prefer a decision given by a "democratic" function to one rendered by an "oligarchic" function, or the contrary), and the message she thinks is the right one (a judge may prefer honest behavior, or not). We contend that the deep preferences of judges or voters *cannot* be the inputs of a practical model of voting. *A judge's input is simply a message, no more no less*. But her input, chosen strategically, depends, of course, on her deep preferences or utilities.

Amartya Sen's model (1970) and the subsequent work on "welfarism" (Blackorby, Donaldson, and Weymark 1984; Bossert and Weymark 2004)—often referred to as social welfare functionals—postulates real number utilities on candidates as the input, a rank-ordering of candidates as the output. The motivation is the study of social welfare judgments in the context of Arrow's framework but with more information in order to avoid the impossibilities. Sen makes no claim that this approach is valid in the context of voting. As with any mathematical model, the mathematical symbols may be given very different interpretations. At a formal level, reinterpreting the symbols of the inputs of Sen's model as the grades of a language yields a social ranking function. But this ignores the essential concept of a common language. By contrast, utilities measured in an absolute scale play no significant role in the social welfare functional literature, which focuses instead on weaker information invariance assumptions (although they are often assumed for simplicity, e.g. Blackorby, Bossert, and Donaldson 2005).

Social welfare functionals are not intended to enable a comparison of rankings. For, how are two outputs—two rankings of the candidates—to be compared by a voter or judge on the basis of his utilities for individual candidates? If the answer is by looking at the first-place candidate of the rankings, then all $(m-1)!$ rankings that have the same first-place candidate must be taken

as giving him equal satisfaction. This is too restrictive for a theory (or practice) that designates winners and orders of finish.

The language of grades has nothing to do with utilities (viewed as measures of individual satisfaction). Grades are absolute measures of merit. In the context of voting and judging, utilities are relative measures of satisfaction. The 2002 French presidential elections offer a perfect example of the difference. The voters of the left would have hated to see Jacques Chirac defeat Lionel Jospin: their utilities for a Chirac victory would have been the lowest possible. The same voters were delighted to see Chirac roundly defeat Le Pen in the second round: their utilities for a Chirac victory were the highest possible. On the other hand, these same voters would probably have given Chirac a grade of *Acceptable* or *Poor* (on a scale of *Excellent, Very Good, Good, Acceptable, Poor, To Reject*) were he standing against Jospin, Le Pen, or anyone else.

Formally, a distinction must be made between two different types of scales of measurement. An *absolute* scale measures each entity individually (height, area, merit). A *relative* scale measures each entity with respect to a collective of like entities (velocity, satisfaction). Were voters to be asked their satisfaction as inputs, adjoining or eliminating candidates would alter their answers, provoking the possibility of Arrow's paradox. A common language must be an absolute scale of measurement.

Utility plays another, important role in voting and judging. A decision maker is routinely assumed to behave in such manner as to try to maximize his utility. But what is it? In theory the utility of a judge or voter j may be imagined to be a function $u_j(\Phi^*, \Phi, f, \mathcal{C}, \Lambda)$, where $\Phi^* = (\alpha^*_{ij})$, with α^*_{ij} the grade judge j believes candidate i merits, $\Phi = (\alpha_{ij})$, with α_{ij} the grade judge j actually gives candidate i, f is the aggregation function, $\mathcal{C}$ the set of competitors, and Λ the common language that is used. The utility of judge j could include a term $-|\alpha^*_{ij} - \alpha_{ij}|$ if she wished to grade candidate i honestly; it could include a term $-\sum_{k\neq j} |\alpha^*_{ik} - \alpha_{ik}|$ if she wished that the other judges graded i honestly; it could include a term $-|\Lambda - \Lambda^*_j|$ if she wished that the common language was Λ^*_j rather than Λ; and the reader can no doubt invent many other utility functions that a judge might have, or plausible components of them. One hypothesis is to imagine that a judge's utility is single-peaked in the grade of each candidate i, $u_j = \sum_i -|\alpha^*_{ij} - f(\alpha_{i1}, \ldots, \alpha_{in})|$: the further the final grade $f(\alpha_{i1}, \ldots, \alpha_{in})$ is from what judge j believes it should be, the less her satisfaction. Another is that a judge's utilities depends solely on the winner, which is usually assumed in the analysis of voting games. In fact, of course, judges' utilities, judges' beliefs, their beliefs about the others' beliefs, their likes and dislikes for the decision mechanism or the language are all completely unknown and often

hidden, and they change from one competition to another (indeed, perhaps a voter's utility on a sunny election day differs from that on a rainy election day).

In the terms of the current technical jargon, we are faced with a problem of *mechanism design*: "[Individuals'] actual preferences are not publicly observable. As a result ... individuals must be relied upon to reveal this information ... [The problem is] how this information can be elicited, and the extent to which the information revelation problem constrains the ways in which social decisions can respond to individual preferences" (Mas-Colell, Whinston, and Green 1995, 857). This is often seen as a problem of the theory of games where the information is incomplete. The standard approach postulates that every individual is of a certain *type* and associates to each type a utility function that depends only on the outcome. Typically, the individuals' types are drawn from a set of types by some known prior probability distribution, and the utility functions are all of some common analytical form (whose parameters vary with the different types). The methods are then, of course, dependent on the utilities that are postulated.

In contrast, the methods we develop make no overall assumptions concerning utilities. They are similar, in this regard, to Vickrey's "second price" mechanism, which allocates the good up for auction to the highest bidder at a price equal to the second highest bid (Vickrey 1961): it depends only on the bidders' bids—their "private values"—not their utilities. Our mechanisms depend *only* on what in practice can be known. Knowing the judges' or voters' true utilities is unnecessary to much of the analysis. The mechanisms that emerge as the only ones that separately satisfy each of several desirable properties are strategy-proof for large classes of reasonable utility functions, though not all. When they are not strategy-proof, they are unique in being the least manipulable methods in several well-defined senses.

10 Strategy in Grading

It is not true that men can be divided into absolutely honest persons and absolutely dishonest ones. Our honesty varies with the strain put on it.
—George Bernard Shaw

The members of a jury assign grades. A social grading function defines a mechanism for transforming the individual grades of several or many judges into one final grade of the jury. The issues addressed in this chapter focus on the question, What strategies will judges use in the game of assigning grades? Later chapters consider other strategic games that judges may play, notably, how they may act and react to giving grades when these are also used to rank competitors.

Experience clearly establishes the fact that assigning grades *is* a game, because the players—the judges—may assign their grades strategically. A judge may assign a grade that is well above or well below what he believes is the correct grade so that the jury's final grade approaches, as much as possible, what he believes it should be. Worse, he may assign a grade dishonestly for reasons that are totally extraneous to the performance or alternative being judged. A device used in many competitions (e.g., sports and music contests) to counter such temptations is to eliminate extreme grades, the highest and the lowest or the two highest and the two lowest. In the 2002 figure skating scandal at the Olympic games in Salt Lake City, the grade of a judge seems to have been exaggerated beyond what it should have been for reasons having to do with nationality rather than performance. As a result, the new International Skating Union system of grading randomly discards several scores and then eliminates the highest and lowest. Strategic manipulation in wine competitions is of a different order because the wines are almost invariably tasted blind: careful procedures make sure that a judge has no information whatsoever concerning the identity of the wines he tastes. Nevertheless, the problem remains for, as was pointed out by an organizer of wine competitions (see chapter 7), the rules

may give one judge the power of forcing a final grade to be well below what the majority believes it should be.

The potential for a judge to manipulate grades in the face of a mechanism that determines the jury's final grade is, of course, a mirror image of similar behavior in the traditional model where a judge has the potential to manipulate his rank-ordering knowing what method of voting will be used, or of the behavior of a bidder in an auction knowing the rules that determine the winner and the price. When Borda's method is the rule in the traditional model, a voter may well list the strongest opponent of her favorite last, though she believes that the opponent is the second best among all candidates. As was recognized by Laplace, and indeed by Borda himself, his method does not encourage honest voting. Vickrey showed that a similar phenomenon occurs in auctions: the usual mechanism—the highest bidder wins and pays the price he bid—does not encourage honest bids. Galton realized an analogous property when a jury is to decide on a money amount or weight: taking the average of their choices as the mechanism "would give a voting power to 'cranks' in proportion to their crankiness."

A judge assigns a grade to a competitor. A very complex set of wishes, opinions, expectations, and anticipations—in theory, his utility function, a complicated expression involving many variables—determines the grade he gives. Note, in particular, that the final grade he wishes a competitor to be awarded, the final grade he believes the competitor merits, and the grade he gives may all be different. In many cases it is natural to assume that the further the jury's grade strays from what a judge wishes the grade to be, the less he likes it. That is, the preferences of each judge over the grades are single-peaked (as implicitly assumed by Galton, explicitly by Black and Moulin). In this case a judge seeks a strategy that will bring the *final grade*—meaning the jury's grade—as close as possible to what he wishes it to be. Whereas the assumption of single-peaked preferences in the traditional model is a very restrictive and unnatural assumption for most applications (Galton's budget problem stands out as a notable exception), assuming that a judge has single-peaked preferences over the grades a competitor is to have is, in contrast, a very reasonable and natural assumption. But it is not necessarily true of all judges. There is a difference between a judge who either honestly believes or simply wishes that a competitor receive a final grade close to his personal assessment, and a judge whose mission is to distort the final grade whatever the value of the performance of a competitor (the judge may have been bribed, or he may be a member of a clique having a particular agenda—recall the two blocs of nations in international figure skating competitions described in chapter 7).

To begin, observe that there is no social grading function F (with aggregation function f) that can prevent every one of the judges from raising the final

grade. For suppose the contrary. By unanimity, $f(0, 0, \ldots, 0) = 0$. Since judge 1 cannot raise the grade, then $0 = f(R, 0, \ldots, 0)$; judge 2 cannot either, so $0 = f(R, 0, \ldots, 0) = f(R, R, \ldots, 0)$; continuing, $0 = f(R, R, \ldots, R) = R$, a contradiction. By a similar argument, all the judges cannot be barred from lowering a final grade. So the potential for a judge to manipulate the final grade certainly exists. This is hardly surprising because otherwise judges would have no influence on the outcome at all.

In the chapters that follow, since every social grading function F has a unique associated continuous aggregation function f, and every such f has a unique F, each is referred to intermittently, the choice falling to that which seems most appropriate for the purpose. The continuity, explicit in the definition, is not always recalled and not always needed.

10.1 Strategy-Proofness in Grading

Experience shows that juries may well include judges who are bribed (recall the Olympic brouhaha over figure skating). However, juries almost certainly contain judges who honestly wish grades to be assigned according to merit, and in certain cases it is perfectly reasonable to assume that all the judges of a jury share this intent.

Juries in wine competitions when the tasting is blind cannot do otherwise, for it is in practice impossible for them to identify particular wines. In a competition to be named the world's best sommelier (or wine master), described in the October 2004 issue of the *Revue du Vin de France*, contestants were asked to identify the country of origin and type of grape of two wines. The first was a Riesling from New Zealand. Four candidates responded as follows: a Chardonnay from South Africa, a Sauvignon from South Africa, an Albariño from Galicia, and an Assyrtico from the Cyclades Islands. The second wine was a Carménère from Chile. The same four candidates identified it as a Merlot from Chile, a Cabernet Franc or Sauvignon from the Loire Valley, a Merlot from Chile, and a Cabernet Sauvignon/Merlot from Chili. Another piece of evidence attesting to the difficulty of identifying wines is an ongoing project devoted to classifying all the world's wines; there are well over 3,000 different categories. Judges of sports competitions who are obliged to publicly announce the grades they give are seriously constrained by the opinion of the public and expert commentators.

Suppose that r is a jury's final grade. A social grading function (SGF) is *strategy-proof-in-grading* if, when a judge's input grade is $r^+ > r$, any change in his input can only lead to a lower grade; and if, when a judge's input grade is $r^- < r$,

any change in his input can only lead to a higher grade. (The specification "in-grading" is often dropped when there can be no confusion as to the meaning of "strategy-proof.")

When it is the case that the more a final grade deviates from the grade a judge wishes it to be, the less he likes it (single-peaked preferences over grades), the utility function is

$$u_j(\mathbf{r}^*, \mathbf{r}, f, \mathcal{C}, \Lambda) = -|r_j^* - f(r_1, \ldots, r_n)|.$$

This implies that it is a *dominant strategy* for a judge to assign the grade he wishes. This means that it is at least as good as any other strategy, and it is strictly better than others in some cases. The reverse implication is not true for some preferences, as will be illustrated in an example below. Moreover, when the judges have preferences that are single-peaked, but they also lean toward being honest rather than not, so the utility function is

$$(u_j(\mathbf{r}^*, \mathbf{r}, f, \mathcal{C}, \Lambda) = -|r_j^* - f(r_1, \ldots, r_n)| - \epsilon|r_j^* - r_j|,$$

for $\epsilon > 0$, then this implies that it is a *strictly dominant strategy* for a judge to assign the grade he wishes; it is a strictly better strategy in all cases.

The use of a strategy-proof-in-grading SGF permits a judge who seeks to grade honestly—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 even know what his preference is between two other grades, when one of them is lower than and the other higher than his true grade. It is a very desirable property.

10.2 Order Functions

There is a class of SGFs whose functions are easily shown to be strategy-proof-in-grading. We call them the *order* SGFs (in another context they are known as the order statistics of the grades given a competitor). Their aggregation functions f are the first or highest grade, the second or second highest grade, ..., the kth highest grade, ..., the nth highest or worst grade.

The kth highest grade is the *kth-order function* f^k.

It is, of course, an aggregation function. To see the truth of the claim of strategy-proofness, suppose the kth-order value, or final grade, is r. A judge who wishes the final grade to be higher can only hope to improve it by increasing the grade he gives, but increasing it changes nothing. Similarly, if he wishes the final grade

to be lower, he can hope to improve it only by decreasing the grade he gives, but decreasing it changes nothing. Finally, if the judge gave the grade r, he is delighted: the final grade is always at least as good and sometimes strictly better than it would have been if he had assigned any other grade. Thus no judge with single-peaked preferences has any incentive to assign a grade different from the honest one, showing that the kth-order function is strategy-proof-in-grading.

In fact, the order functions are the only strategy-proof SGFs.

Theorem 10.1 *The unique strategy-proof-in-grading SGFs are the order functions.*

Proof Let $f(r_1, \ldots, r_n) = r$. Unanimity and monotonicity imply that the value of r must fall between the worst and best grades, $\max r_j \geq r \geq \min r_j$. This fact is used repeatedly here and in the chapters to come.

Suppose the judges assigned the grades $r_1 \geq \cdots \geq r_n$. We claim that

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

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

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

This is true for two reasons. First, when r_j is increased to a higher grade r_j^*, the value of f cannot increase because 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 $f(r_1, \ldots, r_n) = r$,

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

Define $\mathbf{r} = (r_1, \ldots, r_n)$ with $r_1 \geq \cdots \geq r_n$. 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 $r_1 \geq \cdots \geq r_n$, 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 $r^+ > r > r^-$,

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

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 that

$f(r_1, \ldots, r_n) = r_k \quad$ when $r_1 \geq \cdots \geq r_n$

implies

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

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

It must still be shown that this will be true whatever the magnitude of r_k, that is, 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. ■

In fact, the theorem is true without continuity assumptions: the language may be finite—the important practical case—and the aggregation function need not be continuous.

Theorem 10.2 *The unique strategy-proof-in-grading SGFs are the order functions when the language is any finite set of grades.*

Proof Identify the lowest grade with 0, the highest grade with R.

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

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

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

Suppose the language contains at least three 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, $r = 0$, 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 seen in the previous proof,

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

k does not depend on the input $\mathbf{r}$ because it is defined uniquely, so this completes the proof. ∎

As a practical matter it is important to note that even if *all* the judges who think that the final grade should be higher manipulate by increasing their input grades, and *all* those who think that the final grade should be lower decrease their input grades, the final grade remains the same with order functions. Consequently, order functions are strategy-proof for groups of judges having the same interests acting together: they are—in the vocabulary of the literature on voting—*nonmanipulable* by any judge or any coalition of judges, or *group-strategy-proof*. For if a group of judges has given grades above the final grade, a concerted decision to increase all their grades changes nothing; and similarly when a group has given grades below the final grade.

As a technical matter, the central result in (Moulin 1980) may be used to prove theorem 10.1, though the model and spirit are altogether different. Moulin seeks to single out a candidate; we seek to single out a grade (see chapter 5). He characterizes a wider class of methods. When monotonicity is added and candidates are interpreted as grades, Moulin's proof provides another characterization of the order functions, though his proof is considerably more involved.

To see why asking for an SGF that makes honesty a dominant strategy for every judge is less general than asking for strategy-proofness, consider the following example. There are three grades, $\alpha \succ \beta \succ \gamma$, and two judges, and the SGF is defined by

$$f(\delta, \delta) = \delta, \qquad f(\alpha, \beta) = \alpha, \qquad f(\alpha, \gamma) = \alpha, \qquad f(\beta, \gamma) = \gamma,$$

for $\delta = \alpha, \beta, \gamma$. By definition, f is anonymous. f is an aggregation function but not an order function. When $f(\beta, \gamma) = \gamma$, what is the point of view of the judge who assigned the grade β? If she maintains the β, the final grade is γ, whereas if she changes it to α, the final grade jumps to α. What does she prefer, α or γ? If she prefers α, she should obviously assign the grade α. Strategy-proofness bars her from increasing the final grade when she believes it should be higher (and decreasing it when she believes it should be lower), however she may compare α to γ. In other words, an SGF is strategy-proof-in-grading if and only if it is a dominant strategy for a judge to be honest for *every* utility function that is consistent with her single-peaked preferences.

10.3 Minimizing Manipulation

Not all judges send messages (vote) according to their beliefs, as has been amply documented. The problem remains: what social grading functions should be used to determine the grades of figure skaters, divers, pianists, and others when judges may manipulate the grades they announce? The ideal is a strategy-proof SGF, one that encourages every judge, honest or not, to assign the grade he thinks is the correct one. Regrettably, this ideal is impossible to achieve. So, we naturally turn to the question, How can the potential impact of manipulating in the assignment of grades best be countered? Since the facts make attaining the ideal impossible, the demands of the practical world suggest that the ideal be realized "as near as may be."[1]

To manipulate successfully a judge must be able to raise or lower the final grade by changing the grade he assigns. In some situations a judge can only change the final grade by increasing his grade, in others only by decreasing his grade. When grades are sufficiently different, the judge who gives the lowest grade can always increase the final grade, whereas the judge who gives the highest grade can always decrease it. For suppose $f(r_1, \ldots, r_n) = r$ and $R > r_1 > \cdots > r_n > 0$. Then,

$$f(r_1, \ldots, r_{n-1}, R) = f(R, r_1, \ldots, r_{n-1}) > f(r_1, r_2, \ldots, r_n) = r,$$

since f is anonymous and monotonic, showing the first claim. The second is established similarly.

Judges 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 judges. It is important, therefore, to identify methods of grading that eliminate as much as possible the existence of such judges. The order functions clearly have the property that at most one judge is able to both raise and lower a final grade. Moreover, if two or more judges happen to assign the final grade, then no judge is able to both raise and lower the final grade. It happens that they are the only SGFs with this property.

Theorem 10.3 *The order functions are the unique SGFs for which, for any* $\mathbf{r}$*, at most one judge may both increase and decrease a final grade.*

1. This was precisely Daniel Webster's point when he addressed the Senate in 1832 on another problem: "The Constitution, therefore, must be understood, not as enjoying an absolute relative equality, because that would be demanding an impossibility, but as requiring of Congress to make the apportionment of Representatives among the several States according to their respective numbers, *as near as may be*. That which cannot be done perfectly must be done in a manner as near perfection as can be" (1832, 107).

Proof Suppose that f is an aggregation function that permits at most one judge to both increase and decrease a final grade. Consider

$$f(r_1, \ldots, r_n) = r \quad \text{for some } r_1 \geq \cdots \geq r_n,$$

with $\mathbf{r} = (r_1, \ldots, r_n)$.

If judge j is able to decrease the final grade by decreasing his grade r_j, then anonymity and monotonicity together imply that any judge k can do this when $r_k \geq r_j$. Similarly, if judge j is able to increase the final grade by increasing his grade, then any judge k can do this when $r_k \leq r_j$. Let

$$I^-(\mathbf{r}) = \{j \in \mathcal{J} : j \text{ can decrease the final grade}\}$$

and

$$I^+(\mathbf{r}) = \{j \in \mathcal{J} : j \text{ can increase the final grade}\}.$$

By hypothesis, $I^-(\mathbf{r}) \cap I^+(\mathbf{r})$ contains at most one judge.

It will be shown that when $j \in I^-(\mathbf{r})$, then $r_j \geq r$, and symmetrically, when $j \in I^+(\mathbf{r})$, then $r_j \leq r$. Thus, $r_j < r$ implies $j \notin I^-(\mathbf{r})$, and $r_j > r$ implies $j \notin I^+(\mathbf{r})$: by definition f is strategy-proof-in-grading. So theorem 10.1 implies the result.

Suppose i is the judge who gave the lowest grade among those able to decrease the final grade, that is, $r_i = \min_j r_j$, $j \in I^-(\mathbf{r})$. It must be shown that $r_i \geq r$. By hypothesis,

$$f(r_1, \ldots, r_{i-1}, 0, r_{i+1}, \ldots, r_n) = r^0 < r.$$

If $r_{i-1} > r_i$, let $\tilde{r}_{i-1} > 0$ be the smallest grade of judge $i-1$ for which $f(r_1, \ldots, \tilde{r}_{i-1}, r_i, \ldots, r_n) = r$.

Suppose $\tilde{r}_{i-1} > r_i$. The continuity of f implies that there exists a small $\epsilon > 0$ so that

$$f(r_1, \ldots, r^*_{i-1}, r_i, \ldots, r_n) = r' > r^0,$$

for $r^*_{i-1} = \tilde{r}_{i-1} - \epsilon > r_i$. Thus for $(r_1, \ldots, r^*_{i-1}, r_i, \ldots, r_n)$ judge $i-1$ can both increase and decrease the final grade, implying that all those giving lower grades can only increase the final score. Therefore,

$$f(r_1, \ldots, r^*_{i-1}, 0, r_{i+1}, \ldots, r_n) = r'.$$

But the monotonicity of f implies

$$r' = f(r_1, \ldots, r^*_{i-1}, 0, r_{i+1}, \ldots, r_n) \leq f(r_1, \ldots, r_{i-1}, 0, r_{i+1}, \ldots, r_n) = r^0,$$

a contradiction.

In consequence $\tilde{r}_{i-1} \leq r_i$, so replacing r_{i-1} by r_i yields

$$f(r_1, \ldots, r_{i-2}, r_i, r_i, r_{i+1}, \ldots, r_n) = r.$$

Repeat the same argument one judge at a time to conclude that

$$f(r_i, \ldots, r_i, r_{i+1}, \ldots, r_n) = r,$$

where judges $1, 2, \ldots i$ all give the same grade r_i. If $r_i < r$, then $r > r_i \geq r_{i+1} \geq \cdots \geq r_n$, implying $f(r_i, \ldots, r_i, r_{i+1}, \ldots, r_n) < r$, a contradiction. Thus $r_i \geq r$, as was to be shown.

Suppose h is the judge who gave the highest grade among those able to increase the final grade, that is, $r_h = \max_j r_j$, $j \in I^+(\mathbf{r})$. A symmetric argument shows that $r_h \leq r$, and completes the proof. ∎

An immediate consequence of theorem 10.3 is the following:

Corollary *There exists no SGF that, for every profile of grades, prevents every judge from both increasing and decreasing the final grade.*

Theorem 10.3 and its corollary are not true when the language is finite. Take a problem with three grades, $\alpha \succ \beta \succ \gamma$, and two judges, and let the SGF be defined by

$$f(\delta, \delta) = \delta, \qquad f(\alpha, \beta) = \beta, \qquad f(\alpha, \gamma) = \beta, \qquad f(\beta, \gamma) = \gamma.$$

f is not an order function. Whatever the two grades assigned, no judge can both increase and decrease the final score. For when the vote is unanimous, the only possible exception would be for $f(\beta, \beta)$, but in this case neither judge can raise the final grade. Otherwise, a judge whose input is α can only lower the final grade; a judge whose input is γ can only raise it; and when $f(\alpha, \beta) = \beta$ or $f(\beta, \gamma) = \gamma$, the judge with input β can only raise it.

There is another way of looking at the problem of limiting the possibility of manipulation. Given an aggregation function f and a profile of grades $\mathbf{r} = (r_1, \ldots, r_n)$, let $\mu^-\big(f(\mathbf{r})\big)$ be number of judges who can decrease the final grade, $\mu^+\big(f(\mathbf{r})\big)$ be the number of judges who can increase the final grade, and $\mu\big(f(\mathbf{r})\big) = \mu^-\big(f(\mathbf{r})\big) + \mu^+\big(f(\mathbf{r})\big)$ be their sum.

The *manipulability of* f is $\mu(f)$, the largest value of $\mu\big(f(\mathbf{r})\big)$ over all possible profiles $\mathbf{r} = (r_1, \ldots, r_n)$,

$$\mu(f) = \max_{\mathbf{r}=(r_1,\ldots,r_n)} \mu\big(f(\mathbf{r})\big).$$

At worst, a judge can both increase and decrease the final grade, so the manipulability of f can be no more than $2n$, that is, $\mu(f) \leq 2n$. In particular, when

f is taken to be the arithmetic mean of the grades (as with Borda's method) the manipulability is maximized, $\mu(f) = 2n$. On the other hand, when f is the kth-order function, $\mu(f) = n + 1$. When it is impossible to entirely eliminate manipulation, a natural idea is to try to minimize the possibility of manipulation.

Theorem 10.4 *The only SGFs that minimize manipulability are the order functions.*

Proof Let f be any aggregation function whose manipulability is at most $n + 1$, and let $\mathbf{r}$ be any input. As noted, if a judge can reduce (can augment) the final grade, then so can any judge who assigns a higher (a lower) grade. Take $I^-(\mathbf{r})$ and $I^+(\mathbf{r})$ as defined in the proof of theorem 10.3, respectively the sets of judges who can decrease and increase the final grade. If these sets have more than one judge in common, $|I^-(\mathbf{r}) \cap I^+(\mathbf{r})| \geq 2$, then $\mu(f) \geq n + 2$ because each judge in the intersection can both increase and decrease the final grade, and each one of the other judges can at least increase or decrease the final grade. Therefore, at most one judge can be in both sets, $|I^-(\mathbf{r}) \cap I^+(\mathbf{r})| \leq 1$, and theorem 10.3 implies f must be an order function. ■

10.4 Implications

The aim of this chapter is to identify methods of grading that are immune to strategic manipulation of the judges whenever possible; and when not, to limit strategic manipulation as much as possible. In each case the answer is to choose the order functions. They reflect the intuition of many practical people who have felt that the best and worst grades, or the best two and worst two, should be eliminated to combat manipulation. When a judge seeks to award final grades that are merited—in some situations, such as tasting wines, the assumption is reasonable; in others, judges may in fact simply prefer being honest—the order SGFs are totally immune to manipulation. There exists no SGF that completely prevents strategic manipulation, but the use of an order function minimizes the probability that a judge can manipulate. For, a priori, a judge who manipulates wishes to increase the final grade with probability $\frac{1}{2}$ and wishes to decrease it with probability $\frac{1}{2}$. And, a priori, any one of the n judges may cheat. Thus, when the input of grades is $\mathbf{r}$, the probability for a judge to manipulate successfully when using an SGF with aggregation function f is

$$\left(\frac{1}{2}\right)\left(\frac{\mu^-\big(f(\mathbf{r})\big)}{n}\right) + \left(\frac{1}{2}\right)\left(\frac{\mu^+\big(f(\mathbf{r})\big)}{n}\right) = \frac{\mu\big(f(\mathbf{r})\big)}{2n}.$$

One would like this probability to be as small as possible in the worst case; this amounts to finding an f that solves

$$\min_{f} \max_{\mathbf{r}=(r_1,\ldots,r_n)} \mu\big(f(\mathbf{r})\big) = \min_{f} \mu(f),$$

that is, minimizes the manipulability. Thus the minimum probability is slightly over one-half, $\frac{n+1}{2n}$, and is achieved only when f is an order function.

When the judges of a jury are obligated to publicly announce their grades, it is reasonable to suppose that whatever exaggeration takes place is prudent: the increase is small when the intent is to raise the grade, the decrease is small when the intent is to lower the grade. If the first-order function (the maximum of all grades) is used and all the judges give the same grade, then all n of them can manipulate to increase the final grade. Similarly, when the nth-order function is used and all give the same grade, then all n of them can manipulate to decrease the final grade. However, when all the grades are different, $r_1 > \cdots > r_n$, the kth-order function allows only one judge to manipulate prudently, namely, the judge with the kth highest grade. In this case, the probability of manipulation is reduced to $\frac{1}{n}$.

The order functions necessarily give answers that are in the original language, so no enrichment of language is necessary. This is very important for many applications where using numbers may be unnatural or where the numbers are confusing to judges of different cultures (recall the profusion of scales used by different nations in reporting the grades of students). Moreover, the properties that characterize them in the context of a continuous language obviously hold in the context of a discrete language.

These ideas have an immediate application. They provide a partial justification for the Borda-majority method (see chapter 5), which was actually used by the International Skating Union (see chapter 7). In the context of this chapter, that method is the SGF determined by the kth-order function, where k is the middlemost grade (when there are $2k + 1$ judges) and judges or voters are forced to grade every competitor differently with a number that depends on the number of competitors. The middlemost grade is, of course, particularly appealing because a majority of the judges agree that at least that grade should be given, and at the same time a majority agree that at most that grade should be given. It is given special attention in the chapters that follow.

11 Meaningfulness

The establishment of a common measure between mind and mind.
—J. H. Newman

The languages used to grade students vary from nation to nation, the ranges of numbers used to evaluate flautists, pianists, and wines change from competition to competition, and the highest and lowest scores earned by Olympic competitors differ from one to another athletic discipline. When numbers are used, the differing languages are not related by a simple change in scale. For example, a school grade of 10 on a scale of [0, 20] in France has an entirely different meaning than a 50 on a scale of [0, 100] in the United States (indeed, it may be more accurate to say that the scale in the United States is [50, 100], as Parker the wine critic suggests). When the language is defined by letters or descriptive phrases, any association with numbers is arbitrary. Social grading functions transform the measures assigned by each member of the jury into a single measure of the jury. Is this single measure meaningful? Why should a particular SGF be accepted as faithfully representing the will of the jury? Several different qualitative properties are advanced that address this question. Each uniquely characterizes the order functions among all possible aggregation functions.

11.1 Reinforcement and Conformity

Suppose that judges seek to assign the true grade, and imagine that after the members of a jury have assigned their grades, some judge wishes to revise her grade by assigning a grade closer to the final grade of the jury. Surely then the jury's revised grade should be the same because there is a greater consensus for that grade.

A social grading function (SGF) with aggregation function f is *reinforcing* when

$$f(r_1, \ldots, r_{k-1}, r_k, r_{k+1}, \ldots, r_n) = r$$

and

$$r_k > \hat{r}_k \geq r \quad \text{or} \quad r \geq \hat{r}_k > r_k$$

implies

$$f(r_1, \ldots, r_{k-1}, \hat{r}_k, r_{k+1}, \ldots, r_n) = r.$$

It is immediately obvious that the order functions are reinforcing.

Theorem 11.1 *The unique reinforcing SGFs are the order functions.*

Proof Suppose f is reinforcing, and let $f(r_1, \ldots, r_k, \ldots, r_n) = r$. Then

$$f(r_1, \ldots, r_{k-1}, \hat{r}_k, r_{k+1}, \ldots, r_n) = r \quad \text{for } r_k > \hat{r}_k > r.$$

If for some $\delta > 0$,

$$f(r_1, \ldots, r_{k-1}, r_k + \delta, r_{k+1}, \ldots, r_n) = r + \epsilon, \ \epsilon > 0,$$

then since f is reinforcing,

$$f(r_1, \ldots, r_{k-1}, r + \epsilon, r_{k+1}, \ldots, r_n) = r + \epsilon,$$

contradicting the next-to-last equation. Therefore, no matter how much r_k is increased above r, the value of f remains the same. If $r > r_k$, a similar argument shows r_k can be decreased without changing the value of f. This means that f is strategy-proof-in-grading, and theorem 10.1 shows that f must be an order function. ■

A second property is a natural generalization of unanimity. If every judge assigns a grade in a subset of the grades, then this property asserts that the final grade should belong to that subset. Or it may be seen as saying that no enrichment of the language in which the grades are assigned is necessary to determine the final grade.

An SGF with aggregation function f *conforms* with the assigned grades if

$$\{r_1, \ldots, r_n\} \subset S \quad \text{implies } f(r_1, \ldots, r_n) \in S.$$

Again, this property is obviously true for order functions.

Theorem 11.2 *The unique SGFs that conform with the assigned grades are the order functions.*

Proof That f conforms with the grades implies that for all $r_1 > r_2 > \cdots > r_n$, $f(r_1, \ldots, r_n) = r_k$ for some k. By the continuity of f, k should be the same for all profiles, as was seen in the proof of theorem 10.1, so f must be the kth-order function. ∎

11.2 Language-Consistency

The particular language used in grading should make no difference to the ultimate outcomes. An aggregation function should give equivalent grades when one language is faithfully translated into another. This is the meaningfulness problem of measurement theory in the context of a jury decision (Krantz et al. 1971; Pfanzagl 1971; Roberts 1979).

An SGF with aggregation function f is *language-consistent* if

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

for all increasing, continuous functions

$$\phi : [0, R] \to [\underline{R}, \overline{R}], \qquad \phi(0) = \underline{R}, \qquad \phi(R) = \overline{R}.$$

The fact that the order functions are language-consistent is immediate: the kth largest value remains the kth largest value under increasing, continuous transformations. A particularly simple proof is given here of the reverse implication, although the result is well known. One of the main points we would like to emphasize in developing this theory is the ease of its ideas, its theorems and their proofs. For simplicity, we systematically take a slightly less demanding definition of language-consistency, namely, we assume $\underline{R} = 0$ and $\overline{R} = R$: the results are the same because the order functions clearly satisfy the more demanding definition.

Theorem 11.3 *The unique SGFs that are language-consistent are the order functions.*

Proof Suppose f is language-consistent and $R > r_1 > \ldots > r_n > 0$. The property is used to construct a sequence of continuous, increasing functions that converge to the step function defined by

$$\psi(x) = \begin{cases} R & \text{if } x = R \\ r_1 & \text{if } R > x \geq r_1 \\ \vdots & \vdots \\ r_n & \text{if } r_{n-1} > x \geq r_n \\ 0 & \text{if } r_n > x \geq 0. \end{cases}$$

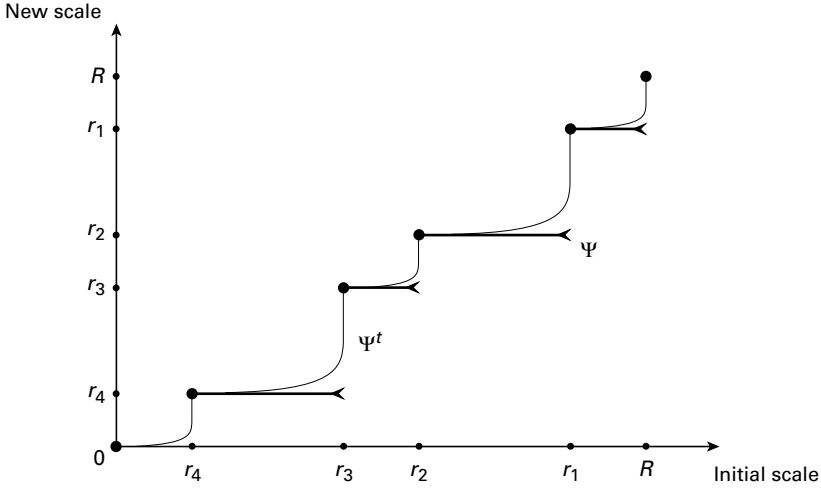

Figure 11.1
The function ψ^t

Define the function ψ^t on each interval $[r_{i+1}, r_i)$, as follows (see figure 11.1):

$$\psi^t(x) = r_{i+1} + \alpha^t(r_i - r_{i+1}) \quad \text{when } x = r_{i+1} + \alpha(r_i - r_{i+1}) \quad \text{for } 0 \leq \alpha < 1,$$

where $r_0 = R$ and $r_{n+1} = 0$. Note that ψ^t is strictly increasing and continuous for every t, and $\psi^t(r_i) = r_i$. By hypothesis

$$f(r_1, \ldots, r_n) = f\big(\psi^t(r_1), \ldots, \psi^t(r_n)\big) = \psi^t\big(f(r_1, \ldots, r_n)\big),$$

whereas in the limit, as $t \to \infty$,

$$\psi^t\big(f(r_1, \ldots, r_n)\big) \to \{r_1, \ldots, r_n\},$$

that is, in the limit ψ^t must be one of the values r_i. But this means that f conforms with one of the assigned grades, so by theorem 11.2 it must be an order function. ■

11.3 Order-Consistency

When do two different aggregation functions induce the same order on the final grades? The answer is simple: when they are the same.

Theorem 11.4 *Two aggregation functions f and g that define the same order on the final grades are identical.*

Proof To say that f and g impose the same order means that $f(r_1, \ldots, r_n) \geq f(s_1, \ldots, s_n)$ for two sets of grades if and only if $g(r_1, \ldots, r_n) \geq g(s_1, \ldots, s_n)$. This is equivalent to the same statement with the inequality ($\geq$) replaced by equality ($=$). Suppose $f(r_1, \ldots, r_n) = r$, so that $f(r, \ldots, r) = r$. Then, since f and g define the same order, $f(r_1, \ldots, r_n) = f(r, \ldots, r)$ if and only if $g(r_1, \ldots, r_n) = g(r, \ldots, r)$. But $g(r, \ldots, r) = r$, so $f(r_1, \ldots, r_n) = r$ if and only if $g(r_1, \ldots, r_n) = r$, that is, $f = g$. ∎

Grades are ordinal. Another way of assuring that an SGF is meaningful is to ask that comparisons between two aggregations should be preserved under any increasing, continuous transformation; in the language of measurement theory, an SGF should be "comparison meaningful." For example, suppose a Franco-American jury is to assign grades to students, and each member is asked to give a grade in both languages, the French and the American grading systems. Then the aggregate French grades should rank the students in the same order as the aggregate American grades. The outcomes classify by comparing the magnitudes on an ordinal scale; this is ordinal measurement in the levels of the hierarchy defined by Stevens. In a word, a transformation from one language to another that preserves order should not change the final outcomes.

An SGF with aggregation function f is *order-consistent* if

$$f(r_1, \ldots, r_n) \geq f(s_1, \ldots, s_n)$$

implies

$$f\big(\phi(r_1), \ldots, \phi(r_n)\big) \geq f\big(\phi(s_1), \ldots, \phi(s_n)\big)$$

for all increasing, continuous functions

$$\phi : [0, R] \to [\underline{R}, \overline{R}], \qquad \phi(0) = \underline{R}, \qquad \phi(R) = \overline{R}.$$

It is obvious that the order functions are order-consistent; again, the property characterizes them.

Theorem 11.5a *The unique SGFs that are order-consistent are the order functions.*

Proof This is really a corollary, but its importance merits calling it a theorem. It suffices to observe, first, that language-consistency implies order-consistency. Second, if $f(r_1, \ldots, r_n) = r$, then $f(r_1, \ldots, r_n) = f(r, \ldots, r)$, implying, when f is order-consistent, that

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

so f is language-consistent. ■

What happens when instead of an SGF, a social ranking function (SRF) is used. An SRF 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 a monotonic transformation ϕ of the grades of each judge. An SRF $\succeq_S$ is *monotonic* if every judge j gives at least as high a grade to a candidate A as to a candidate B, implies A is ranked at least as high B, and if every judge gives A a higher grade than B, then A is ranked above B.

Theorem 11.5b *If an SRF is order-consistent and monotonic, then there exists an order function f such that $f(A) > f(B)$ implies $A \succ_S B$.*

This theorem, established in Hammond (1976) and d'Aspremont and Gevers (1977), is the mirror image of a result concerning social welfare orderings. Repeated application of the theorem shows that an SRF is order-consistent and monotonic if and only if there is a sequence of order functions that decide: if the first does not strictly rank the candidates, the second does; if the second doesn't either, then the third does; and so on.

11.4 The Meaning of Arrow's Theorem

The idea of measuring performances or alternatives to enable them to be compared depends crucially on the judges' using a common language in assigning grades. When each judge uses his own language, the only meaningful information concerning a judge's evaluation is the order of his preferences. What happens when there is no common language? For an SGF to be meaningful in this case, comparisons between two aggregations should be preserved under different increasing, continuous transformations of each of the judge's grades.

An SGF with aggregation function f is *preference-consistent* if

$$f(r_1, \ldots, r_n) \geq f(s_1, \ldots, s_n)$$

implies

$$f\big(\phi_1(r_1), \ldots, \phi_n(r_n)\big) \geq f\big(\phi_1(s_1), \ldots, \phi_n(s_n)\big)$$

for all increasing, continuous functions

$$\phi_j : [0, R] \to [\underline{R}, \overline{R}], \qquad \phi_j(0) = \underline{R}, \qquad \phi_j(R) = \overline{R}.$$

Lemma *There exists no preference-consistent SGF.*

Proof If an SGF is preference-consistent, then it must be language-consistent. The only language-consistent SGFs are the order functions. But the order functions are assuredly not preference-consistent. Q.E.D.[1] ∎

This impossibility encompasses the essence of Arrow's theorem, though it is rather less precise. To see this more clearly, define a *weak social grading function* (WSGF) F to be a continuous *weak aggregation function* $f : (r_1, \ldots, r_n) \to [0, R]$ when the condition of anonymity among the judges is dropped. Moreover, call a weak aggregation function f *dictatorial* if $f(r_1, \ldots, r_n) = r_k$ for all possible $(r_1, \ldots, r_n)$ and some one judge k.

Theorem 11.6a (Arrow's Impossibility for SGFs) *The unique weak social grading functions that are preference-consistent are dictatorial.*

Proof Let F be a WSGF with weak aggregation function f, and consider the grades $s_1, \ldots, s_n$ when $s_i \neq s_j$ for $i \neq j$ and all $s_i \in (0, R)$. Rearrange them in decreasing order, say, $R > s_{(1)} > \cdots > s_{(n)} > 0$. Since preference-consistency implies language-consistency, the proof of theorem 11.3 applied to $R > s_{(1)} > \cdots > s_{(n)} > 0$ shows that $f(s_1, \ldots, s_n) = s_k$ for some k.

Choose any $i \neq k$, and suppose $s_i > s_k$. The continuity of f and the fact that f must take one of the values of its arguments together imply that for any $s_i^* > s_k$,

$$f(s_1, \ldots, s_{i-1}, s_i^*, s_{i+1}, \ldots, s_n) = s_k.$$

In particular,

$$f(s_1, \ldots, s_n) = f(s_1, \ldots, s_{i-1}, s_i + \epsilon, s_{i+1}, \ldots, s_n) = s_k.$$

The same equation is now shown to hold for any $s_i^* \leq s_k$. Let ϕ_i be any strictly increasing, continuous function for which

$$\phi_i(x) = \begin{cases} 0 & \text{if } x = 0 \\ s_i^* & \text{if } x = s_i \\ s_k + \epsilon & \text{if } x = s_i + \epsilon \\ R & \text{if } x = R. \end{cases}$$

Preference-consistency then implies

$$f\big(s_1, \ldots, s_{i-1}, \phi_i(s_i), s_{i+1}, \ldots, s_n\big) = f\big(s_1, \ldots, s_{i-1}, \phi_i(s_i + \epsilon), s_{i+1}, \ldots, s_n\big),$$

1. Language-, order- and preference-consistency are known under different names in the literature on measurement theory; also, theorem 11.4 and its corollaries are known in one guise or another (see, e.g., Orlov 1981). However, the context and spirit of the approach are completely different, and our proofs seem simpler. The theorems we give are in fact true with less demanding hypotheses.

so that for any $s_i^* \leq s_k$,

$$f(s_1, \ldots, s_{i-1}, s_i^*, s_{i+1}, \ldots, s_n) = f(s_1, \ldots, s_{i-1}, s_k + \epsilon, s_{i+1}, \ldots, s_n) = s_k,$$

the last equation having been already established. Therefore, the value of f remains the same whatever may be the choice of value of s_i. On the other hand, if judge k changes the grade s_k, then by continuity (since the value of f is the value of one of its arguments), f must continue to coincide with s_k. Repeating the same argument for every s_i, $i \neq k$, and noting continuity implies the same holds for any set $s_1, \ldots, s_n \in [0, R]$, completes the proof.[2] ■

The theorem is, of course, *not* exactly Arrow's impossibility. There are technical differences of various types. In particular, the language in Arrow's scheme of things has as many words as there are candidates or alternatives, and the theorem holds only when there are at least three candidates or alternatives. And yet, the spirit is the same: the only aggregation of the judges' preferences into a jury's preference that satisfies unanimity and independence of irrelevant alternatives is the dictatorial function.

Consider a *weak social ranking function* (WSRF) to better seize the meaning of Arrow's theorem. It is an SRF for which anonymity is dropped. A WSRF 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 a monotonic transformation ϕ_j of the grades of each judge j. Arrow's theorem with possible ties in the inputs (see section 3.2) implies the following.

Theorem 11.6b (Arrow's Impossibility for SRFs) *A preference-consistent, monotonic WSRF is dictatorial.*

It is not independence of irrelevant alternatives that causes problems. IIA is automatically satisfied when the scale of grades is absolute for each *individual* voter or judge because in that case the absence or presence of any competitor does not change the voters' or judges' true evaluations of any competitor (by theorem 9.2). For the results to be meaningful, however, the scale of grades must be common to *all* voters or judges.

Arrow's theorem shows once again that in order to arrive at meaningful final grades and a final ranking it is *essential* for judges to share an absolute common language; otherwise, the unique meaningful method of election is dictatorial. But that only stands to reason: imagine the leaders of the world's greatest powers negotiating an agreement with no translators and no common language.

2. Theorem 11.6a is proved in Kim (1990). The proof given there is more involved and much longer.

Thus all methods based on the traditional model are meaningless. Hence, the only possible meaningful methods of election must be based on a new paradigm.

In practice, this implies that languages of grades used in the majority judgment should be designed to embody as much as possible absolute meanings for each judge and common meanings among all judges.

12 Majority-Grade

The object of . . . an election is to select, if possible, some candidate who shall, in the opinion of a majority of the electors, be most fit for the post . . . [The] fundamental condition . . . is that the method adopted must not be capable of bringing about a result which is contrary to the wishes of the majority.
—E. J. Nanson

The previous chapters have presented mounting evidence that argues for a jury to arrive at a final grade by using one of the order functions. A completely different set of arguments will single out one function that happens to be an order function. Sir Francis Galton pointed in the right direction a century ago:

I wish to point out that the estimate to which least objection can be raised is the middlemost estimate, the number of votes that it is too high being exactly balanced by the number of votes that it is too low. Every other estimate is condemned by a majority of voters as being either too high or too low, the middlemost alone escaping this condemnation. The number of voters may be odd or even. If odd, there is one middlemost value . . . If the number of voters be even, there are two middlemost values, the mean of which must be taken. (Galton 1907a, 414)

12.1 Middlemost Aggregation Functions

There seems to be no particularly good argument for advancing the mean of the middlemost values in the even case—Galton gave none—rather than any other value in the interval. Given that $r_1 \geq \cdots \geq r_n$, the *middlemost grade(s)* and *interval* are $r_{(n+1)/2}$ when n is odd and $[r_{n/2}, r_{(n+2)/2}]$ when n is even. Recall that the kth-order function is f^k.

A social grading function (SGF) is *middlemost* if it is defined by a *middlemost aggregation function* f, where for $r_1 \geq \cdots \geq r_n$,

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

and

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

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

An aggregation function f *depends only on the middlemost interval* if $f(r_1, \ldots, r_n) = f(s_1, \ldots, s_n)$ when the middlemost intervals of the grades $\mathbf{r} = (r_1, \ldots, r_n)$ and the grades $\mathbf{s} = (s_1, \ldots, s_n)$ are the same.

Four separate and quite different justifications for choosing the middlemost aggregation functions are given, and then it is shown why there is a single best choice—the majority-grade—whatever may be the parity of n.

12.2 Majority Decision

It is at once evident that if a majority of the judges assign a grade r, then the middlemost order functions assign a final grade of r. This is a very important property. If an absolute majority of the judges agree on a grade, then that should surely be the final grade. The objection to the Sydney Wine Competition's system raised by its director (see chapter 7) focused on precisely this point: a majority of the jury could favor bestowing a gold medal, yet the grade assigned by one judge could deny it. The fact is that the middlemost are the only aggregation functions that satisfy this criterion.

Theorem 12.1 *The unique aggregation functions that assign a final grade of r when a majority of judges assign r are the middlemost.*

Proof The proof only invokes weak monotonicity and anonymity. Suppose f is an aggregation function that assigns r when a majority assigns r, where $r_1 \geq \cdots \geq r_n$. If n is odd,

$$r_{(n+1)/2} = f(\overbrace{r_1, \ldots, r_{(n-1)/2}}^{(n-1)/2}, \overbrace{r_{(n+1)/2}, \ldots, r_{(n+1)/2}}^{(n+1)/2}) \geq f(r_1, \ldots, r_n),$$

and

$$f(r_1, \ldots, r_n) \geq f(\overbrace{r_{(n+1)/2}, \ldots, r_{(n+1)/2}}^{(n+1)/2}, \overbrace{r_{(n+3)/2}, \ldots, r_n}^{(n-1)/2}) = r_{(n+1)/2},$$

so $f(r_1, \ldots, r_n) = r_{(n+1)/2}$.

If n is even,

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

and

$$f(r_1, \ldots, r_n) \geq f(\overbrace{r_{(n+2)/2}, \ldots, r_{(n+2)/2}}^{(n+2)/2}, \overbrace{r_{(n+4)/2}, \ldots, r_n}^{(n-2)/2}) = r_{(n+2)/2},$$

so $r_{n/2} \geq f(r_1, \ldots, r_n) \geq r_{(n+2)/2}$. ■

12.3 Minimizing Cheating

When a judge's objective is to give the grades he believes competitors merit, as is invariably the case in wine competitions when tastings are completely blind, all the order functions are for him strategy-proof, including, of course, the middlemost order functions. But when a judge's objective is not necessarily to give the grades he believes competitors merit, it is important to limit his strategic manipulation as much as possible. The first-order function—or the highest of the judges' grades—allows every judge to cheat upward, that is, ascribe however high a grade he wishes, but bars every judge from cheating downward, that is, from lowering any grade. In contrast, the nth-order function—or lowest of the judges' grades—allows every judge to cheat downward, that is, enforce however low a grade he wishes, but bars every judge from cheating upward, that is, from increasing any grade. The kth-order function allows $n - k + 1$ judges to increase the final grade and k to decrease it (by perhaps greatly exaggerating the grade assigned either upward or downward). A person tempted to approach a judge to ask that a grade be modified (a briber) presumably wishes the final grades of some competitors or alternatives to be either increased or decreased. It is clearly desirable to thwart potential bribers to the greatest extent possible. This is accomplished by an aggregation function f for which the probability that a briber can find a judge who can successfully do his bidding—increase the final grade when the briber wishes it increased, decrease the final grade when the briber wishes it decreased—is minimal.

Let $\mu^-\big(f(\mathbf{r})\big)$ be the number of judges who can decrease the final grade and $\mu^+\big(f(\mathbf{r})\big)$ be the number of judges who can increase the final grade for a profile of grades $\mathbf{r} = (r_1, \ldots, r_n)$. The fact that $\max_{\mathbf{r}} \{\mu^-\big(f(\mathbf{r})\big) + \mu^+\big(f(\mathbf{r})\big)\}$ is minimized by the order functions was shown in theorem 10.4.

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

function is sought that minimizes the probability that a judge may be found who can effectively raise or lower the grade in the worst case.

The probability of cheating $Ch(f)$ with an aggregation function f is

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

Theorem 12.2 *The unique aggregation 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 aggregation function f,

$$\max_{\mathbf{r}} \max_{0\le\lambda\le 1} \left\{\lambda\mu^+\big(f(\mathbf{r})\big) + (1-\lambda)\mu^-\big(f(\mathbf{r})\big)\right\}$$
$$\ge \max_{\mathbf{r}} \left\{\frac{1}{2}\mu^+\big(f(\mathbf{r})\big) + \frac{1}{2}\mu^-\big(f(\mathbf{r})\big)\right\} \ge \frac{n+1}{2}.$$

(The last inequality is established in section 10.3.) Thus it suffices to show that $Ch(f^{(n+1)/2}) \le \frac{n+1}{2n}$. But that follows because neither $\mu^+\big(f^{(n+1)/2}(\mathbf{r})\big)$ nor $\mu^-\big(f^{(n+1)/2}(\mathbf{r})\big)$ is greater than $\frac{n+1}{2}$ (equality holding when $r_1 > \cdots > r_n$).

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

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

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 f is any aggregation function for which $Ch(f) \le \frac{n+2}{2n}$, so that

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

Take $r_1 > \cdots > r_n$, 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 $r_k \ge r_j$ 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 $Ch(f) = \frac{n+2}{2n}$. ■

Notice that 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 cuts cheating in half.

12.4 Maximizing Social Welfare

Classical utilitarianism calls for the greatest good for the greatest number. In that view of the world, the utility of individuals may be measured, and society's utility is merely the sum of the utilities of its members. Can that message or approach be interpreted in the context of grading?

Focus on a single judge. Suppose there exists a distance function d that measures the judge's discontent: when the judge assigns the grade r and the final grade is s, his disutility is $d(r, s)$. Thus d enjoys the following properties: $d(r, r) = 0$, $d(r, s) = d(s, r)$, and $r < s < t$ implies $d(r, s) + d(s, t) = d(r, t)$. The last equation simply says that the disutility of a judge who believes the grade should be r when the final grade is t equals his disutility when the final grade is s plus his disutility when he believes it should be s when the final grade is t. This accommodates the possibility that, for example, on a scale of 0 to 100, $d(98, 99) = 5d(75, 76)$. Another interpretation focuses on the performances of the competitors: the improvement in going from a grade of r to s plus the improvement of going from s to t equals that of going from r to t. It is a simple matter to define a continuous, strictly increasing function $\phi : [0, R] \to \Re$ such that $d(r, s) = |\phi(r) - \phi(s)|$, namely,

$$\phi(s) = \begin{cases} d(\frac{R}{2}, s) & \text{if } s \geq \frac{R}{2} \\ -d(\frac{R}{2}, s) & \text{if } s \leq \frac{R}{2}. \end{cases}$$

If one's intuition believes decreasing low grades to 0 is not that costly, then it suffices to define ϕ by $\phi(s) = d(0, s)$. The definition permits situations where going from a low grade to 0 is worth $-\infty$ and from a high grade to R is worth $+\infty$. But that would require an unbounded language, which has not been considered (and is not a practical idea). The function $|\phi(r) - \phi(s)|$—think of r as the final grade and s as a judge's input—is single-peaked. On the other hand, other single-peaked functions are not distance functions, for example, $(r - s)^2$, and so are not justified.

Assume that distance is the same to all the judges.

An SGF with aggregation function f *maximizes the social welfare* when the final grade $f(r_1, \ldots, r_n) = r$ minimizes the total disutility of all the judges, $\Delta(r) = \sum_{j \in \mathcal{J}} d(r, r_j)$.

Theorem 12.3 *The unique aggregation functions that maximize the social welfare are the middlemost.*

Proof Suppose $r_1 \geq \cdots \geq r_n$ and $f(r_1, \ldots, r_n) = r$. If r is smaller than the middlemost grade when n is odd, or smaller than the two middlemost grades when n is even, then increasing it by $\epsilon > 0$ decreases a majority of the terms $d(r, r_j)$ by $\delta = d(r + \epsilon, r) > 0$ and increases a minority of those terms by δ, so $\Delta(r)$ is not a minimum. If r is greater than the middlemost grade when n is odd, or greater than the two middlemost grades when n is even, a similar majority/minority comparison shows that $\Delta(r)$ is not a minimum. Therefore r must be the middlemost grade when n is odd, or belong to the middlemost interval when n is even. In the latter case, every r in the middlemost interval minimizes $\Delta(r)$ because there are the same number of r_j greater than r as there are r_j smaller than r, so $\Delta(r)$ has the same value for all r in the middlemost interval. ∎

If, instead, total disutility is defined by $\sum_{j \in \mathcal{J}} (r - r_j)^2$, then the aggregation function that emerges is the arithmetic mean, $\sum_{j \in \mathcal{J}} r_j / n$, which has no justification. Doing this is tantamount to using Borda's method.

An ideal language for grading would be, in the jargon of measurement theorists, an interval measurement, where equal intervals have the same significance. Many physical measures have that property. The grades given to competing skaters or pianists could have it but almost certainly do not. Typically, as grades

approach "perfection," each additional point represents much more than an additional point added to a merely middling-to-good grade; and at the other end of the scale, the same phenomenon exists, taking off still another point from a bad grade is increasingly difficult. How to construct an interval measure to meet a percentage distribution of grades was discussed in chapter 8: for some distributions they exist, for some they do not. Of course, in many instances grades are *in fact* routinely added and averaged and thus treated as if they were interval measures. However, if for some application (e.g., wine) the utilities of judges were single-peaked in the grades and could be measured with some distance function d for all judges, then an interval measure could be constructed. For voting, the utilities are considerably more complex and can vary wildly from voter to voter.

12.5 Crankiness

Many practical mechanisms of grading eliminate extremes: music competitions often eliminate the top and bottom grades; in figure skating, three of the twelve grades are eliminated randomly and the top and bottom grades of the nine remaining judges are eliminated; in diving, there are either five or seven judges and either the top and bottom or the top two and bottom two are eliminated to end up with exactly three grades. There are three rationales for doing this. First, there is an intuitive sense that extremes may well represent cheaters whose grades should be eliminated. Second, as Galton expressed so well, cranks' grades—outliers by definition—should have limited impact. Extreme grades are not necessarily cranks' grades; they may also be given by judges with idiosyncratic tastes. Both must count—otherwise, why seek their opinions?—yet both should have a limited impact; *consensus* means "an opinion or position reached by a group as a whole." Third, the true significance of grades is ordinal: it is the place on a scale of measurement that counts, not the magnitude in and of itself.

The idea may be defined precisely.

An SGF *counters crankiness* if for $r_1 \geq \cdots \geq r_n$, $n \geq 3$, its aggregation function f satisfies

$$f(r_1, r_2, \ldots, r_{n-1}, r_n) = f(r_2, \ldots, r_{n-1}),$$

where in going from left to right the highest and lowest grades have been dropped.

Strictly speaking, the aggregation functions f in this definition should be distinguished, but that would encumber the notation: f depends on the number of

judges and their grades, so the two fs are different because their arguments are different numbers of grades. Iterating,

$$f(r_1, r_2, \ldots, r_{n-1}, r_n) = f(r_+, r_-),$$

where

$$r_+ = r_- = r_{(n+1)/2} \quad \text{when } n \text{ is odd,}$$

and

$$r_+ = r_{n/2}, \qquad r_- = r_{(n+2)/2} \quad \text{when } n \text{ is even,}$$

so the final grade must belong to the middlemost interval.

Theorem 12.4 *The unique aggregation functions that counter crankiness are the middlemost that depend only on the middlemost interval.*

12.6 Majority-Grade

The middlemost aggregation functions are repeatedly the only ones that meet the evident desiderata. But in the case of an even number of judges, they are infinite in number and thus can yield different results. What should be the unique choice?

An SGF *respects consensus* when all of A's grades belong to the middlemost interval of B's grades, and this 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, to take Galton's perspective, crankiness must not be respected. When there are only two judges, this is closely related to Hammond's equity principle (Hammond 1976).

The *majority-grade* f^{maj} is the lower middlemost order function,

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

Theorem 12.5a *An SGF f respects consensus if and only if $f \leq f^{maj}$. The unique middlemost SGF that respects consensus is the majority-grade f^{maj}.*

Proof Suppose f respects consensus, and consider any profile $r_1 \geq \cdots \geq r_n$. 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 equal at least to the one given to B, so f respects consensus.

The only middlemost SGF f for which $f \leq f^{maj}$ is the majority-grade f^{maj}, and it is clear that the majority-grade respects consensus. This ends the proof. The theorem and its proof are valid if the language is finite. ■

There is an obvious, opposed alternative.

An SGF *respects dissent* when all of A's grades belong to the middlemost interval of B's grades, and this implies that A's final grade is not above B's final grade.

Though this is not a compelling idea (particularly when the jury is small), respecting dissent leads to implications that parallel those obtained with respecting consensus. The proofs are similar.

The *other-majority-grade* $f^{o/maj}$ is the upper middlemost order function,

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

Theorem 12.5b *An SGF f respects dissent if and only if $f \geq f^{o/maj}$ The unique middlemost SGF that respects dissent is the other-majority-grade $f^{o/maj}$.*

12.7 Implications

Together the results point to the majority-grade as the best of all aggregation functions.

First, the middlemost aggregation functions emerge as the best possible choices to determine the final grade for many theoretical, practical, and intuitive reasons. They always agree with a majority decision. They best avoid manipulation. They meet the utilitarian test of maximizing the social welfare of the jury. They counter cheaters and cranks.

Second, when the choice is narrowed to the lower and upper middlemost aggregation functions, respectively f^{maj} and $f^{o/maj}$, all the good properties

of the order functions are inherited, properties proven for different reasons and with independent arguments.

Third, several properties single out the majority-grade f^{maj} among the middlemost functions. The first is respecting consensus, a purely symmetric condition that gives more credence to agreement than to disagreement. The second is a property of the majority-grade that is perhaps not at once evident: it is a Rawlsian criterion. Within the class of all the middlemost functions it is the one that singles out the best grade among many competitors or alternatives by the max-min criterion; it assigns to each competitor or alternative a final grade that is the minimum in the middlemost interval, the best among them being the one which is the highest. The third is its simplicity: the method is intuitive and easily understood. The fourth is that f^{maj} emphasizes the positive: an absolute majority is for the grade it assigns or better. Contrast this with $f^{o/maj}$: an absolute majority is for the grade it assigns or worse. But when there are a large number of voters, the majority-grade and the other-majority-grade are for all intents and purposes the same, $f^{maj} = f^{o/maj}$ (except for a set of measure zero).

To be able to determine a final grade when the number of judges is even is of paramount importance for several reasons, though a priori a jury constituted of an odd number is preferable. To begin, it is not always possible to determine the size of the jury; for instance, in an election the parity of the number of voters is unknown. Next, in practice it often happens that one or several judges may be absent or may be declared ineligible for one or another reason. Last, as is shown in chapter 13, to be able to establish complete rank-orderings of candidates or alternatives, it is necessary to be able to determine final grades for even-numbered as well as odd-numbered juries.

One major general problem remains: How are ties to be resolved? Equivalently, how are alternatives or competitors to be ranked?

13 Majority-Ranking

We may think of the political process as a machine which makes social decisions when the views of representatives and their constituents are fed into it. A citizen will regard some ways of designing this machine as more just than others. So a complete conception of justice is not only able to assess laws and policies but it can also rank procedures for selecting which political opinion is to be enacted into law.
—John Rawls

Some applications do not seek complete rank-orderings of the competitors: wine competitions come to mind, the aim being to give gold, silver, and bronze medals to certain percentages of the entries. In other applications, notably sports and elections, an ordered list from first to last and a clear winner are necessary.

A candidate (or alternative) 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 if rank-orderings are the outputs, the strategic behavior of judges and voters may well change. Does this imply that a different aggregation function should be used? If two candidates or alternatives receive the same majority-grade, how are they to be compared? How are competitors who receive different numbers of grades— their juries had different numbers of judges—to be compared? And, given it is known how to compare any two competitors, how is this justified on the basis of their grades? These questions are addressed in this chapter.

The answer is always to use the majority-ranking (first defined in chapter 1). It is justified axiomatically in this chapter. When there are many voters (e.g., national elections), the majority-gauge is sufficient to determine the majority-ranking (see chapter 14). In contrast with the traditional model, in which a fundamental incompatibility exists between choosing and ranking, in the new model grading and ranking are two sides of the same coin.

13.1 Strategy-Proofness in Ranking

Strategy raises its head again, but in a slightly altered guise. In a sports competition or an election, for example, how the final grades compare may take on greater importance than the final grades themselves. On the other hand, when a language is well established and accepted, the grades themselves inevitably take on a meaning and significance of their own. For example, a champion diver whose majority-grade was merely 7 or *good* would undoubtedly be considered a rather modest champion. Usually, both the absolute final grades—the majority-grades—and their order—the majority-ranking—are significant.

Does this change in point of view—the fact that the order-of-finish may be the principal focus—alter the conclusions arrived at earlier?

Given judges $j \in \mathcal{J}$ and candidates or alternatives $I \in \mathcal{C}$, a profile of grades is a matrix (r_j^I) with $r_j^I \in [0, R]$, and the vector of final grades is r^I. Suppose that the final grades of some two alternatives $A, B \in \mathcal{C}$ are $r^A < r^B$, but some judge 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.

A social grading function (SGF) is *strategy-proof-in-ranking* when the final grade of A is lower than that of B, $r^A < r^B$, and when any judge j of the opposite conviction, $r_j^A > r_j^B$, can neither decrease B's final grade nor increase A's final grade. (The specification "in-ranking" is often dropped when there can be no confusion as to the meaning of "strategy-proof.")

Consider an individual judge or 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. Suppose that when judge j's utility for A to rank above B is greater than her utility for B to rank no lower than A, $u_j\{r_A > r_B\} > u_j\{r_B \geq r_A\}$, this is equivalent to a conviction that A merits a higher grade than B, $r_j^A > r_j^B$. Then if the SGF is strategy-proof-in-ranking, it is a dominant strategy for judge j to assign grades according to her convictions since it serves no purpose to do otherwise. But notice that the ability to so distinguish any two competitors A and B means that the language of grades must be sufficiently rich.

Regrettably, this ideal cannot be met.

Theorem 13.1 *There exists no SGF that is strategy-proof-in-ranking.*

Whereas Arrow's impossibility and the incompatibility theorems of the traditional model are overcome in this theory, the counterpart of the Gibbard-Satterthwaite impossibility is not. However, whereas strong monotonicity fails in the traditional model—no method guarantees that when a candidate other

than the winner is lowered in some input, the winner remains the same—all reasonable methods of the new model are strongly monotonic.

Proof It suffices to construct a profile of grades of two competitors for which $r^A < r^B$ and there exists some judge j with $r_j^A > r_j^B$ who can either raise A's final grade or decrease B's.

Suppose $r_1^A > r_2^A > \cdots > r_n^A$. Either there is a judge j with $r_j^A > r^A$ or not.

If $r_j^A > r^A$, take B's grades to be

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

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

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

with a strict inequality holding between every pair of corresponding components. So monotonicity implies that 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.

On the other hand, it may be that there is no judge with $r_j^A > r^A$, that is, $r_j^A \leq r^A$ for every j. This means $r^A = r_1^A > r_2^A > \cdots > r_n^A$, the equality holding because the final grade cannot exceed the maximum grade. Now suppose that judge n increases A's grade to any value $\hat{r}_n^A$ in the interval $(r_1^A, r^B]$. Then

$$(\hat{r}_n^A, r_1^A, \ldots, r_{n-1}^A) > (r_1^A, r_2^A, \ldots, r_n^A),$$

with a strict inequality holding between every pair of corresponding components. So by monotonicity the final grade for the grades on the left must be strictly higher than that for the grades on the right: judge n is able to increase A's final grade. Notice that continuity plays no role in the proof; it is only necessary that the common language be sufficiently rich. ∎

Absolute perfection is unattainable, as so often in life. But perhaps something that limits the ability of judges to manipulate rankings can be achieved.

An SGF is *partially strategy-proof-in-ranking* when $r^A < r^B$ and when any judge j is of the opposite opinion, $r_j^A > r_j^B$, 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.

Happily, this condition can be realized.

Theorem 13.2 *The unique SGFs that are partially strategy-proof-in-ranking are the order functions.*

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

The proof begins the same way as that of the last theorem. Suppose that $r_j^A > r^A$ for some judge j, and let B's grades all belong to the open interval (r^A, r_j^A). Then $r^A < r^B$ and $r_j^A > r_j^B$, and as before judge j can decrease B's final grade by decreasing his grade. Therefore, by partial strategy-proofness, he cannot increase A's final grade. But then any judge who gave A a higher grade than r_j^A cannot increase the final grade as well as judge j.

A completely symmetric argument shows that if $r_j^A < r^A$ for some judge j, then he cannot decrease the final grade, nor can any judge 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 10.1, 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,

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

and

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

where $\{(1), \ldots, (n)\}$ 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 strategic-proof-in-ranking, completing the proof. ∎

This shows that the order functions are very robust.

What becomes of these two theorems when the language of grades is not sufficiently rich? The extreme case is, of course, approval voting, when there are only two grades. If a judge j approves of A and not B, then he cannot raise

A in the ranking, nor lower B. This does not mean, however, that approval voting is nonmanipulable in ranking. For a judge k could well approve of two candidates A and B, yet—because of the poverty of the language—not be able to express his preference for A to finish above B. In this case he could be tempted to disapprove of B. Thus, whatever the language of grades, rich or poor, every method is manipulable in ranking.

Approval voting is partially strategy-proof-in-ranking. The judge k just mentioned can contribute to lowering B in the ranking, but he can do nothing to raise A. Symmetrically, a judge who disapproves of two candidates can contribute to raising one but can do nothing to lower the other.

Chapter 19 shows how the majority judgment effectively limits strategic manipulation in a real example (see table 19.3 and the accompanying discussion).

Of course, judges' utilities may not depend only on the final rankings: the final grades may enter into their utilities as well. Since it is not always a dominant strategy for a judge to give grades according to his convictions when rankings are in the offing, what may be expected to happen at equilibrium in this game becomes an interesting question (see chapter 20).

13.2 Majority-Value

Recall the basic definitions and results. A profile of grades Φ is a matrix of m rows, each corresponding to a candidate or alternative, and n columns, each corresponding to a judge:

$$\Phi = \begin{pmatrix} r_1^A & \dots & r_j^A & \dots & r_n^A \\ \vdots & & \vdots & & \vdots \\ r_1^I & \dots & r_j^I & \dots & r_n^I \\ \vdots & & \vdots & & \vdots \\ r_1^Z & \dots & r_j^Z & \dots & r_n^Z \end{pmatrix}.$$

A social ranking function satisfies all the basic axioms if and only if each row of Φ may be permuted independently without changing the result. This simply means that given a jury's grades, there are $(n!)^m$ other profiles that give the identical final grades (though some of these may be the same since some of the entries in a row of the matrix Φ may be equal). Each contains exactly the same information, no more, no less. Among them, the *ordered profile* Φ^* is particularly convenient: it satisfies (with the subscripts in each row renamed)

$$\begin{array}{ccccccc} r_1^A \geq & \cdots & \geq r_j^A \geq & \cdots & \geq r_n^A \\ \vdots & & \vdots & & \vdots \\ r_1^I \geq & \cdots & \geq r_j^I \geq & \cdots & \geq r_n^I \\ \vdots & & \vdots & & \vdots \\ r_1^Z \geq & \cdots & \geq r_j^Z \geq & \cdots & \geq r_n^Z \end{array} .$$

Although this profile almost certainly does not correspond to the original jury's profile—it may well be that no judge of the jury has assigned the grades of a column of Φ^*—it is perfectly valid to think of the profile as that of a virtual, equivalent jury. The majority-grades $f^{maj}(I)$ (note the simplified notation) of the contestants $I \in \mathcal{C}$ are the entries of the $(n+1)/2$-th column when n is odd and those of the $(n+2)/2$-th column when n is even.

When the majority-grades of two contestants A and B differ, the one with the higher majority-grade ranks ahead of the other. When the majority-grades of two contestants are equal, there is a (virtual) judge who gives the same majority-grade to both: no more useful information concerning these two contestants can be drawn from this judge. Equivalently, when they are equal, each contestant's majority-grade is given by some judge (they may be different judges or one and the same judge): those equal grades yield no further information as to how to compare A with B.

The *majority-ranking* ($\succ_{maj}$) between two contestants evaluated by a common jury is determined by repeated application of the majority-grade to useful information:

1. If $f^{maj}(A) > f^{maj}(B)$, then $A \succ_{maj} B$.
2. If $f^{maj}(A) = f^{maj}(B)$, one majority-grade is dropped from the grades of each of the contestants, and the procedure is repeated.

There is another way to see the pertinence of this definition. Two contestants are tied according to the majority-grade. How is one to be ranked ahead of the other? All of the theory answers, by looking at the values given by the order functions, in particular, the ones that are situated in the middle.

This definition leads to a simple procedure. It helps to look at an example to see exactly how it works in practice. Suppose two contestants A and B are graded by a jury of seven as follows:

$$\Phi = \begin{pmatrix} A: & 85 & 73 & 78 & 90 & 69 & 70 & 73 \\ B: & 77 & 70 & 95 & 81 & 73 & 73 & 66 \end{pmatrix}.$$

Then the ordered profile is

$$\Phi^* = \begin{pmatrix} A: & 90 & 85 & 78 & \mathbf{73} & 73 & 70 & 69 \\ B: & 95 & 81 & 77 & \mathbf{73} & 73 & 70 & 66 \end{pmatrix},$$

and it may be seen that both candidates have a final majority-grade of 73 (boldface). Dropping it from both candidates' profiles,

$$\Phi^{(2)} = \begin{pmatrix} A: & 90 & 85 & 78 & \mathbf{73} & 70 & 69 \\ B: & 95 & 81 & 77 & \mathbf{73} & 70 & 66 \end{pmatrix},$$

the second majority-grade is obtained. It is 73 for both candidates. It is dropped to obtain

$$\Phi^{(3)} = \begin{pmatrix} A: & 90 & 85 & \mathbf{78} & 70 & 69 \\ B: & 95 & 81 & \mathbf{77} & 70 & 66 \end{pmatrix}.$$

The third majority-grades differ; A with a 78 has a higher grade than B with a 77, so $A \succ_{maj} B$.

Theorem 13.3 *The majority-ranking always ranks one contestant ahead of another unless the two are assigned an identical set of grades by the judges.*

The proof is completely trivial. But remember, the usual practice is to compute the mean or average value of the grades of each contestant and to rank them accordingly. This theorem is assuredly *not* true in that case. As a practical matter, the theorem is very important: complete rankings are very definitely wanted in many applications. Witness, for example, the complicated (and unjustified) rules used to settle ties in skating.

Define, for every candidate or alternative, the *first majority-grade* to be the majority-grade of the entire jury; the *second majority-grade* to be the majority-grade of the grades that remain after the first majority-grade is dropped; the *third majority-grade* to be the majority-grade of the grades that remain after the first two majority-grades have been dropped; ... ; and the *nth majority-grade* to be the majority-grade of the grades that remain after the first $n-1$ majority-grades have been dropped.

A candidate's (or alternative's) *majority-value for a jury of n members* is a vector of n components that assigns, in order, his first, second, third, ... , nth majority-grades. $A \succ_{maj} B$ if and only if A's majority-value is lexicographically higher than B's.

In the previous example, A's majority-value is (73, 73, 78, 70, 85, 69, 90) and B's is (73, 73, 77, 70, 81, 66, 95). A's is higher because A's is higher in the first entry at which they differ. Both candidates have a final grade of 73, but one of the two is slightly better. Notice, however, that in this example it suffices

to know the first three majority-grades to determine that $A \succ_{maj} B$; the kth majority-grades for $k \geq 4$ are superfluous.

What, it is fair to ask, are the lowest grades that A could earn and the highest grades that B could earn from a seven-member jury to end up with the same first three majority-grades and thus with $A \succ_{maj} B$. The answer is

$$\hat{\Phi}^* = \begin{pmatrix} A: & 78 & 78 & 78 & 73 & 73 & 0 & 0 \\ B: & 100 & 100 & 77 & 73 & 73 & 73 & 73 \end{pmatrix}.$$

The majority-ranking systematically eliminates crankiness as much as possible. The important grades are the ones in the middle. In the example, B's 100s and A's 0s are clearly suspect. Indeed, how is one to evaluate the competence of a judge who is a member of a jury? Since the members of the jury are—or must be considered—experts, those whose grades are in the middle *should* be the most significant. With the majority-ranking this wish is realized: the further a judge's grades are from the middle, the less will be their impact on the majority-value.

The majority-ranking is different than the ranking procedure that was once used by the International Skating Union (see tables 7.1b, 7.1c). The ISU's ordinal system gave the majority-grade but resolved ties differently, using increasingly ad hoc devices: first, if the majority-grade is the same, by the size of the majority in favor of at least the majority-grade; second, if the majorities are of the same size, by the magnitude of the sum of the corresponding places (considered as points or grades); third, if those magnitudes are the same, by the magnitude of the sum of all the places (considered as points or grades). The majority-ranking of the particular example happened to agree with that of the ISU. However, the majority-ranking is, in general, more precise than the ISU's. There are cases when the ISU's declares a tie but the majority-ranking discerns an order, whereas a tie in the majority-ranking (rare) implies a tie in the ISU's ranking.

On the other hand, the majority-ranking for the SCW, election Society (table 13.1) does not agree with the ISU's tie-breaking rules (compare with table 5.1a).

13.3 Characterization

Given the input grades of two competitors A and B, how should they be ranked? Write $A \succ_S B$ to mean A is ranked ahead of B, and $A \succeq_S B$ to mean either A is ahead of B or they are tied. Recall that a social ranking function (SRF) assigns a transitive rank-order $\succeq_S$ to any set of competitors that respects ties and grades (in comparing the lists of the grades of two candidates, who gave what grade is forgotten). This is a consequence of asking that it avoid the Arrow and Condorcet paradoxes (see chapter 9).

Table 13.1
Majority-Values and Majority-Ranking with Borda-Majority Method, SCW Society Election

	Points							
	2	*1*	*0*	Score	No. for	Majority-Value		ISU Rule
A	24	11	17	*1*	35	1st	*1,11112121...*	2d
B	9	21	22	*1*	30	3d	*1,11111110...*	3d
C	19	20	13	*1*	39	2d	*1,11111111...*	1st

Note: The majority-grade is given before the comma of the majority-values.

Consider an ordered set of input grades $r_1 \geq \cdots \geq r_n$. The *first middlemost interval* is the middlemost interval previously defined. The *second middlemost interval* is the middlemost interval when the grades of the first middlemost interval have been dropped. The kth *middlemost interval* is the middlemost interval when the grades of the jth middlemost intervals for $j < k$ have been dropped. For example, take the set of grades {11, 10, 10, 9, 6, 5, 4}: then the first middlemost interval is [9, 9], the second is [10, 6], the third is [10, 5], and the fourth is [11, 4].

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 (SRF) is a *middlemost SRF* if $A \succ_S B$ depends only on the set of grades that belong to the first of the kth middlemost intervals where they differ.

For example, if A's grades are those of the example just given and B's are {13, 12, 10, 9, 6, 3, 2}, then the first interval where they differ is the third interval: A's is [10, 5] and B's is [12, 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 middlemost intervals where A's and B's grades differ is the kth interval. An SRF *rewards consensus* when all of A's grades strictly belong to the kth middlemost interval of B's grades, and this implies that A is ranked above B, $A \succ_S B$.

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

An SRF is *choice-monotonic* if $A \succeq_S B$ and one judge raises the grade he gives to A, then $A \succ_S B$.

This is a natural idea that helps to resolve potential ties.

Theorem 13.4 *The majority-ranking is the unique middlemost, choice-monotonic 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 first middlemost interval and n is odd, the statement is true. If n is even, suppose the first middlemost intervals of A and B are $[r_-^A, r_+^A] \neq [r_-^B, r_+^B]$. The properties imply

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

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 first 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 first 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 kth middlemost intervals of their grades for $k > 1$, the proof is the same. ∎

The majority-ranking satisfies Arrow's *unanimity principle*: if every judge gives a strictly higher grade to a candidate A than to B, then A is ranked higher than B. The same is true for any social ranking function that is choice-monotonic.

This simple characterization of the majority-ranking parallels the characterization of the majority-grade and in this sense shows that it is its natural generalization. In marked contrast with the traditional model, grades and ranks agree. Moreover, the majority judgment is rank-compatible *and* rank-monotonic (impossible in the traditional model; recall theorem 4.4). There is a fundamental *compatibility* between grading, ranking, and electing.

Why are the middlemost SRFs so important? Because they best resist manipulation, as the following argument shows.

A social ranking function must be meaningful, that is, it must be order-consistent. Therefore, by theorem 11.5b (and sequel), an order function must be used to decide on a final grade; if there are ties, a second order function must resolve them; if ties remain, a third order function must be invoked; and so on. If the social ranking function must be choice-monotonic or if the practical situation

demands a resolution of ties as much as possible, then all order functions will be invoked.

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

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

There are $n!$ lexi-order SRFs. Which ones minimize the probability of cheating?

Recall that the probability of cheating with a grading function f is

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

Since the decision is taken lexicographically, the probability of cheating must be measured lexicographically as well. Thus the *lexi-probability of cheating* is

$$LCh(f^\sigma) = \left(Ch(f^{\sigma(1)}), \ldots, Ch(f^{\sigma(1)}, \ldots, f^{\sigma(k)}), \ldots, Ch(f^{\sigma(1)}, \ldots, f^{\sigma(n)})\right),$$

and the aim is to find the lexicographic minimum of $LCh(f^\sigma)$ over the functions f^σ.

Theorem 13.5 *The only lexi-order SRFs that minimize the lexi-probability of cheating are those that are middlemost SRFs. In particular, the majority-ranking is the unique choice-monotonic, meaningful SRF that minimizes cheating and rewards consensus.*

Proof To obtain a lexicographic minimum implies first minimizing the first term, then—given the first—minimizing the two first terms, then—given the two first—minimizing the three first terms, and so on.

First, $Ch(f^{\sigma(1)}, \ldots, f^{\sigma(k)})$ must be calculated. Suppose $\mathbf{r} = (r_1, \ldots, r_n)$, where $R > r_1 > \cdots > r_n > 0$. The number of judges who can increase the k-dimensional vector $\left(f^{\sigma(1)}(\mathbf{r}), \ldots, f^{\sigma(k)}(\mathbf{r})\right)$ is $\left(n + 1 - \min_{1\le i\le k} \sigma(i)\right)$, and the number of judges who can decrease it is $\left(\max_{1\le i\le k} \sigma(i)\right)$. So the probability of cheating is $\max_{1\le i\le k} Ch(f^{\sigma(i)})$.

Therefore to minimize the lexi-probability of cheating is equivalent to minimizing lexicographically $\left(Ch(f^{\sigma(1)}), \ldots, Ch(f^{\sigma(k)}), \ldots, Ch(f^{\sigma(n)})\right)$. As in the proof of theorem 12.2, the least manipulable order functions are those in the first middlemost interval, the next least manipulable are in the second middlemost interval, and so forth. So the lexi-order SRFs must be middlemost SRFs. ∎

There are about $2^{\binom{n}{2}}$ middlemost lexi-order SRFs. The majority-ranking is the one that rewards consensus. When there are many judges or voters, all of them

define almost surely the same ranking. In this sense, the majority-gauge-ranking is the unique meaningful SRF that minimizes cheating in a large electorate.

13.4 Juries of Different Sizes

In practice, as has already been mentioned, competitions with many contestants field many juries. *Les Citadelles du Vin*, an annual international wine and spirits competition sponsored by the OIV and held in the Bordeaux region, graded 1,247 different wines in June 2006, with a dozen juries chosen from some sixty experts, each meant to consist of five judges; in fact, several juries had fewer than five members. How are two alternatives evaluated by juries of different sizes to be compared? If the majority-grade of one is higher than that of the other, the answer is known. But what if they have the same majority-grade?

Postulate one jury of five persons evaluating three alternatives, and another jury of four persons evaluating two alternatives, with grades ranging from a high of 5 to a low of 1.

$$\Phi_1^* = \begin{pmatrix} A: & 4 & 4 & \mathbf{4} & 3 & 3 \\ B: & 5 & 4 & \mathbf{4} & 3 & 3 \\ C: & 5 & 4 & \mathbf{4} & 4 & 3 \end{pmatrix} \quad \text{and} \quad \Phi_2^* = \begin{pmatrix} D: & 4 & 4 & \mathbf{4} & 3 \\ E: & 5 & 4 & \mathbf{4} & 2 \end{pmatrix}.$$

They all have a majority-grade of 4. Clearly $C \succ_{maj} B \succ_{maj} A$, since the order agrees with weak monotonicity and the sets of grades are different, so there can be no ties. That $D \succ_{maj} E$ is also clear because the grades of D are more consensual than those of E. The only question is how to obtain a ranking of all five alternatives.

Two procedures naturally suggest themselves.

Annex the majority-grade of each alternative considered by the smaller jury to its grades as many times as necessary to equal the number of grades of the larger jury.

The rationale for this is that the most reliable collective information concerning the grade of a contestant given by any jury is its majority-grade, so it is adjoined. For the example, this yields

$$\Phi^* = \begin{pmatrix} A: & 4 & 4 & \mathbf{4} & 3 & 3 \\ B: & 5 & 4 & \mathbf{4} & 3 & 3 \\ C: & 5 & 4 & \mathbf{4} & 4 & 3 \\ D: & 4 & 4 & \mathbf{4} & 4 & 3 \\ E: & 5 & 4 & \mathbf{4} & 4 & 2 \end{pmatrix},$$

and calculating the majority-values gives

$$C \succ_{maj} D \succ_{maj} E \succ_{maj} B \succ_{maj} A.$$

The new implications are $C \succ_{maj} D$, which agrees with monotonicity, and $E \succ_{maj} B$, which recognizes a greater consensus around E's grades than around B's.

The second procedure takes the dual point of view.

Remove the majority-grade of each alternative evaluated by the larger jury from its grades. If either their new or second majority-grades distinguish them from the alternatives considered by the smaller jury, or their numbers of grades are the same as those of the smaller jury, they can be ranked. Otherwise, repeat.

For the example this gives

$$\hat{\Phi}^* = \begin{pmatrix} A: & 4 & 4 & \mathbf{3} & 3 \\ B: & 5 & 4 & \mathbf{3} & 3 \\ C: & 5 & 4 & \mathbf{4} & 3 \\ D: & 4 & 4 & \mathbf{4} & 3 \\ E: & 5 & 4 & \mathbf{4} & 2 \end{pmatrix}.$$

Observe that $\hat{\Phi}^*$ is exactly equal to Φ^* without the column of the majority-grades, so the two procedures result in identical rankings. They must. A moment's reflection shows why: annexing the identical majority-grade to the alternatives of the smaller jury gives all five the same majority-grade in Φ^*, so they are ranked according to the other grades; but these other grades are precisely the ones obtained when removing the majority-grade from the larger jury found in $\hat{\Phi}^*$. This is clearly true for all problems. This suggests generalizing the majority-ranking to the comparison of pairs of alternatives or candidates evaluated by juries of different sizes.

The *general majority-ranking* ($\succ_{gmaj}$) between two alternatives A and B when B's jury is no larger than A's jury is defined by the majority-ranking ($\succ_{maj}$) applied to two sets of grades of equal size: A's grades, and B's grades supplemented, to the extent necessary, by its majority-grade.

Theorem 13.6 *The general majority-ranking $\succ_{gmaj}$ is a complete, transitive order of all alternatives or candidates.*

Proof Given two alternatives A and B, suppose that $A \succ_{gmaj} B$ and that n_A is the size of A's jury and n_B the size of B's, with $n_B \leq n_A$. If $n_B < n_A$, then adjoin to B its majority-grade $n_A - n_B$ times. This changes B's original majority-value to one that has its majority-grade in the first $n_A - n_B$ components followed by its

previous majority-value. Therefore, from what is already known, $A \succ_{gmaj} B$ if and only if A's majority-value is lexicographically higher than B's (over vectors of length n_A). If A's majority-grade is now adjoined to its grades any number of times, and at the same time B's majority-grade is added the same number of times to its grades, then for these new grades $A \succ_{gmaj} B$. Adjoining the respective majority-grades at the beginning of the lexicographic order does not alter the order between the alternatives. This implies transitivity. ■

There is still another intuitive justification for this solution to the problem of comparing contestants evaluated by juries of different sizes and for the definition of the majority-grade when a jury's size is even. When adjoining a grade to a smaller jury, *any* grade could be adjoined. Which should it be? A jury of even size requires a larger relative majority for a relatively lower majority-grade than a jury of odd size. It seems that an even-sized jury puts contestants at a disadvantage in comparison with contestants evaluated by an odd-sized jury. It is thus reasonable to invoke the following principle:

A procedure that adjoins a grade to an alternative *compensates fairly* if it is the highest possible grade that leaves the majority-grade unchanged when the jury is of even size, and the lowest possible grade that leaves the majority-grade unchanged when the jury is of odd size.

Theorem 13.7 *A continuous procedure for adjoining a grade to an alternative compensates fairly if and only if the majority-grade is adjoined.*

Proof Let $r_1 > \cdots > r_n$. Suppose that n is even, so the majority-grade is $r_{(n+2)/2}$. Then adjoining any grade $r > r_{(n+2)/2}$ increases the majority-grade, whereas adjoining any grade $r \leq r_{(n+2)/2}$ does not change it. Therefore, the majority-grade itself must be adjoined. Suppose now that n is odd, so the majority-grade is $r_{(n+1)/2}$. Then adjoining any grade $r < r_{(n+1)/2}$ decreases the majority-grade, whereas adjoining any grade $r \geq r_{(n+1)/2}$ does not change it. Therefore, the majority-grade itself must be adjoined. Continuity extends the result to any profile of grades. ■

The compensation given in the case of an even jury is real. Compare A, whose grade is determined by a jury of $2n$ members, with B, whose jury has $2n + 1$ members, when their respective majority-grades are equal, so an extra grade is added to the even-sized jury. A is ranked ahead of B unless B's second majority-grade, which cannot be higher than the majority-grade, equals the majority-grade. This gives a small, real advantage to A that compensates for the earlier disadvantage. Now suppose that B's jury has $2n + 1$ members and A's $2n + 2$, so a grade is added to the odd-sized jury. A is ranked ahead of B

unless A's second majority-grade, which cannot be lower than the majority-grade, equals the common majority-grade. Again, a slight but real advantage is given to A.

This procedure seems acceptable as a practical matter in a variety of circumstances. But when there are large numbers of judges, say, one hundred or more, the majority-ranking may be simplified, and this suggests other possible practical approaches to the comparison of results of juries of different sizes.

14 Large Electorates

To the mind, good and evil, above and below, are not skeptical, relative concepts, but terms of a function, values that depend on the context they find themselves in.
—Robert Musil

The majority-values of competitors or candidates—and thus the majority-ranking—may be simplified when a jury has many judges or an electorate many voters, or when the common language contains few grades in comparison with the number of judges. To see why, look again at the example of candidate A in the fifty-two-voter SCW Society election (see table 13.1): twenty-four *2*s, eleven *1*s, and seventeen *0*s:

$$(\overbrace{2_{52}, \ldots, 2_{12}, 2_{10}, 2_8, 2_6}^{24}, \overbrace{1_4, 1_2, 1_1, 1_3, 1_5, 1_7, 1_9, 1_{11}, \ldots, 1_{17}}^{11}, \overbrace{0_{19}, \ldots, 0_{51}}^{17}).$$

The kth majority-grades for $k = 1, \ldots, 52$, are indicated by the subscripts. The majority-value is obtained from the ordered set of grades by beginning with the majority-grade, or the lower middlemost grade, and then taking *alternating grades emanating from the middle*. *A*'s majority-value (of fifty-two digits) begins with two *11*s and seven *12*s, and ends with a sequence of *02*s:

$$\overbrace{1,11112121212121202\ldots0202}^{52 \text{ digits}}.$$

But this may be written much more simply as

$$\overbrace{11}^{2}\ \overbrace{12}^{7}\ \overbrace{02}^{17},$$

where the numbers above the pairs describe how often each pair is repeated.

A general algorithm is given that shows how these repetitions may be used to write a more efficient, abbreviated expression of the majority-value. It applies

to grades reported as numbers or as percentages. However, when there are many judges or voters and relatively few grades, it is almost surely possible to determine an unambiguous order among the competitors with much less information than their majority-values, namely, their majority-gauges (as the electoral experiments testify).

The reader is forewarned: the ideas and procedures described here are very simple—everything relies on the sequence of alternating grades that emanate outward from the middle—but their description may seem complex because it is intricate.

14.1 Majority-Gauge

Recall that the *majority-gauge* of a competitor is a triple (p, α^*, q), where

- p is the number or percentage of the competitor's grades above the majority-grade;
- q is the number or percentage of the competitor's grades below the majority-grade;
- $\alpha* = \begin{cases} \alpha+ & \text{if } p > q, \\ \alpha- & \text{if } p \leq q, \end{cases}$ where α is the competitor's majority-grade.

When the distinction is necessary, $\alpha*$ is called the competitor's *modified majority grade*. Suppose, for example, that $\Lambda = \{A, B, C, D, E, F\}$ is the language of grades, with A the highest to F the lowest, and a competitor's grades are distributed as follows:

$$w = (\overset{A}{13.6\%}, \overset{B}{30.7\%}, \overset{C}{25.1\%}, \overset{D}{14.8\%}, \overset{E}{8.4\%}, \overset{F}{7.4\%}).$$

Then the competitor's majority-gauge is $(44.3\%, C+, 30.6\%)$.

Notice that if voters change their grades by increasing or decreasing them within the grades above the majority-grade α, or increasing or decreasing them within the grades below α, the majority-gauge does not change. In the example, all those who gave the candidate a B could change it to A, and all those who gave him a D could change it to E, yet the majority-gauge would remain the same. Clearly, a change in the triple (p, α, q) where either p increases or q decreases (or both) and the majority-grade α remains the same is better for the candidate.

Strictly speaking, taking the modified majority-grade $\alpha\pm$ instead of the majority-grade α is unnecessary because the information is implicit given p and q, but $\alpha+ \succ \alpha-$ is a useful mnemonic. It simplifies the definition

that follows when a finer order of modified majority-grades is defined, as follows:

$$\alpha* \succ \beta* \quad \text{if} \begin{cases} \alpha \succ \beta, & \text{or,} \\ \alpha = \beta & \text{and} \quad \alpha* = \alpha+, \beta* = \alpha-. \end{cases}$$

Consider two candidates X and Y with majority-gauges $(p, \alpha*, q)$ and $(r, \beta*, s)$. The *majority-gauge-ranking* $\succ_{mg}$ places X ahead of Y, $X \succ_{mg} Y$, or $(p, \alpha*, q)$ ahead of $(r, \beta*, s)$,

$$X \succ_{mg} Y, \quad \text{or} \quad (p, \alpha*, q) \succ_{mg} (r, \beta*, s)$$

when

$$\begin{array}{lll} \alpha* \succ \beta*, & \text{or} & \\ \alpha* = \beta* = \alpha+ & \text{and} & p > r, \quad \text{or} \\ \alpha* = \beta* = \alpha- & \text{and} & q < s. \end{array}$$

With many voters or judges it is almost sure that $p \neq r$ and $q \neq s$, so almost certainly the majority-gauge-ranking will yield an unambiguous order of finish of all the competitors.

Theorem 14.1 *When the majority-gauge-ranking places a competitor X ahead of another Y, so does the majority-ranking, that is,*

$$X \succ_{mg} Y \quad \textit{implies} \quad X \succ_{maj} Y.$$

It may be, however, that the majority-gauge-ranking cannot decide because either

$$\begin{array}{lll} \alpha* = \beta* = \alpha+ & \text{and} & p = r, \quad \text{or} \\ \alpha* = \beta* = \alpha- & \text{and} & q = s, \end{array}$$

yet X and Y are not tied in the majority-ranking. The experimental evidence shows how rarely the majority-gauges are tied. In four independent experiments, with 10,000 random drawings of 101 ballots, ties occurred 8, 9, 10, and 11 times, a rate of about 0.1% (see tables, 6.12a, 6.12b, 6.14a, 6.14b). In another four independent experiments, with 10,000 random drawings of 201 ballots, ties occurred 3, 3, 9, and 10 times (see tables, 14.4a and 14.4b). And as the number of voters increases, ties become still less likely.

Proof It may be assumed that n, the number of judges or voters, is odd. For, if n is even, adjoining one α to X's grades (his majority-grade) leaves $(p, \alpha*, q)$ unchanged; and similarly for Y. Moreover, the majority-order between X and Y remains unchanged as well (see chapter 13).

Take $X \succ_{mg} Y$, that is, $(p, \alpha*, q) \succ_{mg} (r, \beta*, s)$. Then either $\alpha* \succ \beta*$ or $\alpha* = \beta*$.

To begin, suppose $\alpha* \succ \beta*$. Then either X's majority-grade is above Y's, $\alpha \succ \beta$, or they have the same majority-grade α, but $\alpha* = \alpha+$ and $\beta* = \alpha-$. In the first case, the majority-grade of X is above Y's, so $X \succ_{mg} Y$ implies $X \succ_{maj} Y$.

In the second case, $p > q$ and $r \leq s$. The relation $p > q$ implies that beginning at the center and taking X's alternate emanating grades, a higher grade than α is encountered before a lower grade, whereas $r \leq s$ implies that taking Y's alternate emanating grades from the center, a lower grade than α is encountered before a higher grade. This is clearly true when $r < s$. It is also true when $r = s$ because since n is odd, the first majority-grade is squarely in the middle and the second majority-grade is on the side of the lower grades, so the alternating process begins on the lower side. There are two possibilities: (1) a higher grade than α is encountered in X's grades before a lower grade than α is encountered in Y's grades, or (2) a lower grade than α is encountered in Y's grades before a higher grade than α is encountered in X's grades. In either case X's majority-value is lexicographically above Y's majority-value, so $X \succ_{mg} Y$ implies $X \succ_{maj} Y$.

Now suppose $\alpha* = \beta*$. Either $\alpha* = \beta* = \alpha+$, or $\alpha* = \beta* = \alpha-$. Beginning at the center, take the alternate emanating grades of both X and Y. In the first case a grade above α is encountered before a grade below α for X and for Y; and $p > r$ implies it is encountered with X's grades before Y's, so X's majority-value is lexicographically above Y's. In the second case a grade below α is encountered before a grade above α for X and for Y; and $q < s$ implies it is encountered with Y's grades before X's, so again X's majority-value is lexicographically above Y's, completing the proof. ■

For an example of the majority-ranking and the corresponding majority-values and majority-gauges, see table 14.3.

Two important strategic properties of the majority-grade f^{maj} are inherited by the majority-gauge. In comparing majority-gauges here, "higher," "lower," "increase," and "decrease" mean with respect to the majority-gauge-ranking $\succ_{mg}$.

The majority-gauge is strategy-proof-in-grading:

Suppose that a competitor's majority-gauge is $(p, \alpha*, q)$. If a voter's input grade is β and $\beta \succ \alpha$, any change in his input can only lead to a lower majority-gauge; and if a voter's input grade is $\gamma \prec \alpha$, any change in his input can only lead to a higher majority-gauge.

The proof is obvious: the only change that can occur in $(p, \alpha*, q)$ is either for some voter among the p (who gave a grade higher than α) to give as input α or a lower grade, resulting in a lower majority-gauge; or for some voter among the q (who gave a grade lower than α) to give as input α or a higher grade, resulting in a higher majority-gauge.

The majority-gauge is also partially strategy-proof-in-ranking:

Suppose that X's majority-gauge is $(p, \alpha*, q)$, Y's is $(r, \beta*, s)$, and $X \succ_{mg} Y$. Consider a voter j who is of the opposite opinion: for her, X's grade α_j should be below Y's grade β_j, i.e., $\alpha_j \prec \beta_j$. Then if she can decrease X's majority-gauge, she cannot increase Y's majority-gauge, and if she can increase Y's majority-gauge, she cannot decrease X's majority-gauge.

It is again easy to verify this. If j can decrease X's majority-gauge, then she must have assigned $\alpha_j \succeq \alpha$, implying $\beta_j \succ \alpha_j \succeq \alpha \succeq \beta$, so j cannot increase Y's majority-gauge. And if j can increase Y's majority-gauge, then $\beta_j \preceq \beta$, implying $\alpha_j \prec \beta_j \preceq \beta \preceq \alpha$, so j cannot decrease X's majority-gauge. These properties carry over to other rankings that depend only on the triple (p, α, q), as will be seen anon. Moreover, when $p \neq q$ and both are strictly less than 50%, all middlemost lexi-order SRFs yield the majority-gauge-ranking.

14.2 Abbreviated Majority-Value

The majority-gauge can produce a tie between two candidates, though this is highly unlikely, as was observed. In that case, the "full" majority-value may be invoked. But when there are many voters or judges, or the language of grades is small, it is efficient to express the majority-value in an abbreviated form. The first several terms give the majority-gauge, though not obviously: exactly how is explained at the end of this section.

It is assumed that each judge assigns a grade to every competitor. Suppose the common language consists of the grades $\Lambda = \{\alpha_1 \succ \cdots \succ \alpha_r\}$, and that w_i is the percentage of α_i's received by a competitor (the numbers of grades instead of their percentages could, of course, be used). Then the competitor's majority-grade is α_t if

$$w_1 + \cdots + w_{t-1} \leq 50\% < w_1 + \cdots + w_{t-1} + w_t.$$

Each step of the algorithm for calculating the majority-value is accompanied by two lists. The first, in parentheses, consists of numbers (coming from the w_i's) together with a *center*, denoted by $||$:

$$(v_1, \ldots, v_s^- || v_s^+, \ldots, v_r).$$

The second, in square brackets, is an approximation of the majority-value: ordered pairs of grades $(\alpha_e\alpha_d)$, where $\alpha_e \preceq \alpha_d$, each pair accompanied by the frequency with which it appears, $w(e, d)$:

$$\left[\overset{w(e,d)}{(\alpha_e\alpha_d)} \ldots \overset{w(h,g)}{(\alpha_h\alpha_g)}\right].$$

Initial Lists of the Algorithm

$$(w_1, \ldots, w_t^- || w_t^+, \ldots, w_r) \quad \text{and} \quad [\emptyset],$$

where w_t^- and w_t^+ are chosen so that $w_1 + \cdots + w_t^- = w_t^+ + \cdots + w_r = 50\%$.

To see what is going on, take the following example. The common language is $\Lambda = \{A, B, C, D, E, F\}$, A highest, F lowest. The competitor's grades from A down to F are distributed as follows:

$$w = (\overset{A}{13.6\%}, \overset{B}{30.7\%}, \overset{C}{25.1\%}, \overset{D}{14.8\%}, \overset{E}{8.4\%}, \overset{F}{7.4\%}).$$

The initial lists are

$$(\overset{A}{13.6\%}, \overset{B}{30.7\%}, \overset{C}{5.7\%} || \overset{C}{19.4\%}, \overset{D}{14.8\%}, \overset{E}{8.4\%}, \overset{F}{7.4\%}) \quad \text{and} \quad [\emptyset].$$

Recursive Step of the Algorithm

Input

$$v = (v_1, \ldots, v_s || v_{s+1}, \ldots, v_r) \quad \text{and} \quad \left[\overset{w(e,d)}{(\alpha_e\alpha_d)} \ldots \overset{w(h,g)}{(\alpha_h\alpha_g)}\right].$$

Let

k correspond to the lowest grade left of the center with $v_k > 0$,

l correspond to the highest grade right of the center with $v_l > 0$,

$\delta = w(l, k) = \min\{v_l, v_k\}$.

Output

$$v' = (v'_1, \ldots, v'_s || v'_{s+1}, \ldots, v'_r) \quad \text{and} \quad \left[\overset{w(e,d)}{(\alpha_e\alpha_d)} \ldots \overset{w(h,g)}{(\alpha_h\alpha_g)}\overset{w(l,k)}{(\alpha_l\alpha_k)}\right],$$

where $v' = v$ except that $v'_k = v_k - \delta$ and $v'_l = v_l - \delta$.

Return to the example with the input the initial vectors. Then $\delta = w(C, C) = \min\{19.4, 5.7\} = 5.7$, so the new vectors are

Table 14.1
Calculation of the Abbreviated Majority-Value: An Example with Percentages of Grades

Vectors v								
A	B	C	$\|\|$	C	D	E	F	First Terms aM-Vs
(13.6	30.7	5.7	$\|\|$	19.4	14.8	8.4	7.4)	$[\emptyset]$
(13.6	30.7	0	$\|\|$	13.7	14.8	8.4	7.4)	$[\overset{5.7}{(CC)}]$
(13.6	17.0	0	$\|\|$	0	14.8	8.4	7.4)	$[\overset{5.7}{(CC)}\overset{13.7}{(CB)}]$
(13.6	2.2	0	$\|\|$	0	0	8.4	7.4)	$[\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}]$
(13.6	0	0	$\|\|$	0	0	6.2	7.4)	$[\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}\overset{2.2}{(EB)}]$
(7.4	0	0	$\|\|$	0	0	0	7.4)	$[\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}\overset{2.2}{(EB)}\overset{6.2}{(EA)}]$
(0	0	0	$\|\|$	0	0	0	0)	$[\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}\overset{2.2}{(EB)}\overset{6.2}{(EA)}\overset{7.4}{(FA)}]$

Note: aM-V = abbreviated majority-value.

$$(\overset{A}{13.6}\%, \overset{B}{30.7}\%, \overset{C}{0}\% \,||\, \overset{C}{13.7}\%, \overset{D}{14.8}\%, \overset{E}{8.4}\%, \overset{F}{7.4}\%) \quad \text{and} \quad \left[\overset{5.7}{(CC)}\right].$$

The algorithm *terminates* when the vector in parentheses is all zeros, $v = (0, \ldots, 0||0, \ldots, 0)$. Each succeeding step of the algorithm for this example is given in table 14.1.

The final vector in square brackets is the *abbreviated majority-value* (aM-V). It is

$$\left[\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}\overset{2.2}{(EB)}\overset{6.2}{(EA)}\overset{7.4}{(FA)}\right].$$

The first grade C is the majority-grade. The sum of the w's is exactly 50, exactly half of the total percentage or number of voters, since the grades are rearranged into couples. Its first terms are CC with frequency 5.7, followed by CB with frequency 13.7—so altogether there are $2 \times 5.7 + 13.7 = 25.1\%$ of C's—then DB with frequency 14.8, ... Each successive vector in square brackets gives a longer part of the first terms of the majority-value. So each is an approximation of the majority-value written in an abbreviated way.

When numbers of grades are given in terms of percentages or relative frequencies, there can be any number of voters, including an infinite number. When, instead, results are given in numbers of voters, a small modification is necessary. This is most easily understood via an example.

Take the first example of this section. The numbers of grades from *2* down to *1* then to *0* are $(\overset{2}{24}, \overset{1}{11}, \overset{0}{17})$. There is an even number of voters, so the algorithm

Table 14.2
Calculation of the Abbreviated Majority-Value: An Example with Fifty-two grades

Vectors v					
2	*1*	||	*1*	*0*	First Terms aM-Vs
(24	2	||	9	17)	$\left[\emptyset\right]$
(24	0	||	7	17)	$\left[\overset{2}{(11)}\right]$
(17	0	||	0	17)	$\left[\overset{2}{(11)}\overset{7}{(12)}\right]$
(0	0	||	0	0)	$\left[\overset{2}{(11)}\overset{7}{(12)}\overset{17}{(02)}\right]$

Note: aM-V = abbreviated majority-value.

proceeds as before (see table 14.2). The competitor's majority-grade is 1. The competitor's majority-value is *11* twice, followed by *12* seven times, then by *02* seventeen times, that is,

$$\overbrace{1111}^{4}\overbrace{1212\ldots12}^{14}\overbrace{0202\ldots02}^{34}.$$

Compare it with the abbreviated majority-value in table 14.2: it is the same.

When the number of voters is odd, the algorithm cannot even begin. But every competitor has an odd number of grades, so adjoin to each her majority-grade and proceed; doing so maintains the majority-order among them (as was seen in the discussion of juries of different sizes).

Table 14.3
Abbreviated majority-values and Majority-Gauges of six sets of grades, Listed from Best to Worst according to Majority-Ranking

	% Grades					
	A	*B*	*C*	*D*	*E*	*F*
1st	21.6	26.3	19.1	19.9	6.5	6.6
2d	13.6	30.7	25.1	14.8	8.4	7.4
3d	18.9	25.4	25.1	16.3	9.6	4.7
4th	16.9	22.5	19.1	16.6	12.2	12.7
5th	17.0	22.7	18.5	16.8	12.3	12.7
6th	5.1	14.9	38.2	12.1	13.5	16.2

The lexicographic order of the majority-values (M-Vs) or the abbreviated majority-values (aM-Vs) determines the majority-ranking: $v \succ_S \overline{v}$ if (going from left to right) at the first grade where they differ, v's is the higher grade. It is easy to check that the order of the abbreviated majority-values in table 14.3 goes from best to worst. (1) After CC's, a B is encountered in the first aM-V before the second; (2) after CC's, CB's, and DB's, an E is encountered in the third before the second; (3) after CC's, a B is encountered in the third before the fourth; (4) after CC's, a D is encountered in the fifth before the fourth; and (5) after CC's and DC's, a B is encountered in the fifth before the sixth. The majority-gauge rank orders all competitors with these grades except that it is unable to distinguish between the second and third.

Given the aM-V of a candidate, it is easy to obtain his majority-gauge $(p, \alpha*, q)$. The prescription is this:

- The first grade of the aM-V is the majority-grade α.
- $\alpha* = \begin{cases} \alpha+ & \text{if the first grade of aM-V} \neq \alpha \text{ is above } \alpha, \\ \alpha- & \text{if the first grade of aM-V} \neq \alpha \text{ is below } \alpha. \end{cases}$
- $\alpha* = \alpha+$ implies $\begin{cases} p = 50- \text{ frequency of } (\alpha\alpha), \\ q = 50- \text{ frequency of all pairs including } \alpha. \end{cases}$
- $\alpha* = \alpha-$ implies $\begin{cases} q = 50- \text{ frequency of } (\alpha\alpha), \\ p = 50- \text{ frequency of all pairs including } \alpha. \end{cases}$

Take, for example, the sixth set of grades shown in table 14.3. The majority-grade is C, the first grade. $C* = C-$ because the first grade encountered $\neq C$ is D, which is below C. The frequency of (CC) is 8.2, so $q = 41.8$. The sum

Table 14.3
(cont.)

	Abbreviated Majority-Value	Majority-Gauge $(p\%, \alpha^*, q\%)$
1st	$\overset{2.1}{(CC)}\overset{14.9}{(CB)}\overset{11.4}{(DB)}\overset{8.5}{(DA)}\overset{6.5}{(EA)}\overset{6.6}{(FA)}$	$(47.9, C^+, 33.0)$
2d	$\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{14.8}{(DB)}\overset{2.2}{(EB)}\overset{6.2}{(EA)}\overset{7.4}{(FA)}$	$(44.3, C^+, 30.6)$
3d	$\overset{5.7}{(CC)}\overset{13.7}{(CB)}\overset{11.7}{(DB)}\overset{4.6}{(EB)}\overset{9.6}{(EA)}\overset{4.7}{(FA)}$	$(44.3, C^+, 30.6)$
4th	$\overset{8.5}{(CC)}\overset{2.1}{(DC)}\overset{14.5}{(DB)}\overset{8.0}{(EB)}\overset{4.2}{(EA)}\overset{12.7}{(FA)}$	$(39.4, C^-, 41.5)$
5th	$\overset{8.2}{(CC)}\overset{2.1}{(DC)}\overset{14.7}{(DB)}\overset{8.0}{(EB)}\overset{4.3}{(EA)}\overset{12.7}{(FA)}$	$(39.7, C^-, 41.8)$
6th	$\overset{8.2}{(CC)}\overset{12.1}{(DC)}\overset{9.7}{(EC)}\overset{3.8}{(EB)}\overset{11.1}{(FB)}\overset{5.1}{(FA)}$	$(20.0, C^-, 41.8)$

of the frequencies of all pairs including C is 30.0, so $p = 20.0$. The foregoing prescription assumes that a candidate's grades are given in percentages. If instead they are given in numbers, it is assumed that the number of judges n is even, and the prescription is the same except that 50 is replaced by $n/2$.

14.3 Other Rules

Tie-Breaking Rules

The majority-gauge-ranking may be viewed as a rule to break a tie when two candidates X and Y have the same majority-grade α. The outcome as concerns the two leading candidates in Orsay's 12th precinct in the experiment of 2007 is an interesting example. The two leading candidates' majority judgment grades and resulting (p, α, q) were

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	(p, α, q)
Royal	16.4%	26.0%	17.5%	16.1%	10.1%	13.9%	(42.4%, *Good*, 40.1%)
Bayrou	10.8%	30.0%	27.7%	15.5%	8.0%	7.9%	(40.8%, *Good*, 31.4%)

The majority-gauge places Royal ahead of Bayrou (so the majority-value and majority-ranking do, too).

Given the triple (p, α, q), various different seemingly reasonable rules are possible for designating which of two candidates leads the other when their majority-grades α are the same. One naive idea is the *upper tie-breaking rule*:

When two candidates have the same majority-grade α, the candidate with the greater percentage of grades above α (namely, p) is first; and in case of a tie, (p, α, q) and (p, α, s), the first is ranked higher if $q < s$.

It ranks Royal ahead of Bayrou. But then, why not instead take the *lower tie-breaking rule*:

When two candidates have the same majority-grade α, the candidate with the greater percentage of grades below α (namely, q) is last; and in case of a tie, (p, α, q) and (r, α, q), the first is ranked higher if $p > r$.

It ranks Bayrou ahead of Royal.

Imagine the grades of the candidates ordered from best to worst on seesaws or teeterboards that are balanced exactly at the middlemost grade: that which weighs more heavily toward the better-than grades than toward the worse-than grades should correspond to the candidate who is ranked higher. This is the *difference tie-breaking rule*:

When two candidates have the same majority-grade α, the candidate with the greater difference between the percentages of grades above and below α (namely, $p - q$) is first; and in case of a tie, (p, α, q) and (r, α, q) with $p - q = r - s$, the first is ranked higher if $p + q < r + s$ because it expresses a greater consensus for α.

It places Bayrou comfortably ahead of Royal. David Gale, believing at first that this rule agrees with the majority-ranking, suggested it for large electorates. But it does not agree with the majority-ranking, as the Bayrou-Royal example shows. However, the difference tie-breaking rule agrees with the majority-gauge-ranking to the extent that both rank a candidate with an $\alpha+$ above a candidate with an $\alpha-$. It may be observed that the majority-gauge-ranking and the difference-rule-ranking are the same for the Orsay French presidential election experiment and for the INFORMS U.S. presidential election experiment (see table 1.5), and this will often be the case.

The majority-gauge-ranking, deduced from the underlying principles of the majority judgment, is more subtle. If the weight of one candidate leans toward the better-than side of the seesaw and the weight of the other leans toward the worse-than side, then the better-than side is, of course, ahead. If both candidates lean toward the better-than side, then the one whose better-than side is the weightier is ahead of the other—so Royal with 42.4% is ahead of Bayrou with 40.8%—because there are more voters who really care about Royal than there are who really care about Bayrou. Symmetrically, if both candidates lean toward the worse-than side, then the one whose worse-than side is the weightier is behind the other, because there are more voters who really are opposed.

The three tie-breaking rules—upper, lower, and difference—share important properties with the majority-gauge-ranking: all are strategy-proof-in-grading and partially strategy-proof-in-ranking. The proofs are the same as those given for the majority-gauge, where comparisons of majority-gauges are made according to each of the three rules. The essential reason is that all are based on the triple (p, α, q) of the majority-gauge. Each of the three have the serious defect of disagreeing with the majority-ranking; nevertheless, they sometimes behave better than other methods, and they may be compared among themselves. Statistical evidence suggests that the majority-gauge-ranking (for all intents and purposes, the majority judgment, when there are many voters) is slightly less manipulable than the majority judgment with the difference tie-breaking rule, and both are clearly less manipulable than the majority judgement with the upper or lower rules. This confirms the theoretical result because the

Table 14.4a
Number of Wins among Royal, Bayrou, and Sarkozy, 2007 Orsay Experiment

	Left ⟵ Royal		Bayrou		⟶ *Right* Sarkozy		Tie	
First-past-the-post	918	*622*	0	*0*	9,068	*9,311*	14	*67*
Two-past-the-post	1,124	*1,139*	92	*155*	8,244	*8,117*	540	*589*
MJ: upper rule	613	*662*	922	*871*	8,462	*8,464*	3	*3*
MJ: majority-gauge	603	*564*	4,252	*4,397*	5,142	*5,036*	3	*3*
MJ: difference rule	285	*255*	6,840	*6,906*	2,385	*2,326*	490	*513*
Condorcet	152	*138*	8,381	*8,416*	941	*915*	403	*415*
MJ: lower rule	380	*373*	9,540	*9,554*	49	*42*	31	*31*
Borda	13	*40*	9,973	*8,731*	0	*1,086*	14	*143*

Note: MJ = majority judgment.

For each candidate, the (nonitalic) left-hand column shows the number of wins among all twelve candidates, with the winner always Royal, Bayrou, or Sarkozy; and the (italic) right-hand column shows the number of wins among Royal, Bayrou, or Sarkozy only.

Ten thousand samples of 201 ballots, which were drawn from a sample of 501 ballots representative of the national vote. The same approach was used as in estimating first-round results on the basis of majority judgment ballots; the percentage of votes of each candidate in the sample of 501 ballots came close to that of the candidate's national vote. In this sample, Sarkozy had 30.7%, Royal had 25.5%, and Bayrou had 18.7%. (Le Pen had 9.3%.)

With Cordorcet's method, there were 123 cycles with twelve candidates, and 116 cycles with three candidates.

majority-gauge is the unique meaningful SRF that minimizes cheating in large electorates.

The majority judgment with the difference tie-breaking rule is seductive. It is simple and intuitive, and its behavior is closer to the majority-gauge-ranking's than any of the other methods. However, it is of little or no use when juries are relatively small, for it may well happen that exactly one judge gives the majority-grade, implying the rule contributes nothing to breaking ties. Moreover, it ends in ties much more frequently than does the majority-gauge (tables 14.4a and 14.4b). It seems to favor centrist candidates in comparison with the majority-gauge-ranking because it only contrasts grades better than the majority judgment with grades worse than the majority-grade, ignoring the absolute weights of each of the three categories, the percentages of majority-grades, of better-than majority-grades, and of worse-than majority-grades. These points are verified in the experimental evidence of tables 14.4a and 14.4b.

The database for these tables is again the 1,733 ballots of the 2007 Orsay French presidential election experiment. These tables repeat and extend to the three tie-breaking rules the evidence given in tables 6.12a, 6.12b, 6.14a, and 6.14b in which experiments generated 10,000 random samples of 101 voters

Table 14.4b
Number of Wins among Royal, Bayrou, and Sarkozy, 2007 Orsay Experiment

	Left ⟵ Royal		Bayrou		⟶ *Right* Sarkozy		Tie	
First-past-post	1,844	*4,275*	1,850	*1,478*	6,271	*3,825*	35	*397*
Two-past-post	3,425	*3,959*	5,527	*4,242*	685	*1,098*	663	*701*
MJ: upper rule	832	*872*	8,324	*8,250*	835	*868*	9	*10*
MJ: lower rule	1,454	*1,454*	8,369	*8,380*	144	*122*	33	*44*
MJ: majority-gauge	724	*765*	8,830	*8,785*	437	*440*	9	*10*
Condorcet	173	*176*	9,497	*9,507*	2	*3*	327	*309*
MJ: difference rule	119	*119*	9,662	*9,631*	31	*45*	188	*205*
Borda	69	*31*	9,930	*9,889*	0	*3*	1	*77*

Note: MJ = majority judgment.

For each candidate, the (nonitalic) left-hand column shows the number of wins among all twelve candidates, with the winner always Royal, Bayrou, or Sarkozy; and the (italic) right-hand column shows the number of wins among Royal, Bayrou, or Sarkozy only.

Ten thousand samples of 201 ballots, which were drawn from 1,733 ballots.

Cordorcet's method yielded 1 cycle with twelve candidates and 5 cycles with three candidates.

were generated. In the experiments whose results are given in tables 14.4a,b, 10,000 random samples of 201 voters were generated independently for each of the four experiments (two per table). It may be observed that the qualitative properties are the same (where comparisons are possible). The reader may check that the number of wins and ties is 10,000 for each of the methods of election; for instance, for the majority-gauge in table 14.4a, $603 + 4, 252 + 5, 142 + 3 = 10,000$. This fails to be true only for Condorcet's method because cycles occur—the Condorcet paradox—that is, in table 14.4a, $138 + 8{,}416 + 915 + 415 = 9{,}884$, and there were 116 cycles.

In chapter 6 it was shown that Bayrou is in fact the centrist candidate. The methods of election used to designate the winners are listed from least favorable for the centrist candidate to most favorable for the centrist candidate. The majority judgment with the lower tie-breaking rule tends to favor the centrist because he is infrequently given low grades; symmetrically, the majority judgment with the upper tie-breaking rule tends to penalize the centrist because he is infrequently given high grades. The majority-gauge-ranking tends to be in between, as is the difference tie-breaking rule, though the latter tends to be more favorable for the centrist than the former, as may be seen in each of the four experiments of tables 14.4a and 14.4b. Recall, however, that in the three precincts of the Orsay experiment, voters gave quite different percentages of their actual votes than the national percentages: much more to Bayrou, more to Royal, and slightly less to Sarkozy; thus in table 14.4b, Bayrou wins much more often with any method.

Again, Condorcet's method strongly favors the centrist candidate, and Borda's is overwhelmingly favorable to the centrist candidate, though less so when there are only three candidates; and first-past-the-post and two-past-the-post overwhelmingly penalize the centrist candidate.

We conclude that the majority-gauge-ranking should be used in elections with a great many voters, though the majority judgment with the difference tie-breaking rule may be a bit easier to understand and shares some of the same desirable properties. Regrettably, it does not agree with the majority-ranking.

Electorates of Different Sizes

What should be done when candidates or competitors receive different numbers of grades? This can happen with large electorates and juries as well as small ones. The answer depends very much on the particular situation. Three separate procedures naturally commend themselves.

The first procedure was developed in chapter 13. It is the choice when juries are small, but it may also be used when juries are large. Suppose the competitor with the most grades has n grades. This procedure adjoins to any competitor's set of grades her majority-grade as many times as necessary to give her a total of n grades. Why this is reasonable has already been explained; essentially it is because the majority-grade is the best evaluation of a jury's collective opinion.

The second procedure was used in the Orsay and other voting experiments. Every voter or judge *must* assign a grade to each candidate, so every candidate necessarily has the same number of grades. This was accomplished by stating very clearly that when a voter gives no grade or has "no opinion," it means the voter chooses *To Reject* the candidate. If "no opinion" were permitted and the first procedure was used, it would be possible for a relatively unknown candidate to receive "no opinion" from the overwhelming majority of the electorate and a very few very high grades from family, friends, and other acquaintances, giving him a high majority-grade, which, when repeatedly adjoined, would make him the winner or place him high in the ranking. That is clearly unacceptable. The rationale for insisting that a voter must evaluate every candidate is on the one hand civic—a voter *should* take the trouble to know who each candidate is and what he stands for—and on the other hand realistic—a voter who has "no opinion" about a candidate has implicitly rejected him, having not been sufficiently motivated to invest the time necessary to evaluate him.

The third procedure is an immediately evident option when there are several large juries evaluating different sets of competitors (for if exactly the same sets

of competitors are to be evaluated, the different juries may be coalesced into one very large jury). Imagine, for example, two juries, one with 150 judges, another with 200. The first procedure would replicate the majority-grades of candidates evaluated by the smaller jury fifty times. But with these numbers it is also perfectly reasonable to simply translate the numbers of grades into percentages and then compare them.

Weakly Monotonic Aggregation Functions

The theory developed in this book has assumed that social grading functions, which assign a final grade as a function of the judges' or voters' input grades, are monotonic: (1) if the input grades are replaced by the same or higher grades, then the final grade can be no lower, and (2) if the input grades are all replaced by strictly higher grades, then the final grade must be strictly higher. We believe that this makes practical sense. But it is of interest to investigate the consequences of dropping the second condition. An aggregation function is *weakly monotonic* when only the first of the two conditions is satisfied. Andrew Jennings (2009) has identified and characterized the family of weakly monotonic aggregation functions that are strategy-proof-in-grading.

For simplicity, think of percentages. Assume that the language is infinite and belongs to the closed real interval [0, 100]. The *linear median* (weakly monotonic aggregation function), $\mathcal{LM} : [0, 100]^n \to [0, 100]$, is defined by

$$\mathcal{LM}(\alpha_1, \ldots, \alpha_n) = \sup\left\{ y \in [0, 100] : 100\left(\frac{\#(\alpha_i \geq y)}{n}\right) \geq y \right\},$$

where $\#(\alpha_i \geq y)$ means the number of grades α_i that are greater than or equal to y, and n is the total number of grades (or voters). If a candidate's linear median is, say, 87, then at least 87% of voters assigned the candidate an 87 or higher. This idea makes sense only when there are many voters. The linear medians determine the ranking of the candidates. Jennings argues that this approach is of interest for polarized elections—when all voters submit minimal or maximal grades—because an order function returns an extreme grade, whereas $\mathcal{LM}$ returns the fraction of the voters who submit a maximal grade. The mechanism is intriguing, but the majority-gauge and majority-ranking do take into consideration such fractions.

Jennings characterizes the weakly monotonic aggregation functions that are strategy-proof-in-grading: they are the strategic medians. A function $f : [0, 100]^n \to [0, 100]$, is a *strategic median* if there exists a positive increasing function $g : [0, 100] \to [0, 100]$ and

$$f(\alpha) = \sup \left\{ y \in [0, 100] : 100 \left(\frac{\#(\alpha_i \geq y)}{n} \right) \geq g(y) \right\}.$$

The linear median is meaningful only when the common language is an interval measure (which is very rare for large electorates). However, if the language of grades is an interval measure, then the linear median is certainly a lot better than the mean (range voting).

15 Common Language: Voting

All this is strongly reminiscent of the conditions existant at the beginning of the theory of heat: that too was based on the intuitively clear concept of one body feeling warmer than another, yet there was no immediate way to express significantly by how much, or how many times, or in what sense.
—John von Neumann and Oskar Morgenstern

The majority judgment relies on a language of evaluation that is common to the judges of a jury or the voters of an electorate. Practice shows that in judging competitions—of wines, divers, skaters, gymnasts—the numbers used by the judges of juries are defined by rules and regulations and constitute a well-understood language of grades that becomes better understood through use. Much the same may be expected to happen in voting with the majority judgment in large electorates: use will increasingly impart meaning; over many elections the language will become common. And when the method is actually used for the first time in a political election, polls, media coverage, campaign materials, debates, and experimental trials will give the public a good sense of what the language of grades means. In any case, everyone is used to questionnaires concerning the quality of services, objects, or candidates in which they are asked to respond in a simple language such as that of the 2007 Orsay experiment. Moreover, the participants in this experiment had no trouble whatsoever in grading the candidates.

In judging competitions, the rules and regulations give the definitions; in voting, the ballots define the language. The problem of evaluating competitors such as pianists, figure skaters, divers, or wines is, however, somewhat different from the problem of evaluating candidates for political office. Judges of competitions are professionally competent, and they have a collective absolute sense of the excellence of the performances or the qualities of the competing entities, but in political elections, however competent the voters may be, they can have very different evaluations of the candidates because of fundamentally different

conceptions of how society should be run and organized. The deep meaning of the grades of a common language for evaluating candidates for political office is an elusive philosophical concept. There is no known procedure for being able to assert that the words *Excellent* or *Acceptable* mean exactly the same thing to different voters (but, of course, that is also true of the word *green*, for instance).

And yet, people talk, they write, they evaluate, they communicate. Words *do* carry meanings, meanings that depend on the language—English, French, Spanish—and also on the culture—British versus American versus Indian, French versus Canadian versus Swiss, Spanish versus Mexican versus Chilean.

The ballot posed the solemn question, To be president of France, after having taken every consideration into account, I judge in conscience that this candidate would be:

and invited the voter to evaluate every candidate. The words of the language of grading that they were to use,

Très Bien (Excellent), *Bien* (Very Good), *Assez Bien* (Good), *Passable* (Acceptable), *Insuffisant* (Poor), *à Rejeter* (To Reject),

are intimately linked with the question posed and have clear cultural meanings independent of voting that are, by and large, shared by French voters (as well as by English speakers). The choice of those six words was made after considerable discussion and consultation. The first five are known to every French school child, to all those that sat baccalauréat examinations marking the end of secondary school, and to university students. There may at times be a slight hesitation between two successive grades—*Poor* and *Acceptable*, or *Excellent* and *Very Good*—but there is surely none between grades that are two or more apart, such as *Good* and *Poor*. Voters faced with such dilemmas would prefer a still richer language—the possibility of assigning a *Poor*+ or an *Acceptable*−, an *Excellent*− or a *Very Good*+. Imagine a population asked to evaluate colors: reds, yellows, greens, blues are flashed on a screen, and individuals are to identify them. The deep greens of the Amazonian forests and the fragile greens of the new leaves of springtime are *green* to most people, but many would prefer identifying them with additional adjectives. There may be hesitation as greens gradually fade into blues, for example, but as Wittgenstein said, "the meaning of a word is its use in the language."

One fact is clear. A voter is better able to express his opinion by assigning to candidates one of the six grades used by the majority judgment than by giving rank-orders of candidates. This fact emerges from various experiments. Asked to rank-order the candidates, over 50% of the voters rank-ordered at most six

of the twelve candidates in the 2007 French presidential election (see table 6.4); rank-ordering is at once too difficult and too constraining. In the 2007 Orsay experiment, 48% of the ballots had no *Excellent*, 11% had no *Excellent* and no *Very Good*, 2% had no *Good* or above. To be listed first in a rank-order carries very different meanings. To be listed second (or third or in any place) in a rank-order also carries very different meanings: two-thirds of the second highest grades are merely *Good* or worse, one-quarter of the second highest grades are *Acceptable* or worse (see table 6.3). The traditional model and methods aggregate inputs that have considerably less common meaning than the inputs aggregated by the majority judgment. Moreover, even with twelve candidates, fully 86% of the majority judgment voters chose to use only five different grades. This suggests that six is the optimal number of grades that permit absolute judgments (less than Miller's seven; in keeping with the six for the distinction of tones). The number was deliberately chosen to be even so that there would be no middle grade, and there are four positive and only two negative grades, in keeping with a sense that candidates for public office should be in any case exceptional persons. The six grades are also meant to represent very distinctive classifications. Hesitation between neighboring grades is more likely due to difficulties of approximation in the scale of grades than to a lack of understanding of their meanings.

The extensive experience with competitions of other types—sports, products, musical performances—validates Wittgenstein's assertion that over time grades acquire clear, commonly held meanings. There is no reason for this not to happen with a language of grades to evaluate candidates. On the other hand, it is impossible to prove that the meanings of the evaluations were in fact shared by French voters (any more than one can be sure about the shared value of any word or concept in any language). What can be asked is circumstantial: Are the results consistent with the hypothesis that the meanings of the grades are shared? Werner Heisenberg, faced with a similar difficulty in establishing the uncertainty principle, explained, "We believe we have gained *anschaulich*[1] understanding of a physical theory, if in all simple cases, we can grasp the experimental consequences qualitatively and see that the theory does not lead to contradictions."[2]

Two distinct answers are given to this question. First, it is shown that the results make sense. They are consistent with themselves and with the known

1. *Anschaulich* means intuitively intelligible, visualizable.

2. "The more precisely the position [of a particle] is determined, the less precisely the momentum is known, and *vice versa*." See Hilgevoord and Uffink (2008).

facts. Had the grades no common, cultural meanings, there would be no reason for the results to make sense nor for them to agree with observed facts. Second, the actual use of the grades is studied (as opposed to their meanings). Like-minded people tend naturally to grade in the same way. Use, however, depends crucially on the set of candidates. A very weak set of candidates will elicit different grades than a very strong set of candidates. A set of candidates with a right rather than a left coloring (or the opposite) will elicit grades that are very different when they come from the left rather than the right. The French presidential election of 2007 presented twelve candidates who covered the entire political spectrum from the far left to the far right, so it became possible for all the voters to use the grades with the same frequencies. Had the French electorate been presented only with the seven candidates of the left or only with the four candidates of the right (see table 6.8), the voters of the left would have used the language of grades very differently from the voters of the right, so the language of grades could not be expected to be approximately the same for all voters. But with an entire spectrum of opinions, the ballots showed that the words were used very much in the same way, even though the three separate voting precincts did not elect the same candidate with the majority judgment. The hypothesis of a common language is reinforced and certainly not contradicted.

15.1 The 2007 Orsay Experiment: Validation

This was a *field* experiment: real voters were asked to express themselves in a real contest almost at the same time as they cast their votes in the actual election. They came to the experiment with their real opinions or utilities. Nothing in their incentives or beliefs was controlled, and no treatments were assigned or analyzed, as is done in *laboratory* experiments where students are offered artificial monetary incentives under varying conditions. We wished to find out whether real, uncontrollable voters of vastly different opinions and educational backgrounds could easily and intelligently evaluate a wide spectrum of candidates in a common language of grades. This is impossible in a laboratory setting. The fact is that their natural incentive was to answer honestly (though this is totally uncontrollable) because their evaluations would have no real consequence. The analysis of the results suggests that they did respond honestly. If such is the case, the data are an accurate expression of their true opinions, so their preferences between pairs of candidates can be deduced. This allows different methods of voting to be tested, analyzed, and compared on the basis of real data. To our knowledge this is the first database that allows such analyses.

We had no idea as to what would happen: perhaps voters would refuse to participate, perhaps they would use only extreme grades, perhaps a minor or extreme candidate would emerge, perhaps no sense could be made of the distributions of grades. This is why this was a real, risky experiment to test a new method of voting. Much more experimentation would, of course, be salutary.

The background of the presidential elections of 2002 and 2007, the majority judgment ballot that was used, and some of the results of the experiment have already been described (see chapters 1, 2, and 6). The experiment took place in parallel with the first round of the election on April 22, 2007, in three of Orsay's twelve voting precincts, the 1st, 6th, and 12th. Orsay is a small suburban town 22 kilometers from Paris. The three precincts were chosen as the most diverse among five that had been the scene of a previous voting experiment (conducted in 2002; see chapter 18). The experiment—the ballot and the method of ranking—was explained to potential participants well before election day in individual letters, an article in the town's quarterly magazine, posters, and an evening presentation open to all. Did all voters understand, or take the time to understand, the majority judgment method? Surely not. What is important is that voters understood the ballot. But do all voters understand the subtleties of the first- or two-past-the-post method, not to speak of Condorcet, Borda, Arrow, the alternative vote, or approval voting? Surely not.

The three precincts had 2,695 registered voters; 2,383 voted, 1,752 (74%) of those who voted participated in the experiment, and 19 majority judgment ballots were invalid,[3] leaving a total of 1,733 valid ballots in all (559 in the 1st precinct, 601 in the 6th, 573 in the 12th). After voting officially, on their way out, voters had no choice but to walk by the experimental voting booths in each of the precincts. Posters explained the experiment, and a team of three or four persons, able to answer any questions voters might have concerning it, encouraged everyone to participate. Voting in the experiment was conducted as is usual in France: voters expressed themselves in the privacy of voting booths, inserted their ballots into envelopes, and deposited them in large transparent urns. Several elected officials and commentators had darkly predicted that the ballots would be much too difficult for the voters. Quite the contrary, most voters completed their ballots in about one minute. Many proclaimed their satisfaction at being able to express their opinions concerning *all* the candidates. Indeed, one of the most effective arguments for persuading reluctant voters to participate

3. Most of the nineteen invalid ballots checked more than one grade for one candidate; several were blank or expressed opposition to the experiment.

Table 15.1
Percentages of Voters, k Grades ($k = 1, \ldots, 6$)

1 grade	1%
2 grades	2%
3 grades	10%
4 grades	31%
5 grades	42%
6 grades	14%

was that the majority judgment allows a much fuller expression of opinion. With twelve candidates, the official vote allowed one of thirteen messages; the majority judgment allowed more than two billion possible messages.[4] A natural question elicits a quick and natural answer, so despite the vast number of possible answers, assigning grades is an easy task (especially compared with ranking). Of the 1,733 valid ballots, 1,705 were different. Those that were the same had typically only *To Reject*, or an *Excellent* for Royal or Sarkozy and *To Reject* for the others.

Six possible grades assigned to twelve candidates implies that a voter was unable to express a preference between every pair of candidates. The number of different grades actually used by voters shows that in any case they did not wish to distinguish between every pair (table 15.1); only 14% used all six grades. This suggests that six grades were sufficient. A scant 3% of the voters used at most two grades, 13% at most three, suggesting that two or three grades were far from sufficient.

The distribution of the grades of each of the candidates, their majority-grades, and their majority-gauges are given in the order of the majority-ranking in table 15.2. The majority-ranking is very different from the rank-ordering obtained in the three precincts of Orsay with the current first-past-the-post system. Bayrou finished first with the majority judgment, defeating the candidates of the two major parties. Most striking, instead of finishing fourth as in the official vote, the extreme rightist Le Pen finished last: 74% of the electorate assigned him a *To Reject*. Significantly, the Green candidate Voynet placed fourth (instead of her official seventh place): the electorate was able to express the importance it attaches to problems of the environment while giving higher grades to candidates it judged better able to preside over the nation. Sarkozy later recognized this importance: his new government created one superministry, the Ministry of Ecology and Sustainable Development.

4. With twelve candidates and six grades there are $6^{12} = 2,176,782,336$ possible messages. Several participants confessed they had never voted officially before but had done so this time to be able to participate in a vote that enabled them to express their opinions.

The distribution of the grades of the majority judgment was frequently shown in talks and seminars with the names of the four major candidates (Bayrou, Le Pen, Royal, and Sarkozy) hidden. Invariably, someone in the audience was able to identify from the percentages who was who. This shows that the numbers make sense, that they contain meaningful information. What happened is, we believe, what most observers anticipated: Le Pen had an overwhelming percentage of *To Reject*; among the other three, Sarkozy had at once the highest percentages of *Excellent* and of *To Reject*; Bayrou at once the lowest percentages of *Excellent* and of *Poor* plus *To Reject*.

The results of the face-to-face confrontations between every pair of candidates may be estimated from the majority judgment ballots by comparing the grades of each. When two candidates are face-to-face, the ballot's vote goes to the one with the higher grade; when their grades are the same, each receives $\frac{1}{2}$ (to obtain these data the ballots themselves must be inspected; the data of table 15.2 do not suffice). The estimates are given in table 15.3. In particular, Royal defeats Sarkozy with 52.3% of the vote, a prediction of the outcome of the second round within 1%: in the three Orsay precincts Royal actually obtained 51.3% of the official second-round votes to Sarkozy's 48.7%. The participants seem to have expressed themselves in the majority judgment ballots in conformity with the manner in which they actually voted. The 1% difference is easily explained: 26% of the voters did not participate in the experiment; and the last two weeks of the campaign may have changed perceptions (in particular, Sarkozy clearly dominated Royal in a televised debate).

The closeness of the estimate to the outcome shows that the majority judgment ballots are consistent with the observed facts. From the face-to-face results it is possible to obtain the Condorcet- and Borda-rankings (the face-to-face majority rule gives a transitive ordering; there is no Condorcet paradox). Given at the bottom of table 15.3, they are identical with the majority-ranking except for the four last places.

That all three methods rank Bayrou first, Royal second, and Sarkozy third is not surprising: except for the *Excellent*s, whose percentages taken alone give the opposite rank-ordering, the percentages of at least *Very Good*, at least *Good*, . . . , at least *Poor* all agree with that order (table 15.4). Almost any reasonable election mechanism should agree with this ranking of the three important candidates; the first- and two-past-the-post mechanisms are an exception, but they are not reasonable.

Still another explanation for Bayrou to be the winner with all three methods is to compare how voters who gave their highest grade to each of the top three evaluated the other two (table 15.5). Sarkozy voters—voters who gave Sarkozy

Table 15.2
Majority Judgment Results, Three Precincts of Orsay, April 22, 2007

		Excellent	*Very Good*	*Good*	*Acceptable*
1st	Bayrou	13.6%	30.7%	25.1%	14.8%
2d	Royal	16.7%	22.7%	19.1%	16.8%
3d	Sarkozy	19.1%	19.8%	14.3%	11.5%
4th	Voynet	2.9%	9.3%	17.5%	23.7%
5th	Besancenot	4.1%	9.9%	16.3%	16.0%
6th	Buffet	2.5%	7.6%	12.5%	20.6%
7th	Bové	1.5%	6.0%	11.4%	16.0%
8th	Laguiller	2.1%	5.3%	10.2%	16.6%
9th	Nihous	0.3%	1.8%	5.3%	11.0%
10th	de Villiers	2.4%	6.4%	8.7%	11.3%
11th	Schivardi	0.5%	1.0%	3.9%	9.5%
12th	Le Pen	3.0%	4.6%	6.2%	6.5%

Table 15.3
Face-to-Face Elections, Percentages of Votes Estimated from Majority Judgment Ballots, Three Precincts of Orsay, April 22, 2007

	Bayrou	Royal	Sarkozy	Voynet	Besancenot	Buffet	Bové
Bayrou	–	56	60	77	77	81	83
Royal	44	–	52	73	74	78	81
Sarkozy	40	48	–	59	61	64	66
Voynet	23	27	41	–	56	59	67
Besancenot	23	26	39	44	–	53	60
Buffet	19	22	36	41	47	–	57
Bové	17	19	34	33	40	43	–
Laguiller	17	20	34	33	39	41	49
de Villiers	16	23	23	34	38	39	44
Nihous	10	15	25	25	31	32	38
Schivardi	10	13	25	21	26	27	34
Le Pen	14	19	20	26	30	31	35
Condorcet-ranking	1st	2d	3d	4th	5th	6th	7th
Borda-ranking	1st	2d	3d	4th	5th	6th	7th
Majority-ranking	1st	2d	3d	4th	5th	6th	7th

Note: For instance, Royal obtains 52% against Sarkozy; symmetrically, Sarkozy obtains 48% against Royal.

Table 15.2
(cont.)

	Poor	*To Reject*	Majority-Gauge	Official Order
Bayrou	8.4%	7.4%	(44.3%, $Good+$, 30.6%)	3d
Royal	12.2%	12.6%	(39.4%, $Good-$, 41.5%)	1st
Sarkozy	7.1%	28.2%	(38.9%, $Good-$, 46.9%)	2d
Voynet	26.1%	20.5%	(29.7%, $Accept-$, 46.6%)	7th
Besancenot	22.6%	31.1%	(46.3%, $Poor+$, 31.2%)	5th
Buffet	26.4%	30.4%	(43.2%, $Poor+$, 30.5%)	8th
Bové	25.7%	39.5%	(34.9%, $Poor-$, 39.4%)	9th
Laguiller	25.9%	40.1%	(34.2%, $Poor-$, 40.0%)	10th
Nihous	26.7%	55.0%	(45.0%, *To Reject*,–)	11th
de Villiers	15.8%	55.5%	(44.5%, *To Reject*,–)	6th
Schivardi	24.9%	60.4%	(39.7%, *To Reject*,–)	12th
Le Pen	5.4%	74.4%	(25.7%, *To Reject*,–)	4th

Table 15.3
(cont.)

	Laguiller	de Villiers	Nihous	Schivardi	Le Pen
Bayrou	83	84	90	90	86
Royal	80	77	85	87	81
Sarkozy	66	77	75	75	80
Voynet	67	66	75	79	74
Besancenot	61	62	69	74	70
Buffet	59	61	68	73	69
Bové	51	56	62	66	65
Laguiller	–	56	62	66	64
de Villiers	44	–	54	56	59
Nihous	38	46	–	53	56
Schivardi	34	44	47	–	54
Le Pen	36	41	44	46	–
Condorcet-ranking	8th	9th	10th	11th	12th
Borda-ranking	8th	9th	10th	12th	11th
Majority-ranking	8th	10th	9th	11th	12th

Table 15.4
Cumulative Majority Judgment Grades, Three Precincts of Orsay, April 22, 2007

	At Least					
	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
Bayrou	13.6%	43.3%	69.4%	84.2%	92.6%	100%
Royal	16.7%	39.4%	58.5%	75.3%	87.5%	100%
Sarkozy	19.1%	38.9%	53.2%	64.7%	71.8%	100%

Table 15.5
Grades for Top Three Candidates, by Voters Giving Their Highest Grade to One of the Other Two Candidates, Three Precincts of Orsay, April 22, 2007.

	Bayrou's Grades		Sarkozy's Grades		Royal's Grades	
	Royal Voters	Sarkozy Voters	Royal Voters	Bayrou Voters	Bayrou Voters	Sarkozy Voters
Excellent	7%	6%	3%	6%	7%	3%
Very Good	33%	28%	10%	22%	26%	13%
Good	29%	30%	16%	24%	26%	22%
Acceptable	16%	19%	15%	17%	20%	24%
Poor	9%	9%	11%	6%	13%	18%
To Reject	6%	8%	45%	25%	9%	21%

Note: TNS Sofres poll, March 14–15, 2007, showed 72% of Royal voters giving their votes to Bayrou in a second round against Sarkozy, and 75% of Sarkozy voters giving their votes to Bayrou in a second round against Royal.

their highest grade—strongly preferred Bayrou to Royal; Royal voters strongly preferred Bayrou to Sarkozy; and Bayrou voters evaluated Royal and Sarkozy about equally.

The face-to-face estimates were also used to predict the results of the second round between Royal and Sarkozy in each of the three precincts (table 15.6); they are all very close to the actual outcomes. Royal's scores are consistently though slightly overestimated. The difference, as noted, reflects changes in opinions in the two weeks that separated the two rounds of voting. This adds to the evidence that the language of grades permitted voters to correctly express their preferences and indifferences.

The ballots may also be used to predict the outcome of the official first-past-the-post vote in the first round. Assuming there was little or no strategic manipulation in the majority judgment ballots, they may be used to estimate what manipulation took place in the official first-round voting. The model of estimation assumes that the vote of a ballot goes to the candidate with the highest grade. But since the highest grade of one-third of the ballots went to

Table 15.6
Second-Round Results, Percentages of Votes Estimated from Majority Judgment Ballots versus Actual Outcomes, Three Precincts of Orsay, April 22, 2007

	Three Precincts		1st Precinct		6th Precinct		12th Precinct	
	Estimate	Outcome	Estimate	Outcome	Estimate	Outcome	Estimate	Outcome
Royal	52.3%	51.3%	48.2%	47.2%	54.4%	53.7%	54.3%	52.6%
Sarkozy	47.7%	48.7%	51.8%	52.8%	45.6%	46.3%	45.7%	47.4%

Table 15.7
First-Round Votes, Percentages of Votes Estimated from Majority Judgment Ballots, Three Precincts of Orsay, April 22, 2007

	Left					
	Buffet	Laguiller	Bové	Schivardi	Besancenot	Voynet
Estimate 1	2.6	1.6	1.6	0.4	4.9	3.5
Actual	1.4	0.8	0.9	0.2	2.5	1.7
Estimate 2	2.5	1.5	1.5	0.3	4.6	3.4
	Major					
	Royal	Bayrou	Sarkozy			
Estimate 1	25.6	25.6	28.4			
Actual	29.9	25.5	29.0			
Estimate 2	25.4	25.3	27.4			
	Right					
	Nihous	de Villiers	Le Pen			
Estimate 1	0.5	2.3	2.9			
Actual	0.3	1.9	5.9			
Estimate	0.4	1.9	5.8			

Note: In estimate 1 the vote of a ballot is split evenly among the candidates with the highest grade.
Estimate 2 is the same except when Le Pen is among those who received the highest grade, in which case he is accorded the entire vote.

more than one candidate, a further behavioral assumption is necessary. Estimate 1 assumes that the votes of ballots with several same highest grades are split equally among the relevant candidates. Estimate 2 takes into account Le Pen's peculiar niche at the far right: it is the same as estimate 1 except when Le Pen is one of the candidates with a highest grade, in which case Le Pen gets the vote. The results are given in table 15.7. Estimate 2 seems to be the better model. It explains almost perfectly the scores of the far right; 6.3% of the vote predicted

Table 15.8
Majority Judgment Results, 1st Precinct of Orsay, April 22, 2007

		Excellent	*Very Good*	*Good*	*Acceptable*
1st	Bayrou	13.2%	31.3%	25.4%	14.7%
2d	Sarkozy	22.7%	19.3%	16.3%	10.6%
3d	Royal	14.8%	19.5%	20.8%	19.5%
4th	Voynet	2.5%	8.8%	15.6%	22.0%
5th	Besancenot	3.6%	8.9%	12.3%	16.5%
6th	Buffet	1.8%	6.6%	12.3%	20.4%
7th	Bové	1.3%	5.7%	10.8%	14.0%
8th	Laguiller	2.1%	4.8%	6.4%	15.7%
9th	de Villiers	2.3%	5.2%	8.9%	11.4%
10th	Nihous	0.2%	1.6%	4.5%	8.6%
11th	Schivardi	0.4%	0.5%	3.4%	7.5%
12th	Le Pen	2.1%	2.9%	6.4%	7.5%

Table 15.9
Majority Judgment Results, 6th Precinct of Orsay, April 22, 2007

		Excellent	Very Good	Good	Acceptable
1st	Bayrou	16.6%	30.8%	22.3%	14.1%
2d	Royal	18.6%	22.6%	19.1%	15.0%
3d	Sarkozy	16.8%	20.0%	13.3%	7.5%
4th	Voynet	3.8%	9.3%	16.5%	25.6%
5th	Besancenot	3.8%	9.8%	16.0%	15.8%
6th	Buffet	3.3%	7.5%	12.5%	20.0%
7th	Bové	1.3%	5.3%	10.3%	16.1%
8th	Laguiller	2.0%	4.5%	10.1%	16.6%
9th	Nihous	0.3%	1.5%	6.8%	10.5%
10th	de Villiers	2.0%	5.7%	7.0%	10.6%
11th	Schivardi	0.3%	0.8%	2.2%	9.7%
12th	Le Pen	1.5%	4.7%	5.7%	5.8%

Table 15.10
Majority Judgment Results, 12th Precinct of Orsay, April 22, 2007

		Excellent	*Very Good*	*Good*	*Acceptable*
1st	Royal	16.4%	26.0%	17.5%	16.1%
2d	Bayrou	10.8%	30.0%	27.7%	15.5%
3d	Sarkozy	18.0%	20.1%	13.3%	12.6%
4th	Voynet	2.4%	9.9%	20.4%	23.2%
5th	Besancenot	4.9%	11.0%	20.4%	15.5%
6th	Buffet	2.3%	8.7%	12.7%	21.5%
7th	Bové	1.9%	7.0%	12.9%	18.0%
8th	Laguiller	2.3%	6.5%	11.9%	17.5%
9th	de Villiers	2.8%	8.4%	10.1%	11.9%
10th	Nihous	0.5%	2.3%	4.4%	13.8%
11th	Schivardi	0.7%	1.7%	6.1%	11.2%
12th	Le Pen	5.4%	6.3%	6.6%	6.1%

Table 15.8
(cont.)

	Poor	*To Reject*	Majority-Gauge
Bayrou	8.2%	7.2%	(44.5%, *Good*+, 30.1%)
Sarkozy	5.4%	24.8%	(42.0%, *Good*+, 41.7%)
Royal	13.1%	12.3%	(34.3%, *Good*−, 44.9%)
Voynet	30.4%	20.8%	(48.8%, *Poor*+, 20.8%)
Besancenot	23.4%	35.2%	(41.3%, *Poor*+, 35.2%)
Buffet	25.2%	33.6%	(41.1%, *Poor*+, 33.6%)
Bové	25.4%	42.8%	(31.8%, *Poor*−, 42.8%)
Laguiller	26.7%	42.2%	(31.1%, *Poor*−, 42.2%)
de Villiers	16.5%	55.6%	(44.4%, *To Reject*,–)
Nihous	28.6%	56.6%	(43.5%, *To Reject*,–)
Schivardi	24.7%	63.5%	(36.5%, *To Reject*,–)
Le Pen	4.8%	76.2%	(23.8%, *To Reject*,–)

Table 15.9
(cont.)

	Poor	*To Reject*	Majority-Gauge
Bayrou	9.0%	7.2%	(47.4%, *Good*+, 30.3%)
Royal	13.3%	11.3%	(41.3%, *Good*+, 39.6%)
Sarkozy	7.1%	30.9%	(36.8%, *Good*−, 49.9%)
Voynet	24.1%	20.6%	(29.8%, *Accept*−, 44.8%)
Besancenot	24.5%	30.1%	(46.3%, *Poor*+, 31.2%)
Buffet	27.3%	29.5%	(43.3%, *Poor*+, 29.5%)
Bové	28.3%	38.6%	(33.1%, *Poor*−, 38.6%)
Laguiller	27.3%	39.4%	(33.3%, *Poor*−, 39.4%)
Nihous	25.1%	55.7%	(44.3%, *To Reject*, –)
de Villiers	15.8%	58.9%	(41.1%, *To Reject*, –)
Schivardi	26.5%	60.6%	(39.4%, *To Reject*, –)
Le Pen	4.8%	77.5%	(22.5%, *To Reject*, –)

Table 15.10
(cont.)

	Poor	*To Reject*	Majority-Gauge
Royal	10.1%	14.0%	(42.4%, *Good*+, 40.1%)
Bayrou	8.0%	7.9%	(40.8%, *Good*+, 31.4%)
Sarkozy	8.4%	27.7%	(38.0%, *Good*−, 48.7%)
Voynet	23.9%	20.1%	(32.8%, *Accept*−, 44.0%)
Besancenot	19.9%	28.3%	(36.3%, *Accept*−, 48.2%)
Buffet	26.5%	28.3%	(45.2%, *Poor*+, 28.3%)
Bové	23.4%	36.8%	(39.8%, *Poor*+, 36.8%)
Laguiller	23.6%	38.4%	(38.0%, *Poor*−, 38.4%)
de Villiers	15.0%	51.8%	(48.2%, *To Reject*,–)
Nihous	26.4%	52.7%	(47.3%, *To Reject*,–)
Schivardi	23.4%	56.9%	(43.1%, *To Reject*,–)
Le Pen	6.5%	69.1%	(30.9%, *To Reject*,–)

for the left went to others, about three-quarters to Royal, one-quarter to Sarkozy, almost none to Bayrou. That none went to Bayrou counters the conventional wisdom.

The results in each of the three precincts (tables 15.8–15.10) were very similar to the results in all three precincts together. The set of first three candidates is always the same; the fourth-place (Voynet) through eighth-place (Laguiller) candidates are identical; the ninth and tenth-places oscillate between two candidates (Nihous and de Villiers); and the eleventh and twelfth places are identical (respectively, Schivardi and Le Pen).

Among the three, Bayrou is not always first, Sarkozy is not always last. In every case each of the three candidates has a majority-grade of *Good*. The results in the 12th precinct are of particular interest for at least two reasons. The first is that Royal is the majority judgment winner, but the face-to-face estimates show that the Condorcet- and Borda-winner is Bayrou (table 15.11). Here is a clear instance showing that the methods of Condorcet and Borda favor the centrist candidate more than the majority judgment does (see chapter 6).

Could Sarkozy have won in all of France with the majority judgment? The distribution of the official votes in the three precincts of Orsay was very far from the vote in the entire nation (see table 6.13). Sarkozy had 2% less in Orsay than nationally, Royal was the official winner in Orsay with 4% more, Bayrou had 7% more, and Le Pen almost 5% less. But if the scores of Royal and Sarkozy in the 12th precinct are switched, the distribution of the votes is close to that of the national percentages (table 15.12).

Setting aside the obvious—a different method of election induces a different electoral campaign—this suggests Sarkozy could well have been elected with the majority judgment nationally. The statistical evidence (see tables 6.14a and 6.14b) bears this out. Sarkozy is the most probable winner with the majority judgment when samples are drawn from a database whose distribution of the votes approximates that of all of France. This suggests that the majority judgment is not unduly biased in favor of a centrist candidate. On the other hand,

Table 15.11
Projected Second-Round Results, 12th Precinct of Orsay, April 22, 2007

	Bayrou	Royal	Sarkozy
Bayrou	–	53.5%	59.0%
Royal	46.5%	–	54.3%
Sarkozy	41.0%	45.7%	–

Table 15.12
Official First-Round Votes, National and 12th Precinct of Orsay, April 22, 2007

	National	12th Precinct, Orsay		National	12th Precinct, Orsay
Sarkozy	31.2%	*32.0%*	Buffet	1.9%	2.3%
		↑↓	Voynet	1.6%	1.3%
Royal	25.9%	*26.6%*	Laguiller	1.3%	1.2%
Bayrou	18.6%	20.2%	Bové	1.3%	0.8%
Le Pen	10.4%	10.0%	Nihous	1.1%	0.2%
Besancenot	4.1%	2.7%	Schivardi	0.3%	0.0%
de Villiers	2.2%	2.5%			

Note: Sarkozy's and Royal's 12th precinct scores are switched here, for comparison. The actual percentages were Royal 32.0%, Sarkozy 26.6%.

Sarkozy's chances of winning with majority judgment are considerably less than with first- and two-past-the-post.

In all these simple cases the qualitative consequences of the majority judgment make eminent sense.

15.2 Common Use of Grades: Raw Data

Did all the voters use the language in the same way? Did subsets of the voters use each of the words on average about the same number of times? The immediate answer is an unqualified yes. In each of the three precincts, the average numbers or percentages of *Excellent*s, *Very Good*s, down to *To Reject*s were remarkably close to those of all three precincts taken together (table 15.13). But is this exceptional, or are such results to be expected when the subpopulations contain well over 500 ballots (each of the precincts) or consist of 100 or 50 ballots drawn at random?

This regularity holds in the frequency with which every grade is used in each voting precinct (table 15.14). For example (see bold face numbers), in

Table 15.13
Average Number of Grades (and Percentages) per Majority Judgment Ballot, Three Precincts of Orsay, April 22, 2007

Precinct	*Excellent*		*Very Good*		*Good*		*Acceptable*		*Poor*		*To Reject*	
1st	0.7	5.6%	1.2	9.6%	1.5	12.1%	1.7	14.0%	2.3	19.3%	4.7	39.4%
6th	0.7	5.9%	1.2	10.2%	1.4	11.8%	1.7	14.3%	2.3	19.5%	4.6	38.4%
12th	0.7	5.7%	1.4	11.5%	1.6	13.7%	1.8	15.2%	2.2	17.9%	4.3	36.0%
All three	0.7	5.7%	1.3	10.4%	1.5	12.5%	1.7	14.5%	2.3	18.9%	4.5	37.9%

Table 15.14
Frequencies (in Percentages) in the Use of Grades, Three Precincts of Orsay, April 22, 2007

		No. of Times Grades Used in a Ballot (%)								
	Precinct	0×	1×	2×	3×	4×	5×	6×	7×	> 8×
Excellent	1st	47.0	43.1	7.7	1.6	0.2	0.2	0.0	0.0	0.2
	6th	46.6	41.8	8.7	2.0	0.7	0.0	0.2	0.0	0.2
	12th	51.1	37.3	7.9	2.3	0.9	0.2	0.0	0.0	0.3
Very Good	1st	30.2	40.3	**19.7**	6.8	1.1	1.3	0.5	0.2	0.0
	6th	28.8	37.9	**22.0**	7.2	2.7	0.8	0.3	0.3	0.0
	12th	26.0	37.9	**20.4**	8.2	4.4	2.1	0.7	0.3	0.0
Good	1st	24.3	35.1	22.2	11.4	4.7	1.4	0.7	0.2	0.0
	6th	26.3	35.1	20.5	10.1	5.3	2.2	0.3	0.2	0.0
	12th	21.8	30.4	25.5	12.0	7.2	2.3	0.3	0.3	0.2
Acceptable	1st	23.3	29.3	20.0	16.8	6.4	3.6	0.2	0.0	0.4
	6th	22.6	28.8	24.1	13.0	6.5	3.7	0.3	0.5	0.5
	12th	22.5	23.0	24.6	17.1	7.3	3.8	0.5	0.9	0.2
Poor	1st	16.5	20.0	22.9	15.9	14.0	5.5	2.9	1.4	0.9
	6th	16.3	24.0	19.5	17.0	9.5	5.7	5.8	1.0	1.3
	12th	23.2	20.8	18.5	15.2	10.6	6.1	3.1	1.4	1.0
To Reject	1st	3.0	6.1	10.7	12.0	16.3	17.2	10.4	9.3	15.0
	6th	4.7	4.7	9.2	17.0	18.1	14.5	11.0	7.3	13.6
	12th	7.0	7.3	14.5	14.0	14.5	13.8	7.3	7.0	14.7

the 1st precinct 19.7% of the ballots had two *Very Good*s, in the 6th precinct the percentage of two *Very Good*s was 22.0%, and in the 12th it was 20.4%. The corresponding triplets throughout the table are surprisingly close to each other. But again, is this exceptional, or are these results to be expected in any case with subsets that contain well over 500 ballots?

To determine whether the grades of the language of evaluation are really used in the same way—to see that the raw data are exceptional and not simply the natural consequence of the sample sizes—a finer analysis is needed. This analysis is rather technical, and readers may prefer to skip this material to go directly to the conclusions.

Let $X = \{1, \ldots, 6\}$ be a discrete random variable whose value represents the grade a voter gives a candidate: *Excellent* is 6, *Very Good* 5, *Good* 4, *Acceptable* 3, *Poor* 2, and *To Reject* 1. The greater their numerical difference, the greater is the difference between two grades. Let G_X be the underlying (cumulative) distribution function and $\hat{G}_X$ the observed or empirical (cumulative) distribution function for all of Orsay. The vector of percentages (divided by 100)[5] given in

5. The exact numbers sum to 1; rounding them to three significant places may result in a sum slightly above or below 1, as in this case.

table 15.13 gives the observed density function for all of Orsay,

$$\hat{g}_X = (\overbrace{0.379}^{\textit{To Reject}}, \overbrace{0.189}^{\textit{Poor}}, \overbrace{0.145}^{\textit{Acceptable}}, \overbrace{0.125}^{\textit{Good}}, \overbrace{0.104}^{\textit{Very Good}}, \overbrace{0.057}^{\textit{Excellent}}),$$

and so also the (cumulative) distribution of Orsay,

$$\hat{G}_X = (\overbrace{0.379}^{\textit{To Reject}}, \overbrace{0.568}^{\preceq\textit{Poor}}, \overbrace{0.713}^{\preceq\textit{Acceptable}}, \overbrace{0.838}^{\preceq\textit{Good}}, \overbrace{0.942}^{\preceq\textit{Very Good}}, \overbrace{1.000}^{\preceq\textit{Excellent}}).$$

The raw data are comforting because the difference between the distribution of any of the subpopulations and the empirical distribution of the whole is small. In general, fix a subpopulation of some size M, and let $d(M)$ be the distance between the observed distributions of X of the subpopulation and the observed distribution $\hat{G}_X$ of Orsay. Imagine that for every subpopulation of size M the distance is computed, giving the distribution of $d(M)$, call it $F_{d(M)}$. The difficulty is that it is impossible to generate $F_{d(M)}$ because there are far too many subpopulations of any reasonably sized M; for example, if $M = 20$, there are $C_{1733}^{20} \approx 10^{46}$ different subpopulations (an amount far less than a googol, 10^{100}, but far more than the total number grains of sand of all the world's beaches, which is estimated to be less than 10^{19}). In consequence, a Monte Carlo approach is used to estimate $F_{d(M)}$: a large number of subpopulations of size M are drawn randomly and independently from the whole population, and their empirical distribution is taken as an approximation of $F_{d(M)}$.

Two questions at once present themselves: How should $d(M)$ be defined? How many subpopulations must be drawn for the results to be significant?

If $q_X = (q_1, \ldots, q_n)$ and $q'_X = (q'_1, \ldots, q'_n)$ are density functions of grades when the language contains n grades, their respective distribution functions are

$$Q_X = (q_1, q_1 + q_2, q_1 + q_2 + q_3, \ldots, 1)$$

and

$$Q'_X = (q'_1, q'_1 + q'_2, q'_1 + q'_2 + q'_3, \ldots, 1).$$

Define the distance between them to be

$$d(Q_X, Q'_X) = \frac{1}{n-1} \sum_{1}^{n-1} \left| Q_X - Q'_X \right| = \frac{1}{n-1} \sum_{i}^{n-1} \left| \sum_{k}^{i} (q_k - q'_k) \right|.$$

When $n = 6$, the two density functions that are furthest apart are $q = (0, 0, 0, 0, 0, 1)$ and $q' = (1, 0, 0, 0, 0, 0)$, which correspond to the two distribution functions $Q = (0, 0, 0, 0, 0, 1)$ and $Q' = (1, 1, 1, 1, 1, 1)$, so $d(Q, Q') = 1$. If they

Table 15.15
Estimated Distributions of $F_{d(M)}$, 2007 Orsay Experiment

$d(M)$ in Interval	M=100	M=50	M=20	M=10
[0, .10]	100%	100%	100%	98%
[0, .05]	100%	99%	91%	75%
[0, .03]	98%	90%	64%	38%
[0, .02]	88%	68%	35%	14%
[0, .01]	43%	20%	5%	1%
[0, .005]	7%	2%	0%	0%

are different but close, for example, if Q is as before and $q' = (0, 0, 0, 0, 1, 0)$ or $Q' = (0, 0, 0, 0, 1, 1)$, then $d(Q, Q') = \frac{1}{5}$.

If $\hat{Q}_X$ is the observed distribution function of a subpopulation of size M, its *distance from the observed distribution of all of Orsay* is $d(M) = d(\hat{Q}_X, \hat{G}_X)$.

Letting G be the distribution function of a random variable Y, and $\hat{G}_n$ an empirical distribution function obtained from n independent realizations of Y, the Dvoretzky–Kiefer–Wolfowitz (1956) inequality asserts

$$P\{\sup_{x\in\mathbf{R}} |\hat{G}_n(x) - G(x)| \geq \epsilon\} \leq 2e^{-2n\epsilon^2}.$$

It implies that 738 independent samples assure an approximation of G with 95% precision and 95% confidence, whereas 5,756 independent samples assure an approximation of G with 98% precision and 98% confidence. For Orsay, 10,000 random independent samples are taken, so those results assure that $F_{d(M)}$ is estimated with a little over 98% precision and a little over 98% confidence.

The approximations of $F_{d(M)}$ are given in table 15.15. Thus, for example, 90% of the 10,000 computations yield $d(50) \leq .03$. When $M = 100$, the distance is always at most .05, and $d(100) \leq .03$ for 98% of the subpopulations. The distances, of course, increase as M decreases. Yet even for $M = 10$, 98% of the distances are at most .10. But how is this to be evaluated? Are these distances small, or not?

15.3 Measuring Homogeneity of Voters' Grades

Imagine the set $\mathcal{P}(\hat{G}_X)$ of all populations of $N = 1,733$ voters that assign one of the six grades to $L = 12$ candidates according to the distribution function $\hat{G}_X$. Assume that each voter assigns grades to candidates according to his

own distribution of grades independently of the other voters, and assigns a grade to each candidate independently of the other candidates. Thus the population as a whole generates $L \times N$ independent grades. A vast number of different populations—voters behaving in very different ways—may assign grades whose distribution is $\hat{G}_X$.

A *homogeneous population* (H-P) is a population in $\mathcal{P}(\hat{G}_X)$ whose voters use the language of grades independently in the same way: every voter assigns grades according to the distribution $\hat{G}_X$, that is, the grade given to a candidate by a voter is drawn i.i.d. from $\hat{G}_X$. This implies that if two voters were to judge a great many candidates, then the distributions of their grades would be the same even if their preferences over the candidates were completely different. It also implies that with a fixed number of candidates but a large number of voters, the total distribution of the grades would be very close to $\hat{G}_X$.

A *nonhomogeneous population* (non-H-P) is a population in $\mathcal{P}(\hat{G}_X)$ whose voters are at the opposite pole: every voter assigns a same grade to every candidate, so the voters belong to six different mutually exclusive sets, those who always cast an *Excellent*, . . . , those who always cast a *To Reject*. This implies that those who always vote *Excellent* constitute 5.7% of the population, . . . , those who always vote *To Reject* constitute 37.9% of the population.

The idea of this approach is to situate the observed behavior of the actual population on the line going from homogeneous to nonhomogeneous population. If the actual behavior is relatively close to homogeneous behavior, then the language has been used in about the same way; if it is relatively close to nonhomogeneous behavior, then the language has not been used in the same way. Of course, the use of the grades depends on the candidates. When a population includes a wide sprectrum of opinion and so do the candidates (as was the case in the French presidential election of 2007), then if there is a common language, the observed behavior in the use of grades should be homogeneous. But if the candidates do not match the population's diversity of opinion, then if there is a common language, the use of the grades should be expected to be less homogeneous. If all the candidates are of the left, the population to the right will assign grades quite differently from the population to the left.

$F_{d(M)}$ is estimated as it was estimated for the Orsay population (Orsay-P). However, whereas non-H-P and Orsay-P are fixed, well-defined populations of 1,733 voters from which samples may be drawn, there is no fixed, well-defined population of voters for the H-P model, which is defined probabilistically. Accordingly, the estimates of $F_{d(M)}$ are carried out as follows:

· For non-H-P and Orsay-P, draw 10,000 random samples of size M and calculate the $d(M)$s that determine the approximation.

· For H-P, generate ten different base populations of 1,733 ballots, each ballot according to the distribution $\hat{G}_X$; draw 10,000 random samples of size M and calculate the $d(M)$s for each base population; the average values over the ten base populations determine the approximation. Ten sufficed because they were almost the same.

The results are given in table 15.16a. Focus, for example, on the percentage of the subpopulations for which $d(M) \leq .05$.

· When $M = 100$, all the samples from the homogeneous population (H-P) and the Orsay population (Orsay-P) are within that distance, whereas only 88% of those from the nonhomogeneous population (non-H-P) are within it.

· When $M = 50$, all the samples from H-P are within it, 99% of the samples from Orsay-P are within it, whereas only 65% of those from non-H-P are.

· When $M = 20$, 98% from H-P are within it, 91% from Orsay-P are within it, whereas a mere 29% from non-H-P are.

· When $M = 10$, 90% from H-P are within it, 75% from Orsay-P are within it, whereas only 8% from non-H-P are.

Table 15.16a
Estimated Distributions of $F_{d(M)}$, 2007 Orsay Experiment

$d(M)$	M=100			M=50		
in Interval	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	100%	100%	100%	100%	98%
[0, .05]	100%	100%	88%	100%	99%	65%
[0, .03]	100%	98%	54%	98%	90%	26%
[0, .02]	97%	88%	23%	86%	68%	7%
[0, .01]	65%	43%	2%	37%	20%	0%
[0, .005]	16%	7%	0%	5%	2%	0%
$d(M)$	M=20			M=10		
in Interval	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	100%	81%	100%	98%	55%
[0, .05]	98%	91%	29%	90%	75%	8%
[0, .03]	82%	64%	6%	58%	38%	1%
[0, .02]	54%	35%	1%	27%	14%	0%
[0, .01]	10%	5%	0%	3%	1%	0%
[0, .005]	1%	0%	0%	0%	0%	0%

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Table 15.16b
Estimated Quantiles r_k of $F_{d(M)}$, 2007 Orsay Experiment

	M=100			M=50		
Quantile	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0233	.0315	.0804	.0335	.0457	.1153
75%	.0116	.0155	.0393	.0165	.0223	.0564
50%	.0081	.0108	.0284	.0117	.0159	.0404
	M=20			M=10		
Quantile	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0542	.0750	.1884	.0782	.1083	.2532
75%	.0263	.0350	.0913	.0376	.0500	.1313
50%	.0190	.0245	.0658	.0268	.0350	.0916

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay Population.

It is clear that on the spectrum from the homogeneous to the nonhomogeneous population, the Orsay voters are very much closer to H-P than to non-H-P.

These results may be more easily appreciated by considering the kth quantile r_k of the distribution of $d(M)$ in each of the cases. It is defined by $k\%$ of the values of $d(M)$ belonging to the interval $[0, r_k]$. Three quantiles are compared in table 15.16b. They elicit the same remark: the quantiles for the Orsay population are much closer to those of the homogeneous than to the nonhomogeneous population.

Now suppose that the only candidates in the French election were the seven of the left (see table 6.8); that is, ignore the grades given to the other five candidates. Exactly the same analysis may be applied to this situation. In this case the observed density function of all of Orsay is

$$\hat{g}_X = (\overbrace{0.335}^{To\ Reject}, \overbrace{0.234}^{Poor}, \overbrace{0.170}^{Acceptable}, \overbrace{0.130}^{Good}, \overbrace{0.088}^{Very\ Good}, \overbrace{0.043}^{Excellent}),$$

so the observed distribution function is

$$\hat{G}_X = (\overbrace{0.335}^{To\ Reject}, \overbrace{0.569}^{\preceq Poor}, \overbrace{0.739}^{\preceq Acceptable}, \overbrace{0.869}^{\preceq Good}, \overbrace{0.957}^{\preceq Very\ Good}, \overbrace{1.000}^{\preceq Excellent}).$$

As mentioned earlier, there is no reason to expect homogeneous behavior in this case. The results are given in tables 15.17a and 15.17b. For the seven candidates of the left, the quantiles of the Orsay population are relatively much less close to the homogeneous population than they were for all twelve candidates.

Table 15.17a
Estimated Distributions of $F_{d(M)}$, Seven Candidates of the left, 2007 Orsay Experiment

$d(M)$ in Interval	M=100			M=50		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	100%	100%	100%	100%	98%
[0, .05]	100%	99%	91%	100%	93%	69%
[0, .03]	99%	88%	57%	94%	70%	28%
[0, .02]	92%	67%	25%	74%	42%	8%
[0, .01]	47%	22%	2%	20%	8%	0%
[0, .005]	7%	3%	0%	2%	1%	0%
$d(M)$ in Interval	M=20			M=10		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	98%	84%	100%	89%	57%
[0, .05]	95%	71%	31%	81%	48%	8%
[0, .03]	68%	38%	6%	41%	17%	0%
[0, .02]	36%	15%	1%	15%	5%	0%
[0, .01]	4%	1%	0%	1%	0%	0%
[0, .005]	0%	0%	0%	0%	0%	0%

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Table 15.17b
Estimated Quantiles r_k of $F_{d(M)}$, Seven Candidates of the left, 2007 Orsay Experiment

Quantile	M=100			M=50		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0277	.0485	.0745	.0404	.0688	.1055
75%	.0141	.0228	.0375	.0204	.0329	.0545
50%	.0103	.0157	.0274	.0148	.0225	.0392
Quantile	M=20			M=10		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0652	.1108	.1735	.0902	.1608	.2459
75%	.0330	.0523	.0865	.0458	.0749	.1265
50%	.0238	.0351	.0635	.0334	.0510	.0907

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Now suppose that the only candidates in the French election were the four of the right (see table 6.8); that is, ignore the grades given to the other eight candidates, and repeat the same analysis. The observed density function of all of Orsay is in this case

$$\hat{g}_X = (\overbrace{0.533}^{\textit{To Reject}}, \overbrace{0.137}^{\textit{Poor}}, \overbrace{0.101}^{\textit{Acceptable}}, \overbrace{0.086}^{\textit{Good}}, \overbrace{0.082}^{\textit{Very Good}}, \overbrace{0.062}^{\textit{Excellent}}),$$

so the observed distribution function is

$$\hat{G}_X = (\overbrace{0.533}^{\textit{To Reject}}, \overbrace{0.670}^{\preceq \textit{Poor}}, \overbrace{0.771}^{\preceq \textit{Acceptable}}, \overbrace{0.857}^{\preceq \textit{Good}}, \overbrace{0.939}^{\preceq \textit{Very Good}}, \overbrace{1.000}^{\preceq \textit{Excellent}}).$$

The results are given in tables 15.18a and 15.18b. For the four candidates of the right, the quantiles of the Orsay population are again relatively close to the homogeneous population, much as they were for all twelve candidates (table 15.18b). That is a surprise: one would expect them to be less close. Why? One clear difference among the three cases is that as the number of candidates decrease, the quantiles increase. For example, when $M = 50$, r_{99}(H-P) = .0335 for all twelve candidates, r_{99}(H-P) = .0404 for seven candidates, and r_{99}(H-P) = .0578 for four candidates. The same is true for all values of M.

It has already been observed that with a very large number of candidates, the distributions of the grades of any two voters in a homogeneous population would be (almost) the same, namely $\hat{G}_X$. By the law of large numbers, as the number of candidates increases, the distance between the homogeneous population and $\hat{G}_X$ approaches zero. Note in passing that if each of the six values of $12\hat{G}_X$ were integer (e.g., $12\hat{G}_X = (1, 2, 3, 5, 8, 12)$, meaning the average number of *Excellent* is 1, of *Very Good* 1, of *Good* 1, of *Acceptable* 2, of *Poor* 3, and of *To Reject* 4), then it would have been possible to generate a random population with distance from $\hat{G}_X$ exactly 0.

A *perfectly homogeneous population* (P-H-P) is one whose distance $d(M)$ from $\hat{G}_X$ is zero. As was observed, the distance of a homogeneous from a perfectly homogeneous population varies as the number of candidates varies: the fewer the number of candidates, the greater the distance.

What is very striking in the data is that for each M the distributions $F_{d(M)}$ of the distances of the nonhomogeneous populations from the perfectly homogeneous population are almost identical whether it concerns all candidates, the seven candidates of the left, or the four candidates of the right. For example, compare the distributions for $M = 50$ (table 15.19).

The same is true for each of the other values of M (as the reader may verify). It is as if in each case there is an absolute, identical nonhomogeneous wall.

Table 15.18a
Estimated Distributions of $F_{d(M)}$, Four Candidates of the Right, 2007 Orsay Experiment

$d(M)$ in Interval	M=100			M=50		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	100%	100%	100%	99%	97%
[0, .05]	99%	99%	88%	97%	91%	68%
[0, .03]	94%	87%	58%	79%	66%	30%
[0, .02]	77%	65%	28%	52%	40%	9%
[0, .01]	29%	21%	3%	10%	6%	0%
[0, .005]	3%	2%	0%	1%	0%	0%
$d(M)$ in Interval	M=20			M=10		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
[0, 1]	100%	100%	100%	100%	100%	100%
[0, .10]	100%	97%	82%	95%	87%	57%
[0, .05]	83%	70%	33%	59%	47%	9%
[0, .03]	48%	37%	7%	22%	16%	0%
[0, .02]	20%	15%	1%	6%	4%	0%
[0, .01]	2%	1%	0%	0%	0%	0%
[0, .005]	0%	0%	0%	0%	0%	0%

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Table 15.18b
Estimated Quantiles r_k of $F_{d(M)}$, Four Candidates of the Right, 2007 Orsay Experiment

Quantile	M=100			M=50		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0401	.0520	.0814	.0578	.0735	.1147
75%	.0193	.0235	.0386	.0278	.0345	.0557
50%	.0135	.0160	.0272	.0195	.0227	.0387
Quantile	M=20			M=10		
	H-P	Orsay-P	non-H-P	H-P	Orsay-P	non-H-P
99%	.0926	.1108	.1866	.1352	.1608	.2734
75%	.0438	.0523	.0883	.0634	.0749	.1266
50%	.0310	.0351	.0634	.0443	.0510	.0908

Note: H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Table 15.19
$F_{d(50)}$ for Different Sets of Candidates

	$d(M)$ in Interval						
	[0, .005]	[0, .01]	[0, .02]	[0, .03]	[0, .05]	[0, .10]	[0, 1]
All 12 candidates	0%	0%	7%	26%	65%	98%	100%
7 candidates of the left	0%	0%	8%	28%	69%	98%	100%
4 candidates of the right	0%	0%	9%	30%	68%	97%	100%

On the other hand, the distance of the homogeneous from the perfectly homogeneous population varies. Therefore, to situate the observed behavior of the actual populations on comparable lines in each of the three cases, take the ends of the line to be the perfectly homogeneous population and the nonhomogeneous population, and see where they lie in comparison with the homogeneous populations. Define the *measure of closeness* of a population P to a perfectly homogeneous population relative to the nonhomogeneous population for any k to be

$$\mu^{k\%}(\text{P}) = \frac{100r_k(\text{P})}{r_k(\text{non-H-P})}\%.$$

$\mu^{k\%}(\text{P}) = 0\%$ when P is the perfectly homogeneous population, and $= 100\%$ when P is the nonhomogeneous population. The relevant $\mu^{k\%}$ for each case is given in table 15.20.

The data are remarkably stable, as if there were an invariance. The twelve values of $\mu^{k\%}$(H-P) and of $\mu^{k\%}$(Orsay-P) for the different M and $k\%$ are almost the same in each of the three cases: all twelve candidates, the seven of the left, and the four of the right. To see the degree of homogeneity of each of the three populations of ballots, let $\bar{\mu}(\text{P})$ be the respective averages of the twelve values for each P (table 15.21).

The use of the language of grades is most homogeneous—closest to a perfectly homogeneous population—when all twelve candidates are included. The other cases are some 20% more distant. On the other hand, when there are only seven or only four candidates, their distances are necessarily greater because the corresponding homogeneous populations are lower bounds that constrain them. Relatively speaking, with respect to the H-P lower bounds, the Orsay population is $(39.2 - 29.2)/(100 - 29.2)$, or 14.1%, from the most homogeneous possible language for all twelve candidates versus 37.1% for the seven candidates of the left and 19.9% for the four candidates of the right.

Table 15.20
Measures of Closeness $\mu^{k\%}$ of Orsay-P and H-P to a Perfectly Homogeneous Population, relative to the Non-Homogeneous Population, 2007 Orsay Experiment

	All Twelve Candidates					
	$\mu^{99\%}$		$\mu^{75\%}$		$\mu^{50\%}$	
	H-P	Orsay-P	H-P	Orsay-P	H-P	Orsay-P
$M = 100$	29.0%	39.2%	29.5%	39.4%	28.5%	38.0%
$M = 50$	29.1%	39.6%	29.3%	39.5%	29.0%	39.4%
$M = 20$	29.4%	40.7%	28.8%	38.3%	28.9%	37.2%
$M = 10$	30.9%	42.8%	28.6%	38.1%	29.3%	38.2%
	Seven Candidates of the Left					
	$\mu^{99\%}$		$\mu^{75\%}$		$\mu^{50\%}$	
	H-P	Orsay-P	H-P	Orsay-P	H-P	Orsay-P
$M = 100$	37.2%	65.1%	37.6%	60.8%	37.6%	57.3%
$M = 50$	38.3%	65.2%	37.4%	60.4%	37.8%	57.4%
$M = 20$	37.6%	63.9%	38.2%	60.5%	37.5%	55.3%
$M = 10$	36.7%	65.4%	36.2%	59.2%	36.8%	56.2%
	Four Candidates of the Right					
	$\mu^{99\%}$		$\mu^{75\%}$		$\mu^{50\%}$	
	H-P	Orsay-P	H-P	Orsay-P	H-P	Orsay-P
$M = 100$	49.3%	63.9%	50.0%	60.9%	49.6%	58.8%
$M = 50$	50.4%	64.1%	49.9%	61.9%	50.4%	58.7%
$M = 20$	49.6%	59.4%	49.6%	59.2%	48.9%	55.4%
$M = 10$	49.5%	58.8%	50.1%	59.2%	48.8%	56.2%

Note: H-P = homogeneous population; Orsay-P = Orsay population.

Table 15.21
Values of $\bar{\mu}$(P), 2007 Orsay Experiment

	P-H-P	H-P	Orsay-P	non-H-P
All 12 candidates	0%	29.2%	39.2%	100%
7 candidates of the left	0%	37.4%	60.6%	100%
4 candidates of the right	0%	49.7%	59.7%	100%

Note: P-H-P = perfectly homogeneous population; H-P = homogeneous population; non-H-P = nonhomogeneous population; Orsay-P = Orsay population.

Why is the Orsay population closer to the homogeneous population for the four candidates of the right than for the seven candidates of the left? The answer is found by looking at the particular candidacies. The four candidates of the right are Sarkozy, de Villiers, Nihous, and Le Pen; 55%–60% of the voters gave a grade of *To Reject* to each of the last three, and some 70%–80% of the voters graded them *Poor* or worse (see table 15.2). The observed distribution function over the four shows 53% *To Reject* and 67% *Poor* or worse. The voters of the left must have given grades to these three candidates not all that different from the voters of the right. It is accordingly not surprising to find that the language was relatively not very far from homogeneous for the four candidates of the right. In contrast, there is little doubt that the voters of the right gave very different grades to the seven candidates of the left, as the analysis shows.

15.4 Conclusion

The analysis of the 2007 Orsay experiment demonstrates several key facts concerning the language of grades (the inputs of the majority judgment).

First, the outputs of the majority judgment are reasonable:

· Alone the distributions of the grades of candidates are sufficient for knowledgeable observers to deduce their identities.

· The distributions of the grades give accurate estimations of face-to-face confrontations and explain them.

· The distributions of the grades give accurate estimations of the results of first-round first-past-the-post voting.

· The distribution of the grades are consistent across voting precincts (though they yield different majority-orders).

In a word, since the outputs make sense, so must the inputs.

Second, populations make a remarkably homogeneous use of the means they are offered to express themselves. The grades are consistently used in the same manner by the voters. When the political spectrum of the candidates encompasses the spectrum of society's political opinion, the frequencies with which grades are used by voters are substantially the same. In this sense, the grades constitute a common language of evaluation. It must be emphasized that France was a particularly happy choice for experimentation because candidates for president traditionally represent a wide spectrum of political opinion. This finding is not exclusive to the 2007 Orsay experiment. A homogeneous use of the means offered to express opinions may also be observed in the 2002 Orsay experiment with approval voting, and in the 2007 Illkirch–Louvigny–Cigné

experiments with approval and point-summing voting (with points 0, 1, and 2) (see chapters 17 and 18).

Third, the language used and its analysis represent only first steps in the understanding of common languages of evaluation. The number and the nature of the grades that were used in the 2007 Orsay experiment seem to have been good choices, but considerably more experimentation is necessary.

Nevertheless, we believe we have gained *anschaulich* understanding of the concept of a common language. The individual use of a language of grades is almost surely biased. As was observed, different slates of candidates gave rise to different distributions of grades. To determine if there exists a common language of grades, these biases must be canceled out. They have two sources: voters' biases and candidates' biases. One way for them to be canceled is to have many candidates representing a broad political spectrum and similar precincts. This was the case in Orsay. As a consequence, there was an underlying common distribution of grades.

16 Objections to Majority Judgment

Here this or that has happened, will happen, must happen; but he invents: Here this or that might, could, or ought to happen. If he is told that something is the way it is, he will think: Well, it could probably just as well be otherwise.
—Robert Musil

The majority judgment enjoys a host of excellent properties. It is important to discover and understand under what circumstances, if any, it may fail to satisfy other desirable properties. Shortly after the theory was first publicly presented at the 8th International Meeting of the Social Choice and Welfare Society in Istanbul on July 14, 2006, objections began to emerge. The same ones have been rediscovered repeatedly. As James Stephens once remarked, "Nothing is perfect. There are lumps in it."

The first type of "lump" is perceived because of deeply ingrained attitudes anchored in habit and tradition, which often ignore the underlying spirit of the majority judgment approach, where grades and final grades are of importance not only because of the order of finish they determine but also in and of themselves. What judges or voters seek to achieve (their utility functions) no one knows, but they may well include final grades and not only winners and rankings. Some of the "defects" depend on the spirit of the traditional model, where inputs are only comparisons and in no way evaluations, so implicitly voters and judges are assumed to be indifferent to grades. Others depend on perceptions or intuitive notions tied to the idea that when grades are involved, averages (or sums) should determine results, whether the grades be numbers, letters, or words. Situations that some wish to qualify as paradoxes are, on the contrary, instances of rational decisions. These "lumps" are simply not valid objections.

The second type of lumps is related to the no-show paradox of the traditional model. Critics point out that the majority judgment violates one of the three closely related properties of proper cancellation, join-consistency, and participant-consistency. In the rare cases when these do occur with the majority

judgment it may be argued that they are not very paradoxical and indeed perhaps sensible. Moreover, it is shown that the majority judgment *is* consistent in its evaluations. In any case, it is proven in chapter 17 that the only way to avoid any one of these "lumps" implies using a point-summing method.

Averages—that is, point-summing methods—seem eminently reasonable because whenever any one voter or judge raises or lowers the grade of a competitor, the competitor's final grade is raised (a little) or lowered (a little). The same, of course, is true of the majority-values that determine the majority-rankings (though the majority-grades may not change). There are four main arguments against sums or averages: first, the number-grades must have well-defined meanings to constitute a common language; second, for sums or averages to make any sense at all the grades must belong to an interval measure (which is almost never true in voting); third, the use of sums or averages unduly favors centrist political candidates and often eliminates exceptional competitors in favor of competitors who are merely middling in every dimension; and fourth, methods based on sums or averages are by far the most manipulable.

16.1 "Majority" and "Average" Objections

The majority judgment is motivated by the need for a method that avoids Arrow's paradox, avoids Condorcet's paradox, combats strategic manipulation, and, of course, satisfies unanimity and impartiality. The inputs to the basic model are grades instead of comparisons, for otherwise there is no escaping *dependence* on irrelevant alternatives. Indeed, it has been shown that the only way to avoid the Arrow and Condorcet paradoxes is to ignore who gives what grade (theorem 9.2): only a candidate's set of grades counts. It is evident that examples may easily be invented where the outcome with the majority decision of the traditional model differs with the majority judgment or with a method that relies on sums or averages of grades. Observe that only two candidates are necessary to either attack or support the majority judgment because it is independent of irrelevant alternatives: adding a third candidate can change nothing concerning the first two.

Example 16.1: New Model versus Traditional Model The new model uses grades; the intensities of judges and voters count. When the common language is numbers with well-defined meanings, say, from a low of 0 to a high of 20, point-summing methods immediately leap to mind. Point-summing methods and the majority judgment both yield transitive rankings, but when pairs of candidates are compared by a majority of the voters' preferences, the results

are not always transitive. Thus the following example, in which $2k + 1$ voters or judges give grades to two competitors X and Y, should come as no surprise:

	k judges	1 judge	k judges
X :	20, . . . , 20,	10,	0, . . . , 0
Y :	19, . . . , 19,	9,	19, . . . , 19

If the grades are listed according to the order of the judges, and a higher grade implies a preference for that candidate, the majority candidate in the traditional model is X with $k + 1$ votes against Y with k. However, X's majority grade and average grade is 10, whereas Y's majority-grade is 19 and Y's average grade is a shade under 19, so Y is a winner over X with both methods of the new model. The two points of view are simply not compatible.

But do all the voters really see a significant difference between 20 and 19 or between 10 and 9? In a large electorate the distinction is clearly too fine: 20 and 19 are about the same, say, *Excellent*, as are 10 and 9, say, *Acceptable*, and the 0s are *To Reject*. This yields

	k judges	1 judge	k judges
X :	*Excellent, . . . , Excellent,*	*Acceptable,*	*To Reject, . . . , To Reject*
Y :	*Excellent, . . . , Excellent,*	*Acceptable,*	*Excellent, . . . , Excellent*

The difficulty simply disappears: Y is clearly preferred to X. In large elections, or with inexperienced judges, there should be relatively few grades to assure that their meanings are common to all judges or all voters.

Example 16.2: Horizontal Majority versus Vertical Majority The contrast between the majority judgment and the majority decision of the traditional model may be much starker:

	k judges	1 judge	k judges
X :	12, . . . , 12,	12,	4, . . . , 4
Y :	16, . . . , 16,	8,	8, . . . , 8

Here, under the same assumptions as in the last example, the majority candidate of the traditional model is Y with $2k$ votes against X with 1 vote, and Y is also the average-vote winner with a score of slightly under 12 to X's slightly over 8. But X's majority-grade is 12 and Y's is 8, so X is the majority judgment winner. The situation, however, is highly artificial; nothing remotely resembling it has been encountered in practice. And as a mathematical possibility, it is very rare.

Under the impartial culture assumption with $2k + 1$ voters, the probability that half the voters give to two candidates more than their majority-grades is of the order $\frac{1}{2^k}$. Moreover, one judge is able to make the majority-grade of X be any grade from 4 to 12 and the majority-grade of Y any grade from 8 to 16.

In any case, it is perfectly reasonable for X to be the winner: a majority gives X the grade 12 and a majority gives Y the grade 8. Why should this majority decision—a decision reached by looking at the example "horizontally"—be any less valid than the traditional model's majority decision—a decision reached by looking at the example "vertically"? The notion of majority is not an axiom of the traditional model. Unanimity is demanded, but it is satisfied by the majority judgment: when all of X's grades are above Y's grades, then X is ranked above Y.

Example 16.3: Majority Judgment versus Average (or Sum) A more insightful example is the following:

	k judges	1 judge	k judges
X :	$20, \ldots, 20,$	$20,$	$0, \ldots, 0$
Y :	$20, \ldots, 20,$	$19,$	$19, \ldots, 19$

The majority judgment winner is X, but most observers who look casually at this example opt for Y.[1] The habits of a lifetime immediately suggest that the grades of each candidate should be added, or their averages computed, so Y leads X by a factor of almost 2. Alternatively, if the grades are listed in the order of the judges, the traditional model's point of view says that Y wins with k votes against X with 1 vote because the first k judges are indifferent.

But unless the $2k + 1$ judges are a small number of very discerning experts, almost none see a real difference between grades of 20 and 19—they are all *Excellent*—and the 0s are *To Reject*. So the objection disappears, Y is evaluated *Excellent* by all the voters and is the big winner. If, on the other hand, they are a few expert judges who are able to make the fine distinction between 19 and 20, then the 0s are almost surely manifestations of strategic manipulation. In this extreme example it is then perfectly sensible to let the majority-grades decide.

When an election is in the offing with a large number of voters, then, as has been emphasized repeatedly, the majority judgment *must* offer a language that has only a few grades so that it will be understandable to all participants; otherwise, the language will not be common to all, and the results

1. Roberto Cominetti proposed the example in January 2007 and reported that everyone he asked said Y should be the winner.

become meaningless. When the grades used are those of the presidential election experiments—*Excellent* down to *To Reject*—an alternative example meant to express an objection similar to the preceding example is the following:

	k judges	1 judge	*k* judges
X :	*Excellent*, . . . , *Excellent*,	*Excellent,*	*Poor*, . . . , *Poor*
Y :	*Excellent*, . . . , *Excellent,*	*Good,*	*Good*, . . . , *Good*

The mind-sets of sums, of approval voting, or of the traditional model again suggest *Y* should be the winner, whereas *X* is the majority judgment winner. But a majority believes *X* merits an *Excellent* and that *Y* merits a *Good*. The majority decides on grades ("horizontally") in the new model just as it decides on the order between two competitors ("vertically") in the traditional model. However, the "vertical" view completely ignores the information contained in the intensities of the evaluations. Narrow majorities are accepted routinely in elections (e.g., committees and nations) on the basis of the number of votes; there is no reason not to accept them on the basis of grades. Moreover—who knows?—perhaps some or all of the voters or judges care deeply about the grades themselves.

Example 16.4: Majority Judgment and Centrists The examples that follow [2] are given in the spirit of the new model: only the numbers or percentages of each of the grades have meaning. The eye that sums or averages immediately concludes that *X* should be the winner in every case. And yet, the majority judgment winner is *Y* in every case.

Suppose that in the following example voters were asked, Is each of the candidates at least *Very Good*? A majority would respond yes for candidate *Y* and no for candidate *X*. *Y*'s majority-grade is *Very Good*; *X*'s is only *Good*. Either the idea of a majority evaluation is accepted, or not. Majorities of grades are clearly considerably more discerning decisions than are majorities of preferences. There is no reason that these two very different types of majority decision should concur, nor that either should agree with a point-summing-winner.

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
X :	9%	40%	51%	0	0	0
Y :	4%	47%	0	0	0	49%

2. Those given in percentages are very slight modifications of examples due to Monzoor Ahmad Zahid, communicated by H. de Swart, January 2009.

X is the more centrist, more consensual candidate; Y is the candidate with more confirmed political views—and thus has more high grades and more low grades than X—and he happens to be conferred a higher grade by a *majority*, so he wins with the majority judgment.

A very similar example with a jury of eleven expert judges is the following:

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
X :	1	4	6	0	0	0
Y :	0	6	0	0	0	5

Y's majority-grade is again *Very Good*; X's is only *Good*. X is the competitor who has comfortably high grades all around but has failed to arouse the enthusiasm of a majority, whereas Y has won the enthusiasm of a majority. However, the multitude of 0s suggests that some of the grades may well have been deliberately exaggerated.

In the following example, both candidates have the majority-grade *Very Good*+, but Y's majority-gauge (50%,*Very Good*+, 49%) is higher than X's (1%,*Very Good*+, 0) (the same is true of their majority-values). The "summing-eye" once again prefers X, the centrist. The 1% of voters who assigned *Very Good* to Y can, by changing, give Y any one of the six possible majority-grades and so can decide who is the winner.

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
X :	1%	99%	0	0	0	0
Y :	50%	1%	0	0	0	49%

A very similar example with a jury of eleven expert judges that leads to the same ranking is the following:

	Excellent	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*
X:	1	10	0	0	0	0
Y:	5	2	0	0	0	4

Suppose X and Y are wines; which *should* be ranked higher? It is not at all clear that X should be ranked ahead of Y (as the example is meant to suggest). The majority judgment has crowned Y, perhaps the more daring, the more exceptional, controversial wine; most point-summing methods would make X the winner, but not all. If, for example, *Excellent* awards 10 points, *Very Good* 4 points, *Good* 3 points, and so on down to *To Reject* 0 points, then Y wins. What is *wrong* with these points? Once again, the presence of many 0s suggests

Table 16.1
Number of Wins of the Centrist Candidate (Bayrou), 2007 Orsay Experiment

	Entire Population		Representative Population	
	3 candidates	12 candidates	3 candidates	12 candidates
First-past-the-post	1,848	2,328	47	45
Majority judgment	7,786	7,824	4,073	4,019
Point-summing	9,219	9,231	7,762	7,721
Borda	9,462	9,612	7,120	9,586

Note: "Entire population" = ten thousand samples of 101 ballots, which were drawn from all 1,733 ballots of the 2007 Orsay experiment; "representative population" = ten thousand samples from 101 ballots, which were drawn from a sample of 501 ballots representative of the first-round national vote.

strategic manipulation. And here the one judge that gave *Y* a *Very Good* is able to give *Y* whatever majority-grade she wishes.

The four examples just discussed point to another property: the tendency of the "summing-eye" to favor centrist political candidates or middling competitors—competitors that are judged highly by all judges because they have few faults—versus candidates that are more confirmed on the right or the left of the political spectrum or are exceptional competitors—competitors that dare, fail on some points, yet soar so high as to overcome whatever their other defects. In these examples the point-summing view favors the centrists and middling competitors much more than does the majority judgment. Of course, majority judgment in turn favors centrist candidates much more than first-past-the-post. These tendencies have been proven experimentally (and are discussed again in the subsequent chapters).

Table 16.1 extracts the relevant information. The point-summing method assigned a 5 to *Excellent* down to a 0 to *To Reject*. The point-summing method significantly favors the centrist in comparison with the majority judgment. Observe that elections are almost never conducted using Borda's method. Why? One valid reason may well be that there is an aversion to a method that almost always elects the middling competitor and the centrist candidate, though the standard objection is its ease of manipulation.

16.2 No-Show Objections

The second type of "lump" is generic to a family of essentially equivalent phenomena that bear different names in the literature of the traditional theory of

social choice:[3] the failure to satisfy participant-consistency (or the no-show paradox), join-consistency, and proper cancellation. These ideas have their counterparts in the new model.

Briefly, the *no-show paradox* in terms of the new model is this. A jury decides that candidate X is ranked higher than another candidate Y. Then the jury is augmented by one additional judge who assigns a higher grade to X than to Y. Result: Y is ranked ahead of X. Suggested moral: the judge would have done better not to participate. A method is *participant-consistent* if it avoids the no-show paradox.

A method is *join-consistent* when X is ranked above Y in each of two separate electorates and this implies that X is ranked above Y in the combined electorate. When one of the separate electorates consists of one voter, join-consistency is identical to participant-consistency.

A method *cancels properly* when X is ranked higher than Y and a judge or voter gives both of them the same grade and then omitting (equivalently, adjoining) that judge or voter implies that X remains ranked ahead of Y.

One example suffices to show that the majority judgment fails to satisfy any one of these three properties. The argument in defense of the majority judgment in view of these "defects" is twofold. First, the phenomena are of little real importance. Second, those who insist that any one of them must be avoided have no choice other than to use a point-summing method (see chapter 17), a fate that embraces serious defects.

Example 16.5 This example may be used to show that none of the three properties are satisfied by the majority judgment.

X : 20, 17, 15, 15, 12, 11, 7
Y : 18, 17, 16, 14, 13, 10, 5

X is the majority judgment winner with a majority-grade of 15 to Y's 14. It is not participant-consistent, for suppose an eighth judge gives a 6 to X and a 4 to Y; then X's majority-grade becomes 12 and Y's 13, so Y is the winner. It is not join-consistent for the same reason, since X wins in the seven-judge electorate and in the one-judge electorate yet does not win in the entire eight-judge electorate. It does not cancel properly, for suppose the eighth judge gives a 5 to X and to Y; then again Y becomes the winner.

3. The no-show paradox in the majority judgment was first noticed by Jerôme Renault in October 2006. That the majority judgment is neither join-consistent nor cancels properly was pointed out by the authors in 2007, see Balinski and Laraki (2010). The argument given in this section was summarized in that working paper. Others who have noticed one or another facet of this "lump" are W. Smith in RangeVoting.org (2007), Felsenthal and Machover (2008) and Zahid (2009).

The same phenomena occur in this example on the high side. If the eighth judge gives X a 19 and to Y an 18, or if he gives them both a 17, then Y wins.

However, who knows what are the eighth judge's intentions (or his utility function)? If he gives them both low grades, then he seems to have a relatively low opinion of both candidates, may not care much about which of the two wins, and be very pleased to see their grades lowered (X from 15 to 12 and Y from 14 to 13). If he gives them both high grades, he seems to have a relatively high opinion of both candidates, may not care much about which of the two wins, and be pleased to see their grades raised. But if he felt strongly about preferring X to Y and gave X a grade between 20 and 15 and Y a grade between 0 and 14, then X remains the winner.

Lemma *If X with majority-grade α is the winner against Y with majority-grade β, and a new judge assigns α or a higher grade to X and strictly lower than α to Y, or symmetrically, assigns β or lower to Y and strictly higher than β to X, then X remains the winner.*

All of this suggests that the no-show paradox is not of much importance.

The violation of proper cancellation is a positive property, not a negative one. The majority judgment gives to every voter the possibility of altering the ranking, whether she is indifferent between several or all candidates, or not (though giving two 19s is a very different "indifference" than giving two 2s, and either of these "indifferences" may lead to grades the voter prefers). This is a clear inducement to participate. It is not true of point-summing methods that do cancel properly.

Example 16.6: Participant-Inconsistency A rendition of example 16.5 when there are only a few grades is the following:

X:	*Excellent*,	*Excellent*,	*Very Good,*	*Very Good,*	*Acceptable,*	*Acceptable,*	*Poor*
Y:	*Excellent*,	*Excellent,*	*Excellent,*	*Good,*	*Good,*	*Acceptable,*	*Poor*

X is the majority judgment winner with a majority-grade of *Very Good* to Y's *Good*. An eighth judge gives X the grade *Acceptable* and Y *Poor*, or X *Poor* and Y *To Reject*, or both *Poor*. In each case X's majority-grade becomes *Acceptable* and Y's *Good*. Does this judge really care that the order has been reversed? He certainly showed no enthusiasm for either X or Y. Or is he more interested in simply seeing their majority-grades low? Notice that for the reversal to take place at all the grade just below X's *Very Good* must skip a grade and be strictly below *Good*. Or, the eighth judge goes on the high side and gives to both an *Excellent*, and makes Y the winner, but in this case the grade just above Y's *Good* must skip a grade and be strictly above *Very Good*.

What is to be concluded? Suppose this is a wine competition. If the eighth judge really believes X is superior to Y, then it is not reasonable (or very rare) for that judge to bestow *Excellent* to both X and Y or to bestow *Acceptable*, a grade well below the median, to X. So suppose this is an election, and there are many voters. Then one voter's choice will almost certainly not make a difference, but many with the same grades could. X's majority-grade is *Very Good*; to lower it implies that voters who wish X to win would have to give X a grade lower than *Very Good*. This is very rare. In the Orsay experiment, for example, 89% of the voters gave at least *Very Good* to their preferred candidate (and 98% gave at least *Good*). In practice, the no-show paradox is simply not important.

Example 16.7: Join-Inconsistency Join-consistency is not an applicable property for small juries: it makes no sense to cut a jury of five or nine into subjuries and then ask for the property to hold.

The following is an instance of join-inconsistency with many voters. The majority judgment elects Y with majority-grade *Very Good*$-$ in the 51-voter electorate 1 to X's *Good*$+$.

Electorate 1	*Excellent*	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	Majority-Gauge
X:	9	15	17	9	1	0	(24, *Good*$+$, 10)
Y:	11	15	9	2	3	11	(11, *Very Good*$-$, 25)

It elects Y again in the 124-voter electorate 2, though narrowly, for both candidates have the majority-grade *Good*$-$, but X's grades worse than *Good* are more numerous than Y's.

Electorate 2	*Excellent*	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	Majority-Gauge
X:	20	36	8	56	4	0	(56, *Good*$-$, 60)
Y:	28	16	24	0	5	51	(44, *Good*$-$, 56)

However, in the combined 175-voter electorates 1 and 2, X is the winner with a majority-grade of *Good*$+$ to Y's *Good*$-$.

Electorate 1 and 2	*Excellent*	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	Majority-Gauge
X:	29	51	25	65	5	0	(80, *Good*$+$, 70)
Y:	39	31	33	2	8	62	(70, *Good*$-$, 72)

So Y is the winner in each of the electorates 1 and 2 but not in the combined electorate.

Why should Y be the winner in the combined electorate? The two electorates may have agreed on the rankings, but they certainly did not agree on the evaluations. Notice that in electorate 2, X has many more grades above *Good* than does Y. Why should agreement on rankings be more important than the evaluations? The evaluations are much more discerning than mere rankings.

Join-inconsistency is real but rare. Several simple theorems together with experimental evidence explain why. A social grading function (SGF) f defined for any number of voters, is *grade-join-consistent* if

$$f(\alpha_1, \ldots, \alpha_n) \geq \gamma \quad \text{and} \quad f(\beta_1, \ldots, \beta_k) \geq \gamma \quad \text{implies}$$
$$f(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_k) \geq \gamma,$$

and the same holds if the inequalities are reversed, or are strict, or are equations.

Theorem 16.1 *The majority-grade and the modified majority-grade are grade-join-consistent.*

Proof The hypotheses (with $\geq$) say that an absolute majority of α's and an absolute majority of β's are γ or above, so an absolute majority of α's and β's are γ or above as well (and, of course, the same observation holds for $\leq$, $<$, $>$, or $=$), showing that the majority-grade is grade-join-consistent.

Now, observe that when a candidate X's majority-gauges in two separate electorates are (p_1, α, q_1) and (p_2, α, q_2), so that they have the same majority-grade α, then X's majority-gauge in the combined electorate is $(p_1 + p_2, \alpha, q_1 + q_2)$. This fact immediately implies that the modified majority-grade is grade-joint-consistent, too. ■

The theorem says that when two electorates are in agreement on their evaluations of a candidate with the majority judgment, then the combined electorate evaluates the candidate identically. In the new model the important point is to assure a consistency in evaluations, not a consistency in orders. The majority judgment does.

Two observations are now immediately evident.

Corollary 1 *When a candidate X wins in two electorates with a majority-grade of at least α and all other candidates have majority-grades that are strictly lower than α, then X wins in the combined electorate.*

Corollary 2 *When a candidate X wins against Y in two electorates with a same modified majority-grade and Y also has a same modified majority-grade in both electorates, then X wins in the combined electorate.*

Table 16.2
Number of Wins with Modified Majority-Grades, 2007 Orsay, Experiment

	Entire Population			Representative Population		
	Good−	*Good*+	*Very Good*−	*Good*−	*Good*+	*Very Good*−
Royal	1	734	6	25	581	0
Bayrou	32	8,160	638	789	3,537	0
Sarkozy	0	413	6	52	4,901	112
Total	33	9,307	650	866	9,019	112

Note: "Entire population" = ten thousand samples of 201 ballots, which were drawn from all 1,733 ballots of the 2007 Orsay experiment (ten cases ended in ties); "representative population" = ten thousand samples from 201 ballots, which were drawn from a sample of 501 ballots representative of the first-round national vote (three cases ended in ties).

The mathematics shows that the underlying spirit of join-consistency is met by the majority judgment: there is consistency in evaluations. It also shows that in most cases formal join-consistency is satisfied, and, of course, a situation for which it cannot be proven to hold does not imply that it is violated.

What happens in practice? We conducted three separate experiments during the 2008 U.S. presidential election. In each, Barack Obama emerged with a majority-grade of *Very Good*, the other candidates with strictly lower majority-grades. Obama would forcibly be the winner in the "combined electorate" of the experiments. This is a situation where one candidate stands out well ahead of the others everywhere.

The 2007 Orsay experiment is more instructive. Recall that in all the experiments, whatever the method used, the winner was always one of the three principal candidates, Bayrou, Royal, or Sarkozy. Table 16.2 gives the number of times the candidates won with each of the winning modified majority-grades in two statistical experiments where all twelve candidates are present. In each experiment 201 ballots were drawn randomly 10,000 times. In the first, they were drawn from all 1,733 ballots of the Orsay experiment; in the second, they were drawn from 501 ballots representative of all of France.

In the entire population, the winner's modified majority-grade was *Good*+ 93% of the time, and most of time the winner was Bayrou. In the representative population, the winner's majority-grade was *Good* 99% of the time and his modified majority-grade was *Good*+ 90% of the time. In this second case, the winners were often different, so join-consistency is moot. Recall that the majority judgment winner was not the same in the three actual voting precincts of Orsay, so the property was moot there as well. This suggests that the circumstances under which it is possible for join-consistency to be violated are rare (though more experimentation is necessary).

16.3 Conclusion

It is a relatively simple matter to invent extreme examples that seem, at first glance, to make *any* method look bad. Example 16.1 is an instance: anybody who sees it agrees that *Y* should win, but the traditional model concludes otherwise. To be useful, examples must show how one or another important property is violated, for in that case the logical consequences of the property may be investigated. Charles Darwin reminisced, "I had, during many years, followed a golden rule, namely, that whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without fail and at once . . . Owing to this habit, very few objections were raised against my views which I had not at least noticed and attempted to answer" (1958, 123). We have tried to follow Darwin's example by exploring the consequences of the properties that the majority judgment does not satisfy. Examples of the type "everyone would agree that *X* should be elected instead of *Y*—for instance, of the "summing-eye" type—cannot be the basis on which methods are judged and compared: many disappear when placed in a specific practical context and, in any case, many are extreme.

What paradoxes have actually been observed in practice? Arrow's, in elections (e.g., Bush 2000, Chirac 2002) and in competitions (e.g., skating). A mechanism must satisfy independence of irrelevant alternatives. In the traditional model this leaves Condorcet's pair-by-pair majority rule as the only possible rule (see Dasgupta and Maskin 2008). But it is not necessarily transitive in all domains. Except for political elections it is not possible to restrict the domain to obtain transitivity. In political elections prior information allows manipulation and leads to intransitive outputs (see chapter 19). A mechanism must give transitive outputs in all situations. Conclusion: Individual rankings must be ignored; only methods based on distributions of grades may be considered (theorem 9.2). Therefore, one must accept that the winner between two candidates is not necessarily preferred by a majority of voters (theorem 20.7). What is left? Majority judgment and point-summing methods (thus also approval voting).

The three consistency properties that the majority judgment does not necessarily satisfy—participant-consistency, join-consistency, and proper cancellation—are precisely the three properties that *every* rule that is consistent with Condorcet in the traditional model does not satisfy (see chapter 4). Yet almost all the voting rules that have been advanced in past and recent times are either consistent with Condorcet or advance claims that they usually are. Each of these properties is shown to uniquely characterize point-summing methods

in the new model (see chapter 17). Thus those who ardently insist on any one of them must accept all the very bad properties of point-summing methods.

Those who defend the traditional model and criticize the new one have studiously avoided any discussion of the various monotonicity properties. And yet, they are crucial. It is inconceivable to accept a method that penalizes a candidate when voters change their minds and think more highly of her. If a candidate moves up in the estimation of the voters, then she should not lose in the final standings (choice-monotonicity). If voters' estimations remain the same except that the winner moves up, then not only should she still be the winner but the final ranking among all the others should remain the same (rank-monotonicity). Some methods of the traditional model satisfy one or the other of these two properties, but none satisfy both (theorem 4.4). And no method in the traditional model guarantees that when a nonwinner falls in the estimation of the voters, the winning candidate remains the winner (strong monotonicity, chapter 5). The majority judgment is at once choice-monotonic, rank-monotonic and strongly monotonic.

The majority judgment claims to be *a practical method for electing and ranking*. The rebuttals to the supposed "lumps" of the majority judgment were given without invoking a fact of life: judges and voters may behave strategically. When the possibility of strategic behavior is invoked, the "lumps" become even less significant (see chapter 20).

17 Point-Summing Methods

Fie, fie, my brother!
Weigh you the worth and honour of a king,
So great as our dread father in a scale
Of common ounces? Will you with counters sum
The past-proportion of his infinite?
And buckle in a waist most fathomless
With spans and inches so diminutive
As fears and reasons? Fie, for godly shame!
—William Shakespeare

Point-summing methods are much used in practice: a set of numerical points is specified, a voter or judge assigns any point of the set to each candidate, the winner is the candidate whose sum of points or average is highest, and the candidates are ordered according to the sums of points they receive or their averages. Point-summing methods are *not* sum-scoring methods. In the latter, scores or points are associated with places in a voter's rank-order of the candidates; in the former, voters are free to assign any point of the allowable set of points to a candidate. Approval voting in its traditional presentation is a point-summing method with points 0 and 1. In fact, using *any* two different rational numbers $r < s$ is equivalent if voters systematically use r in place of 0 and s in place of 1 (which is doubtful, for if approval voting instructions said approval counts $+1$ and disapproval -1, the behavior would be quite different than with the usual 1 and 0). Bloggers of various nations and languages have inundated the Web with point-summing proposals. A French site advocates "le vote de valeur," which uses five points: 2 "very favorable," 1 "favorable," 0 "neutral," -1 "hostile," and -2 "very hostile" (Le vote de valeur 2007).[1] A U.S. site argues for "range voting" using one hundred points, 0–99, which are given no definition;

1. Before the 2007 Orsay experiment this site gave no definitions as to the meanings of the numbers: one of us suggested that without definitions the numbers were meaningless.

or alternatively, for "single-digit voting," which uses ten points, 0–9, again not defined (RangeVoting.org 2007). Some French bloggers suggested that the six-grade language used in the Orsay experiment was fine but that it would be simpler to associate a 6 with *Excellent* down to a 1 with *To Reject* and add the numbers to determine the winner and the order of finish. An electoral experiment of 2007 used the points 0, 1, and 2.[2]

The principal attraction of point-summing methods over the traditional methods or approval voting is that they permit voters and judges finer scales of distinction by which to distinguish candidates and competitors. This chapter characterizes these methods in several different ways to show that if certain conditions are imposed, they are the only possible methods. The results should not be seen as arguments in favor of point-summing methods. Quite the contrary. They are ill-conceived for several reasons, some apparent, some more subtle. They are all highly manipulable; the wider the gap between the minimum and maximum of the points, the more they are manipulable. When the numbers have no definition, the language of grading may well not be common, so Arrow's impossibility theorem applies. When the numbers have definitions, the definitions may be so formulated that they nonetheless induce voters to make relative comparisons rather than absolute evaluations, so the method may suffer from Arrow's paradox. More fundamentally, once again, summing numbers or, equivalently, taking their average, is meaningless—unless they are drawn from a bona fide interval scale, as shown by the theory of measurement—so has no justification whatsoever. There is no doubt that point-summing methods are to be shunned.

17.1 Point-Summing Methods: Theory

General Point-Summing Methods

Inputs are profiles Φ of grades of a language Λ, assigned by any number of judges to any number of candidates. Unless specified otherwise, the language is assumed to be finite. Recall that a social ranking function (SRF) is a transitive binary relation $\succeq_S$ that respects ties and grades (meaning that only the grades count, not which judge assigned which grade; see chapter 9). In this analysis the numbers of judges may vary, so it is important to keep in mind that $\succeq_S$ only compares candidates who have sets of the same number of grades.

2. They refer to Hillinger (2004). In fact, Hillinger advocates cardinal utility inputs on a scale of $-1, 0, +1$ and a point-summing method. He correctly recognizes the need to abandon the traditional model but confuses the evaluation of merit with the evaluation of utility.

A social ranking function $\succeq_S$ is a *lexicographic point-summing method* if there exist functions $\psi^j : \Lambda \to \mathbf{R}$, $j = 1, \ldots, k$ for which

$\alpha \succ_S \beta$ if and only if

$$\left(\sum_i \psi^1(\alpha_i), \ldots, \sum_i \psi^k(\alpha_i)\right) >_{lex} \left(\sum_i \psi^1(\beta_i), \ldots, \sum_i \psi^k(\beta_i)\right),$$

or equivalently, if and only if

$$\sum_i \big(\psi^1(\alpha_i), \ldots, \psi^k(\alpha_i)\big) >_{lex} \sum_i \big(\psi^1(\beta_i), \ldots, \psi^k(\beta_i)\big),$$

where $>_{lex}$ means lexicographically higher and the sums are taken over the judges. For such methods to be choice-monotonic it is necessary and sufficient that when $\alpha_0 \succ \beta_0$ are any two single grades,

$$\big(\psi^1(\alpha_0), \ldots, \psi^k(\alpha_0)\big) >_{lex} \big(\psi^1(\beta_0), \ldots, \psi^k(\beta_0)\big).$$

A lexicographic point-summing method is a *point-summing method* if there exists a single function $\psi : \Lambda \to \mathbf{R}$ for which

$$\alpha \succ_S \beta \quad \text{if and only if} \quad \sum_i \psi(\alpha_i) > \sum_i \psi(\beta_i).$$

Thus a lexicographic point-summing method is simply a point-summing method together with point-summing tie-breaking rules. Practically speaking, with many voters, there are no ties.

A social ranking function is *join-consistent* if

$$\alpha \succ_S \beta \quad \text{and} \quad \gamma \succeq_S \delta \quad \text{implies} \quad (\alpha, \gamma) \succ_S (\beta, \delta),$$

and

$$\alpha \approx_S \beta \quad \text{and} \quad \gamma \approx_S \delta \quad \text{implies} \quad (\alpha, \gamma) \approx_S (\beta, \delta).$$

This notation assumes that α and β have the same number of grades (say, n), and that γ and δ also do (say, k); and when $n = 1$, $\alpha \succ_S \beta$ means the individual's input. This implies that if a set of voters is indifferent between two candidates, then an additional individual's input is society's output.

Theorem 17.1 *A social ranking function is join-consistent if and only if it is a choice-monotonic lexicographic point-summing method.*

Proof.[3] Suppose the finite language consists of k grades and

$$\alpha = (\overbrace{\alpha_1, \ldots, \alpha_1}^{n_1}, \ldots \overbrace{\alpha_k, \ldots, \alpha_k}^{n_k}) \quad \text{and} \quad \beta = (\overbrace{\alpha_1, \ldots, \alpha_1}^{m_1}, \ldots \overbrace{\alpha_k, \ldots, \alpha_k}^{m_k}),$$

where there are $n = \sum n_i = \sum m_i$ judges. Take Δ_k to be the simplex of dimension $k-1$, and $\Delta_k^{\mathbf{Q}} = \Delta_k \cap \mathbf{Q}^k$ to be the simplex of rational numbers. For the α and β just given, define the binary relation $\succ$ on $\Delta_k^{\mathbf{Q}}$ by

$$\left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right) \succ \left(\frac{m_1}{n}, \ldots, \frac{m_k}{n}\right) \quad \text{if and only if} \quad \alpha \succ_S \beta,$$

and $\approx$ on $\Delta_k^{\mathbf{Q}}$ by

$$\left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right) \approx \left(\frac{m_1}{n}, \ldots, \frac{m_k}{n}\right) \quad \text{if and only if} \quad \alpha \approx_S \beta.$$

Complete the definitions by taking $\succeq$ to mean either $\succ$ or $\approx$. Join-consistency assures that these preference relations are well-defined for any pair in $\Delta_k^{\mathbf{Q}}$.

In effect, $x \in \Delta_k^{\mathbf{Q}}$ is simply a representation of the set of grades α normalized to sum to 1.

Consider now $D = \Delta_k^{\mathbf{Q}} \times \Delta_k^{\mathbf{Q}}$. Let

$$D_1 = \{(x, y) \in D : x \succ y\} \quad \text{and} \quad D_2 = \{(x, y) \in D : y \succ x\}.$$

They are symmetric: $(x, y) \in D_1$ if and only if $(y, x) \in D_2$. Join-consistency implies that D_i (for $i = 1$ or 2) are $\mathbf{Q}$-convex, meaning that for any rational $\lambda \in [0, 1]$, and any (x, y) and (x', y') in D_i, $\lambda(x, y) + (1-\lambda)(x', y')$ is also in D_i.

The essence of the remainder of the proof shows that the closures of the $\mathbf{Q}$-convex sets D_1 and D_2 may be separated by a hyperplane that defines ψ^1 of the lexicographic point-summing method.

Suppose $\overline{D} = \Delta_k \times \Delta_k$, the closure of D, is not equal to $\overline{D_1} \cup \overline{D_2}$. Then their difference D_0 must be an open set in $\overline{D}$. Take an $(x_0, y_0) \in D_0$ whose values are rational. Then $x_0 \approx y_0$. Now take $x_1 \succ y_1$. By join-consistency, $\lambda x_0 + (1-\lambda)x_1 \succ \lambda y_0 + (1-\lambda)y_1$ for any $0 < \lambda < 1$. But this is a contradiction: D_0 is an open set, so for λ rational and close enough to 1, the convex combination belongs to D_0. Therefore, $\overline{D} = \overline{D_1} \cup \overline{D_2}$.

$\overline{D}$ is the union of two symmetric convex sets $\overline{D_1}$ and $\overline{D_2}$, so all three have the same dimension, and so have nonempty interiors. Moreover, $\overline{D_1} \cap \overline{D_2}$ cannot

3. The proof is inspired by that given in Young (1974), characterizing sum-scoring methods in the traditional model.

contain an open set, for then it would contain rational points (x, y). In fact, as will be shown, they belong to both D_1 and D_2, which is a contradiction.

To see that (x, y) belongs to D_1, note first that $ri(cvx D_1) = ri(cvx \overline{D_1}) = ri(\overline{D_1})$, where ri is the relative interior and cvx the convex hull operators. It has been shown that D_1 **Q**-convex implies $D_1 = Q^{2k} \cap cvx D_1$ (see Young 1975, lemma 1). Therefore,

$$(x, y) \in Q^{2k} \cap ri(\overline{D_1}) = Q^{2k} \cap ri(cvx D_1) \subset D_1,$$

so $(x, y) \in D_1$. The same argument applies to D_2.

As a consequence, $ri(\overline{D_1}) \cap ri(\overline{D_2}) = \emptyset$, and so $\overline{D_1}$ and $\overline{D_2}$ may be properly separated by a hyperplane: there exists $\psi = (\psi_1, \ldots, \psi_k)$ and $\psi' = (\psi'_1, \ldots, \psi'_k)$, both in $\mathbf{R}^k$ and $\phi \in \mathbf{R}$ such that $(x, y) \in \overline{D_1}$ if and only if $\sum_i \psi_i x_i \geq \phi + \sum_i \psi'_i y_i$, and symmetrically, $(x, y) \in \overline{D_2}$ if and only if $\sum_i \psi_i x_i \leq \phi + \sum_i \psi'_i y_i$. Since $\sum y_i = 1$, the ϕ may be absorbed into the ψ'_i, which comes to the same thing as taking $\phi = 0$.

What does this say about D? Take any (necessarily rational) $(x, y) \in D$ for which $\sum_i \psi_i x_i > \sum_i \psi'_i y_i$. The point (x, y) is in the relative interior $ri(\overline{D_1})$ implying, as above, that it belongs to D_1, so $x \succ y$. Symmetrically, $\sum_i \psi_i x_i < \sum_i \psi'_i y_i$ for (x, y) rational implies $y \succ x$.

Now note that whenever (x, y) is rational and $x = y$, then since $x \approx y$, $\sum_i \psi_i x_i = \sum_i \psi'_i y_i$. Taking different $x = y = (0, \ldots, 1, \ldots, 0)$ with one 1 in the ith position implies $\psi_i = \psi'_i$ for $i = 1, \ldots, k$, so $\psi = \psi'$.

If $D^2 = \{(x, y) \in D : \sum_i \psi_i x_i = \sum_i \psi_i y_i\}$ contains no point such that $x \succ y$, the proof ends here. The method is a point-summing method. Otherwise, rename ψ to be ψ^1. The dimension of $\overline{D^2}$ is strictly smaller than that of $\overline{D}$. The same reasoning implies the existence of ψ^2 such that when $(x, y) \in D^2$, $\sum \psi_i^2 x_i > \sum \psi_i^2 y_i$ implies $x \succ y$ and $\sum \psi_i^2 x_i < \sum \psi_i^2 y_i$ implies $y \succ x$. If $D^3 = \{(x, y) \in D^2 : \sum_i \psi_i^2 x_i = \sum_i \psi_i^2 y_i\}$ contains no point such that $x \succ y$, the proof ends; otherwise, it ends in at most $2k$ steps because each successive D^j is of smaller dimension. Thus the social ranking function must be a lexicographic point-summing method.

To see that it must be choice-monotonic, consider any list of grades $\alpha = (\alpha_1, \ldots, \alpha_n)$, and two single grades $\alpha_0 \succ \beta_0$. Join-consistency implies $(\alpha, \alpha_0) \succ_S (\alpha, \beta_0)$. Therefore,

$$(\psi^1(\alpha_0), \ldots, \psi^k(\alpha_0)) + \textstyle\sum_i (\psi^1(\alpha_i), \ldots, \psi^k(\alpha_i)) \\ >_{lex} (\psi^1(\beta_0), \ldots, \psi^k(\beta_0)) + \sum_i (\psi^1(\alpha_i), \ldots, \psi^k(\alpha_i)),$$

or

$$\left(\psi^1(\alpha_0), \ldots, \psi^k(\alpha_0)\right) >_{lex} \left(\psi^1(\beta_0), \ldots, \psi^k(\beta_0)\right),$$

so the method must be choice-monotonic.

That any choice-monotonic lexicographic point-summing method is join-consistent, transitive, and respects ties and grades is at once obvious, completing the proof. ∎

A social ranking function $\succeq_S$ *respects large electorates* when α and β are any sets of the same number of grades, and $\gamma \succ_S \delta$ (so also have a same number of grades), there is a large enough integer L for which

$$(\alpha, \overbrace{\gamma, \ldots, \gamma}^{L}) \succ_S (\beta, \overbrace{\delta, \ldots, \delta}^{L}).$$

If $x_\theta \in \Delta_k^{\mathbf{Q}}$ is the representation of the set of grades θ, this definition says that for any x_α and x_β that represent sets of equal numbers of grades and $x_\gamma \succ x_\delta$ there is a large enough integer L for which

$$\frac{x_\alpha + Lx_\gamma}{L+1} \succ \frac{x_\beta + Lx_\delta}{L+1}.$$

Theorem 17.2 *A join-consistent social ranking function respects large electorates if and only if it is a choice-monotonic point-summing method. It is defined uniquely up to a positive affine transformation.*

Proof The proof of the previous theorem ends with one ψ—thus with a point-summing method—when
$(x', y') \in D' = \left\{(x, y) \in D : \sum_i \psi_i x_i = \sum_i \psi_i y_i\right\}$ implies $x' \approx y'$.

So suppose, first, that $(x', y') \in D'$ but $x' \prec y'$. Take any rational $(x, y) \in ri(\overline{D}_1)$, so that $x \succ y$. Then for any $n > 0$,

$$\frac{1}{n}(x, y) + \frac{n-1}{n}(x', y') \in ri(\overline{D}_1),$$

meaning

$$\frac{x + (n-1)x'}{n} \succ \frac{y + (n-1)y'}{n} \quad \text{for any } n > 0.$$

But this does not respect large electorates, for $x' \prec y'$ with n large enough implies the preference must go in the opposite direction.

So suppose that $(x', y') \in D'$ but $x' \succ y'$. Then take any $(x, y) \in ri(\overline{D}_2)$, and repeat the same argument to find the symmetric contradiction, and so conclude that $(x', y') \in D'$ implies $x' \approx y'$.

ψ is unique up to a positive linear (e.g., affine) transformation, for suppose there were two point-summing methods, ψ and ϕ. Take any $(x, y) \in \Delta_k \times \Delta_k$.

Then

$$\sum \psi_i x_i \geq \sum \psi_i y_i \quad \text{if and only if} \quad \sum \phi_i x_i \geq \sum \phi_i y_i,$$

and a standard algebraic argument (such as that used in von Neumann and Morgenstern 1944 in establishing the expected utility theorem) proves the claim.

That any choice-monotonic point-summing method is join-consistent, transitive, and respects large electorates, ties, and grades is obvious, completing the proof. ∎

Since the sum is always taken over a finite number of judges or voters, a finite number of grades is used in every comparison. This permits the previous theorem to be generalized for any language, finite or infinite, denumerable or not, measurable or not.

Theorem 17.3 *Whatever the language* Λ*, a join-consistent social ranking function respects large electorates if and only if it is a choice-monotonic point-summing method defined uniquely up to a positive affine transformation.*

Proof Choose two grades $\alpha_o, \beta_o \in \Lambda$ for which a candidate assigned only α_o's is ranked strictly above a candidate assigned only β_o's. The function $\psi : \Lambda \to \mathbf{R}$ is constructed as follows.

$\psi(\alpha_o) = 1$ and $\psi(\beta_o) = 0$. Take any $\gamma_o \in \Lambda$. It is possible that judges use only the three grades $\alpha_o, \beta_o, \gamma_o$. By the previous theorem, there exists a unique value $\psi(\gamma_o)$ for the point-summing method restricted to these three grades. Define ψ for any grade $\gamma_o \in \Lambda$ in this manner.

Now consider a comparison between two candidates in which a finite number of grades Λ_0 are used. The last theorem implies there exists a unique point-summing method ϕ defined on the finite set of grades $\{\alpha_o, \beta_o\} \cup \Lambda_0$ for which $\phi(\alpha_o) = 1$ and $\phi(\beta_o) = 0$. Take any $\gamma_o \in \Lambda_0$. Since only the three grades $\alpha_o, \beta_o, \gamma_o$ might have been used, $\phi(\gamma_o)$ must be identical to $\psi(\gamma_o)$, completing the proof. ∎

A social ranking function $\succeq_S$ *cancels properly* if when γ_o is a single grade and α and β are sets of the same number of grades,

$$\alpha \succ_S \beta \quad \text{implies} \quad (\alpha, \gamma_o) \succ_S (\beta, \gamma_o), \quad \text{and} \quad \alpha \approx_S \beta \quad \text{implies} \quad (\alpha, \gamma_o) \approx_S (\beta, \gamma_o).$$

Thus join-consistency immediately implies proper cancellation. In fact, they are equivalent.

Theorem 17.4 *A social ranking function cancels properly if and only if it is a choice-monotonic lexicographic point-summing method. If the method respects large electorates, then it is a choice-monotonic point-summing method.*

Proof That a lexicographic point-summing method cancels properly is obvious. The proof shows that the hypotheses imply join-consistency, so the result follows from the last theorems.

Suppose α and β have the same number of grades (say, n), γ and δ also have the same number of grades (say, k). If $\alpha \succ_S \beta$ and $\gamma \succeq_S \delta$, then, applying the cancellation property one γ_i at a time,

$$(\alpha, \gamma) \succ_S (\beta, \gamma),$$

and one β_j at a time

$$(\beta, \gamma) \succeq_S (\beta, \delta),$$

conclude that

$$(\alpha, \gamma) \succ_S (\beta, \delta).$$

Suppose then that $\alpha \approx_S \beta$ and $\gamma \approx_S \delta$. The same argument shows that $(\alpha, \gamma) \approx_S (\beta, \delta)$, so join-consistency holds. ∎

A social ranking function $\succeq_S$ is *participant-consistent* if, when γ_o and δ_o are single grades and α and β are sets of the same number of grades,

$$\alpha \succ_S \beta \quad \text{and} \quad \gamma_o \succ \delta_o \quad \text{implies} \quad (\alpha, \gamma_o) \succ_S (\beta, \delta_o), \quad \text{and} \quad \alpha \approx_S \beta \text{ implies } (\alpha, \gamma_o) \approx_S (\beta, \gamma_o).$$

Participant-consistency is a very weakened form of join-consistency. It is not a sufficiently strong property to enable simple characterizations. However, a conclusion similar to the last two theorems may be reached by invoking another property.

A social ranking function $\succeq_S$ is *step-continuous* if $\alpha \succ_S \beta$ implies that any grade of α (not equal to the minimum grade of Λ) may be lowered to obtain α^- satisfying $\alpha^- \succeq_S \beta$, and that any grade of β (other than the maximum grade of Λ) may be raised to obtain β^+ satisfying $\alpha \succeq \beta^+$. All order functions are step-continuous, as is the majority-ranking; when the language is continuous (e.g., $[0, R]$), so are the point-summing methods.

Theorem 17.5 *A choice-monotonic, step-continuous social ranking function is participant-consistent if and only if it is a choice-monotonic lexicographic point-summing method. Moreover, if the method respects large electorates, then it is a choice-monotonic point-summing method.*

Proof Again, it is clear that lexicographic point-summing methods are participant-consistent. The proof shows that the hypotheses imply proper cancellation, so the result follows from the last theorem. It suffices to show that $\alpha \succ_S \beta$ implies $(\alpha, \gamma_o) \succ_S (\beta, \gamma_o)$.

Suppose γ_o is not the maximum grade of Λ. By choice-monotonicity some α_i must be above the minimum grade. Therefore by step-continuity it may be lowered to obtain $\alpha^- \succeq_S \beta$. There are two possibilities: either $\alpha^- \approx_S \beta$ or $\alpha^- \succ_S \beta$.

In the first case, participant-consistency implies $(\alpha^-, \gamma_o) \approx_S (\beta, \gamma_o)$ and choice-monotonicity then implies that $(\alpha, \gamma_o) \succ_S (\beta, \gamma_o)$.

In the second case, suppose $(\alpha^-, \gamma_o) \prec_S (\beta, \gamma_o)$. Step-continuity implies γ_o may be replaced by $\gamma_o^+ \succ \gamma_o$ to obtain $(\alpha^-, \gamma_o^+) \preceq_S (\beta, \gamma_o)$. Since $\alpha^- \succ_S \beta$, this contradicts participant-consistency. Therefore, $(\alpha^-, \gamma_o) \succeq_S (\beta, \gamma_o)$ so, by choice-monotonicity, $(\alpha, \gamma_o) \succ_S (\beta, \gamma_o)$. This finishes the proof unless γ_o is the maximum possible grade. But in that case a grade of β may be raised, and a similar argument completes the proof. That it is a choice-monotonic point-summing method when it respects large electorates follows from theorem 17.2. ∎

Leximax and Leximin Social-Ranking Functions

It was observed (in chapter 16) that the majority-ranking is not participant-consistent. Is *any* order function participant-consistent? The max *order function* is f^1, the highest of the set of grades; the min *order function* (when there are n grades) is f^n, the lowest of the set of grades.

Theorem 17.6 *The max and min are the only order functions that are participant-consistent social ranking functions (when the language has at least three grades).*

Proof The max order function is participant-consistent. To verify this, first suppose α and β are sets of the same number of grades with $\max(\alpha) \succ \max(\beta)$ and that $\gamma_o \succ \delta_o$ are single grades. Either $\delta_o \preceq \max(\beta)$ or not: in both cases $\max(\alpha, \gamma_o) \succ \max(\beta, \delta_o)$. Next, suppose $\max(\alpha) \approx \max(\beta)$. Either $\gamma_o \preceq \max(\beta)$ or not: in both cases $\max(\alpha, \gamma_o) \approx \max(\beta, \gamma_o)$. A similar verification may be given for the min order function.

Suppose that when n grades are assigned, the order function $f^{k(n)}$ is used, and that $k(n+1) \neq 1, n+1$. Take any three different grades; for simplicity of exposition, let them be 0, 1, and 2, from lowest to highest.

Suppose, first, that $1 < k(n+1) \leq k(n) \leq n$. Consider the following two sets of n grades:

$$\left.\begin{aligned}\alpha &= (1, \ldots, 1, \overbrace{1,}^{k(n)} 0, \ldots, 0)\\ \beta &= (1, \ldots, 1, \overbrace{0,}^{k(n)} 0, \ldots, 0)\end{aligned}\right\} \text{ so } f^{k(n)}(\alpha) \succ f^{k(n)}(\beta).$$

Then $f^{k(n+1)}(\alpha, 2) = 1$ and $f^{k(n+1)}(\beta, 1) = 1$, so $(\alpha, 2) \approx (\beta, 1)$, contradicting participant-consistency.

Suppose, next, that $1 \leq k(n) < k(n+1) < n+1$. Consider the following two sets of n grades:

$$\left.\begin{aligned}\alpha &= (2, \ldots, 2, \overbrace{2,}^{k(n)} 1, \ldots, 1)\\ \beta &= (2, \ldots, 2, \overbrace{1,}^{k(n)} 1, \ldots, 1)\end{aligned}\right\} \text{ so } f^{k(n)}(\alpha) \succ f^{k(n)}(\beta).$$

Then $f^{k(n+1)}(\alpha, 1) = 1$ and $f^{k(n+1)}(\beta, 0) = 1$, so $(\alpha, 1) \approx (\beta, 0)$, contradicting participant-consistency.

Participent-consistency must fail if max and min are used for successive sizes of the electorate. To see this, consider the following sets of n grades:

$$\alpha = (2, 0, \ldots, 0)$$
$$\beta = (1, 1, \ldots, 1).$$

If max is used for n grades and min for $n+1$ grades, then $\alpha \succ \beta$, but adjoining a 2 to α and a 1 to β reverses the order, contradicting participant-consistency. If the min is used for n grades and the max for $n+1$ grades, then $\alpha \prec \beta$, but adjoining a 0 to α and a 1 to β reverses the order, too, completing the proof. ■

The max and min are the most manipulable of the order functions because every judge can raise the final grade in the first instance and every judge can lower the final grade in the second instance. It is the middlemost order functions that best resist manipulation.

Observe that max and min are step-continuous but neither choice-monotonic nor join-consistent. They fail to be choice-monotonic because some grade other than the max or min of a candidate may be raised, but this changes nothing in the grade of the candidate. The order function max fails join-consistency because when $\max(\alpha) \succ \max(\beta)$ and $\max(\gamma) = \max(\delta) = \gamma_o$, but $\gamma_o \succ \max(\alpha)$, it is not true that $\max(\alpha, \gamma) \succ \max(\beta, \delta)$ because the two terms are equal to γ_o. And, for a similar reason, min fails as well. This explains why theorem 17.5 requires more conditions than the earlier theorems.

There may, of course, be many candidates with the same max or the same min final grade. Can order functions be catenated into a lexicographic scheme for

breaking ties, as is done with point-summing? Two possibilities immediately suggest themselves. The *leximax*[4] social ranking function first uses the max order function to determine the first final grade α^1 of a set of n grades α; one α^1 is dropped from α, and max determines the second final grade α^2 among the $n-1$ remaining grades; and this is repeated to obtain the ordered set of n grades $(\alpha^1, \ldots, \alpha^n)$ (where $\alpha^i \succeq \alpha^{i+1}$). Given two sets of n grades, α and β,

$$\text{leximax}(\alpha) \succ_S \text{leximax}(\beta) \quad \text{if } (\alpha^1, \ldots, \alpha^n) \succ_{lex} (\beta^1, \ldots, \beta^n),$$

where $\succ_{lex}$ means lexicographically higher in grades. An alternative description, when n grades are assigned, is to specify the levels of the hierarchy in the lexicography: first use f^1, then $f^2, \ldots$, finally f^n. Clearly all the order functions must be used because using a same one yields no additional information.

Similarly, the *leximin* social ranking function first uses the min order function to determine the first final grade α^1 of a set of n grades α; one α^1 is dropped from α, and min determines the second final grade α^2 among the $n-1$ remaining grades; and this is repeated to obtain the ordered set of n grades $(\alpha^1, \ldots, \alpha^n)$ (where $\alpha^i \preceq \alpha^{i+1}$). Given two sets of n grades, α and β,

$$\text{leximin}(\alpha) \succ_S \text{leximin}(\beta) \quad \text{if } (\alpha^1, \ldots, \alpha^n) \succ_{lex} (\beta^1, \ldots, \beta^n).$$

An alternative description is to specify the levels of the hierarchy in the lexicography: first use f^n, then f^{n-1}, ..., finally f^1 (when n grades are assigned). Recall that a *lexi-order social ranking function* associates to each size of the electorate n, a permutation σ_n of the order functions $f^{\sigma_n(1)}, \ldots, f^{\sigma_n(n)}$, and ranks the candidates by

$$\alpha \succ_S \beta \quad \text{if } (f^{\sigma_n(1)}(\alpha), \ldots, f^{\sigma_n(n)}(\alpha)) \succ_{lex} (f^{\sigma_n(1)}(\beta), \ldots, f^{\sigma_n(n)}(\beta)).$$

Theorem 17.7 *Leximax and leximin are the unique meaningful social ranking functions that cancel properly (when the language has at least three grades). They are therefore join-consistent and participant-consistent.*

Proof To begin, recall that proper cancellation implies choice-monotonicity which, together with meaningfulness (or order-consistency; see theorem 11.5b), implies the SRF must be a lexi-order SRF.

If $n = 1$, there is nothing to prove. It is first shown that $\sigma_n(1) = 1$ for all n or $\sigma_n(1) = n$ for all n.

When $n = 2$, either $\sigma_2(1) = 1$ or $\sigma_2(1) = 2$.

If $\sigma_2(1) = 1$, assume inductively that $\sigma_k(1) = 1$ for $k \leq n$, and suppose $\sigma_{n+1}(1) \neq 1$. Consider the two sets of n grades:

4. Leximax and especially leximin are well-known functions in the literature on welfarism (see Hammond 1976 and H. Moulin 1988).

$$\left.\begin{aligned}\alpha &= (2, 0, \ldots, 0)\\ \beta &= (1, 1, \ldots, 1)\end{aligned}\right\} \text{ for which } \alpha \succ_S \beta.$$

But $\sigma_{n+1}(1) \neq 1$ implies $(\alpha, 0) \preceq_S (\beta, 0)$, contradicting the cancellation property. So $\sigma_n(1) = 1$ for all n.

If $\sigma_2(1) = 2$, assume inductively that $\sigma_k(1) = k$ for $k \leq n$, and suppose $\sigma_{n+1}(1) \neq n+1$. Consider the two sets of n grades:

$$\left.\begin{aligned}\alpha &= (2, \ldots, 2, 0)\\ \beta &= (1, \ldots, 1, 1)\end{aligned}\right\} \text{ for which } \alpha \prec_S \beta.$$

But $\sigma_{n+1}(1) \neq n+1$ implies $(\alpha, 2) \succeq_S (\beta, 2)$, again contradicting the cancellation property. So $\sigma_n(1) = n$ for all n.

If $\sigma_n(1) = 1$ for all n, assume inductively that $\sigma_n(j) = j$ when $j \leq k$ for all n. Necessarily $\sigma_{k+1}(k+1) = k+1$ since the first k order functions have already been assigned. So suppose $n+1$ is the smallest integer for which $\sigma_{n+1}(k+1) \neq k+1$. This means that $\sigma_{n+1}(k+1) > k+1$. Consider the two sets of n grades:

$$\left.\begin{aligned}\alpha &= (2, \ldots, 2, \overbrace{2}^{k+1}, 0, \ldots, 0)\\ \beta &= (2, \ldots, 2, \overbrace{1}^{k+1}, 1, \ldots, 1)\end{aligned}\right\} \text{ for which } \alpha \succ_S \beta.$$

But $\sigma_{n+1}(k+1) > k+1$ implies $(\alpha, 0) \preceq_S (\beta, 0)$, so cancellation is once again violated. Therefore $\sigma_n(j) = j$ when $j \leq k+1$ for all n. So induction implies $\sigma_n(j) = j$ for all j and all n; this is the leximax social ranking function. A similar argument shows the leximin social ranking function must be taken when $\sigma_n(1) = n$ for all n, which completes the proof. ∎

In fact, leximax and leximin are choice-monotonic lexicographic point-summing methods. The language of grades Λ is finite; let it be $\lambda^1 \succ \cdots \succ \lambda^k$. Leximax is the lexicographic point-summing method defined by

$$\left(\sum_i \overline{\psi}^1(\alpha_i), \ldots, \sum_i \overline{\psi}^k(\alpha_i)\right) \quad \text{where } \overline{\psi}^j(\alpha_i) = \begin{cases} 1 & \text{if } \alpha_i = \lambda^j,\\ 0 & \text{if otherwise.}\end{cases}$$

Letting $n_j(\alpha)$ be the number of λ^j's in α $\left(\text{so } n = \sum_1^k n_j(\alpha)\right)$, this may be expressed as

$$\left(\sum_i \overline{\psi}^1(\alpha_i), \ldots, \sum_i \overline{\psi}^k(\alpha_i)\right) = (n_1(\alpha), \ldots, n_k(\alpha)).$$

It has in its first place the number of highest grades λ^1, in its second place the number of second highest grades $\lambda^2, \ldots,$ down to the number of lowest grades λ^k, so gives the exact same ordering as leximax.

An equivalent description is

$$\left(\sum_i \psi^1(\alpha_i), \ldots, \sum_i \psi^k(\alpha_i)\right) \quad \text{where } \psi^j(\alpha_i) = \begin{cases} 1 & \text{if } \alpha_i \succeq \lambda^j, \\ 0 & \text{if otherwise,} \end{cases}$$

or

$$\left(\sum_i \psi^1(\alpha_i), \ldots, \sum_i \psi^k(\alpha_i)\right) = \left(n_1(\alpha), n_1(\alpha) + n_2(\alpha), \ldots, \sum_1^k n_j(\alpha)\right).$$

It has in its first place the number of highest grades λ^1, in its second place the number of the two highest grades λ^1 and $\lambda^2, \ldots,$ and in its last place the number of all grades n. Seen in this guise, leximax is a specific type of lexicographic approval voting: an approval is the highest grade, the candidate with the most wins; if there is a tie, then an approval is either of the two highest grades; if there is still a tie, an approval is any of the three highest grades; and so on.

Leximin is the lexicographic point-summing method defined by

$$\left(\sum_i \overline{\psi}^1(\alpha_i), \ldots, \sum_i \overline{\psi}^k(\alpha_i)\right) = (-n_k(\alpha), -n_{k-1}(\alpha), \ldots, -n_1(\alpha)).$$

A more agreeable equivalent description is

$$\left(\sum_i \psi^1(\alpha_i), \ldots, \sum_i \psi^k(\alpha_i)\right) = \left(\sum_1^{k-1} n_j(\alpha), \sum_1^{k-2} n_j(\alpha), \ldots, n_1(\alpha), 0\right),$$

or

$$\left(\sum_i \psi^1(\alpha_i), \ldots, \sum_i \psi^{k-1}(\alpha_i), 0\right) \quad \text{where } \psi^j(\alpha_i) = \begin{cases} 1 & \text{if } \alpha_i \succeq \lambda^{k-j}, \\ 0 & \text{if otherwise.} \end{cases}$$

It has in its first place the number of $k-1$ highest grades $\lambda^1, \ldots, \lambda^{k-1}$, in its second place the number of the $k-2$ highest grades $\lambda^1, \ldots, \lambda^{k-2}, \ldots,$ and in its last place 0. Seen in this guise, leximin is another specific type of lexicographic approval voting: an approval is any grade but the last, the candidate with the most wins; if there is a tie, then an approval is any grade but the last two; if there is still a tie, an approval is any grade but the last three; and so on.

When the language contains only two grades, leximax and leximin are identical: the social ranking function is approval voting (see chapter 18).

17.2 Point-Summing Methods: Practice

The choice of the language of grades to use in the first real test of the majority judgment—the 2007 Orsay experiment—was the subject of long debates. The first, obvious possibility was the usual 0–20 scale of the French educational system. Its meanings are very different than its linear interpolations in the 0–100 scale often used in the United States (see chapters 7 and 8): a 10/20 is a passing grade in France, but a 50/100 is a very clear failing grade in the United States. In France an 18/20 is excellent, a 15/20 is very good, a 12/20 is good. However, for the population at large, the concern was that numbers could well be understood in very different ways; moreover, how would a voter understand the difference between a 16 and a 17, or a 7 and an 8? Twenty different clear and distinct definitions would have to be given, well beyond Miller's magic seven plus or minus two. A scale of words is much more meaningful to voters than a scale of numbers. An additional difficulty with undefined numbers is that they are abstract: voters usually assume they will be summed or averaged, and in any case undefined numbers constitute a clear invitation to manipulation. That summing points is a bad idea in theory is clear. That it is a bad idea in practice emerges from the four following experiments.

One experiment was conducted in December 2007, when the École Polytechnique student government (Kès) elections took place.[5] The École Polytechnique is one of France's three most elite undergraduate institutions.[6] Admission crucially depends on mathematical ability. In these elections, teams or parties are formed and presented as entire lists, and the students vote for one of the lists, and the list with the most votes is elected. There were two serious lists (the others were organized as larks), called Jukesbox and Kesdelweiss (the names must include the syllable "kes"). In parallel with the official vote, the students were asked to (1) give one of the six grades used in the presidential election experiments,[7] and (2) assign numerical grades between *0* and *100* to each of these two lists on one and the same ballot. They were informed that the highest median grade would determine the winner in both cases.

5. The origin of the Kès dates to a decree of the revolutionary year XII (1804) that made the financial situation of students of modest backgrounds precarious. Their better-off fellows established a cashier's office ("caissier," whence the shortened name) that gathered funds and distributed them. Across the years, the Kès came to represent students in dealing with Polytechnique's administration.

6. The others are the École Normale Supérieure and the Institut d'Études Politiques (IEP), often called Sciences Po.

7. *Très Bien (Excellent), Bien (Very Good), Assez Bien (Good), Passable (Acceptable), Insuffisant (Poor), à Rejeter (To Reject).*

The Jukesbox list was the official winner with 244 (54%) votes to Kesdelweiss's 206 (46%) (ignoring the scattered votes for other lists). Of those who voted officially, 228 (roughly half) participated in the experiment and 221 votes were valid. The Jukesbox list was also the winner with the majority judgment with both of the scales of grades, obtaining a majority-gauge of (31%,*Very Good*+, 23%) to Kesdelweiss's (20%,*Very Good*−, 35%) in the first case, and (32%, *80*−, 48%) to its rival's (47%, *70*+, 42%) in the second case.

Grades on ballots permit deducing the face-to-face vote, assuming that a higher grade for one list implies a vote for that list, and a same grade implies $\frac{1}{2}$ vote for each list. Using the six-grade language makes Jukesbox the winner with 57% of the votes; the [0, 100] scale makes it the winner with 54% of the votes. This together with the outcomes suggests that the grades were assigned in a manner consistent with the actual votes.

What is most interesting about this experiment is the use of the number-grades in the [0, 100] scale: they constitute much too rich a set, and they are not used in the same way by the voters. Table 17.1a gives the numbers of word-grades that were used. To begin, 87% of the 442 number-grades were multiples of 5, showing that a scale of twenty-one levels is already quite sufficient (moreover, 15 and 20 were never used). Since one ballot contained both majority judgment votes, it is possible to give the distributions of the number-grades corresponding to each word-grade along with other relevant information (for this purpose, the fifty-six grades that were not multiples of 5 were rounded to the closest multiple of 5; see table 17.1b). As may be observed, voters who assigned an *Excellent* gave number-grades as low as 40 and as high as 100; those who assigned a *Very Good* gave number-grades as low as 0 and as high as 100; similarly, to *Good* correspond a low of 0 and a high of 90 to *Acceptable* 5 and 62 to *Poor* 0 and 50. Although these are the extremes, the wide distributions corresponding to the three highest word-grades show that the voters ascribed practically no common meaning to the number-grades in the range [0,100]. On the other hand, the medians of the number-grades corresponding to each of the word-grades clearly show that statistically voters had in mind the French [0, 20] scale: dividing by 5 the median attached to *Excellent* yields 18, that attached to *Very Good* yields 15, and that attached to *Good* yields 12.

Another experiment that concerned a point-summing method (in part) was conducted on January 23, 2002, in the main entrance hall (*la péniche*) of the Institut d'Études Politiques of Paris (IEP, popularly referred to as Science Po) from 9:00 A.M. to 5:30 P.M. (Balinski, Laslier and van der Straeten 2002). It was a dress rehearsal for the 2002 Orsay experiment on approval voting that was conducted in parallel with the French presidential election of that year (see chapters 6 and 18). Participation was open to everyone, students, staff, and

Table 17.1a
Number of Word-Grades Used, École Polytechnique, December 2007

Excellent	113
Very Good	201
Good	78
Acceptable	29
Poor	8
To Reject	13
Total	442

Table 17.1b
Distribution of Number-Grades Attached to *Excellent, Very Good*, and *Good*, École Polytechnique, December 2007

Range [40,100]			*Excellent*						
89.2	90	9.2	100	95	90	85	80	75	≤ 70
Average	Median	S.D.	23.0%	10.6%	30.1%	15.9%	15.9%	0.9%	3.6%
Range [0,100]			*Very Good*						
73.4	75	14.3	≥ 90	85	80	75	70	65	≤ 60
Average	Median	S.D.	11.5%	6.5%	27.4%	15.4%	18.4%	7.0%	13.8%
Range [0,90]			*Good*						
56.0	60	15.9	≥ 75	70	65	60	55	50	≤ 45
Average	Median	S.D.	6.5%	14.1%	3.8%	39.7%	6.4%	10.3%	19.2%

Note: S.D. = standard deviation.

faculty, but the vast majority of those who voted were students. The context was the 2002 French presidential election, but the official candidates were not yet known, so the ballot presented fifteen likely candidates (in fact there were sixteen official candidates, three of whom were not on the Sciences Po ballot, whereas two who were did not run officially).

Participants were asked to complete one ballot. It listed the fifteen candidates in alphabetical order. To the line of each candidate corresponded slots in two columns: in the first, the voter was to enter a cross meaning approval (*assentiment*), otherwise leave a blank; in the second, the voter was to enter a number of points between 0 and 10, a blank interpreted as a 0. The total number of crosses given a candidate determined their order with approval voting (the approval voting part of the experiment is discussed in chapter 18); the sum of the points given a candidate determined their order with the point-summing method. A total of 429 persons participated. Tables 17.2a–17.2c give several statistics concerning number-grades.

Table 17.2a
Average Number of Number-Grades per Ballot, Sciences Po, January 2002

Number-Grade	Average No. per Ballot		
0	7.26	7.26	*To Reject*
1	1.33		
2	1.10	2.43	*Poor*
3	1.00		
4	0.86	1.86	*Acceptable*
5	1.18		
6	0.72	1.90	*Good*
7	0.56		
8	0.46	1.02	*Very Good*
9	0.15		
10	0.37	0.52	*Excellent*

Table 17.2b
Distribution, Ballots' Frequency of Maximum Number-Grades, Sciences Po, January 2002

Number-Grade	Frequency in Ballots
0	0.9%
1	0.5%
2	0.0%
3	1.2%
4	1.9%
5	3.0%
6	11.2%
7	16.3%
8	25.9%
9	9.1%
10	30.1%

Their distribution (table 17.2a) makes little sense; for example, half as many 9s are used as 10s. They were probably used to mean the same thing. Reinterpreting the number-grades with 0 meaning *To Reject*, 1 and 2 *Poor*, 3 and 4 *Acceptable*, 5 and 6 *Good*, 7 and 8 *Very Good*, and 9 and 10 *Excellent* makes a lot more sense. The majority-ranking of the fifteen candidates with the eleven grades is identical to the majority-ranking with the six grades except that the third- and fourth-place candidates are inverted. It is obvious that examples may be invented in which regrouping grades yields very different outcomes, but in practice such reasonable regroupings will lead to very similar outcomes.

Table 17.2c
Distribution, Ballots' Number of Maximum Number-Grades, Sciences Po, January 2002

No. of Maximum Number-Grades	Percent of Ballots	No. of Maximum Number-Grades	Percent of Ballots
1	76.0%	9	0.0%
2	18.9%	10	0.0%
3	2.8%	11	0.0%
4	0.9%	12	0.0%
5	0.2%	13	0.0%
6	0.0%	14	0.0%
7	0.0%	15	0.9%
8	0.2%		

Table 17.2d
Distribution, Ballots' Number of Different Number-Grades, Sciences Po, January 2002

No. of Different Number-Grades	Percent of Ballots
1	0.9%
2	5.8%
3	12.1%
4	17.0%
5	22.1%
6	22.8%
7	12.1%
8	6.3%
9	0.5%
10	0.2%

The confusion over the meaning of number-grades is borne out by table 17.2b: in 30.1% of the ballots the highest grade was a 10, in 16.3% the highest grade was a 7.

Table 17.2c shows once again that voters are often loathe to name one preferred candidate: the highest number-grade was given to two or more candidates in 24% of the ballots.

In conformity with Miller's 7 ± 2 conclusions (see chapter 8), eleven levels constituted too rich a language (even with fifteen candidates): Table 17.2d gives the distribution of the number of different points used in the ballots. A majority of the voters used five or fewer different number-grades; 80.9% of the ballots assigned six or fewer different number-grades to the candidates; and 93.0% assigned seven or fewer.

An experiment also conducted in France (Baujard and Igersheim 2007), tested two mechanisms at once, a point-summing mechanism with points 0, 1, or 2 and

approval voting (see chapter 18). It was realized in six different voting precincts of three towns: three in Illkirch (Alsace), two in Louvigny (Basse-Normandie), and one in Cigné (Mayenne). There were 2,836 participants (62% of those who voted officially). The ballot stated,

> *Instructions:* You give a grade to each of the 12 candidates: either *0*, or *1*, or *2* (*2* the best grade, *0* the worst). To do so, place a cross in the corresponding box . . . The candidate elected with [this] method is the one who receives the highest number of points.

The instructions are neutral: nothing is said concerning the meaning of 0, 1 or 2. The numbers induce relative and thus strategic behavior. Other numbers could have been given, for example, −1, 0, and +1; mathematically there is strictly no difference, but if these numbers had been used, the behavior of the voters would almost surely have been different.

On average a ballot contained 1.68 2s, 2.69 1s, and 7.63 0s. Behavior throughout the six precincts was very similar (table 17.3). Faced with a relatively simple means for expressing themselves, voters of a common culture seem to use those meanings in the same way. This is, we believe, an important observation.

On the other hand, the evidence suggests that voters used the numbers in a relative sense, not an absolute sense. In the ILC experiment 2s were used 1.68 times per ballot. Those voters share a common culture with the voters of Orsay, and the official first-round percentages of votes were not very different. If voters used the 2s as an absolute indication of merit, then their use should correspond approximately to an evaluation of *Excellent* or at least *Very Good*, or at least *Good*, . . . , as in the Orsay experiment (see table 6.1). But there are on average 0.69 *Excellent*s, 1.94 at least *Very Good* s, 3.44 at least *Good* s; none comes close to agreeing with 1.68, so the behavior seems not to be purely absolute. There is, however, an explanation for this. The average number of highest grades given by the voters of the Orsay experiment was 1.64 (for other related statistics, see table 18.6), which is in substantial agreement with the average number of the highest points—namely, 2, regularly given in the ILC experiment, 1.68.

Table 17.3
Rates of Use of *2*s, *1*s, *0*s, ILC Experiment, April 2007

	Cigné	Louvigny	Illkirch	All
No. of ballots	227	1,022	1,489	2,738
Average no. of *2*s	1.71	1.72	1.65	1.68
Average no. of *1*s	2.70	2.80	2.61	2.69
Average no. of *0*s	7.59	7.48	7.75	7.63

Table 17.4a
Number of Wins of the Centrist Candidate (Bayrou) with a Point-Summing Method, 2007 Orsay Experiment

	Royal		Bayrou		Sarkozy		Tie	
	(3)	(12)	(3)	(12)	(3)	(12)	(3)	(12)
First-past-the-post	4,420	1,855	1,455	1,924	3,922	6,204	203	17
Majority judgment	718	735	8,880	8,836	401	429	1	0
Point-summing	113	116	9,878	9,871	0	0	9	13
Borda	25	56	9,973	9,943	1	0	1	1

Note: Ten thousand samples of 201 ballots, drawn from all 1,733 ballots of the 2007 Orsay experiment. (3) indicates the experiment with three candidates; (12) indicates the experiment with all twelve candidates. One of the three candidates Royal, Bayrou, or Sarkozy, always wins among the twelve candidates.

This suggests that the 2s are purely relative; they are awarded to the voters' favorites. This seems quite natural, for when number-grades are not associated with a common language, their only real meaning is found in their strategic impact, which induces comparisons versus evaluations and immediately leads to a serious defect: Arrow's paradox.

In the traditional model Arrow's paradox arises when a candidate enters the ring or drops out, which may change the order of finish among the other candidates. In this case it may arise because one more or one less candidacy may alter the strategies of the voters, thus provoking a change in the order of finish among the others. When points 0, 1, 2 are used, a voter who gave a 2 to a candidate who dropped out, for example, may well decide to change a 1 to a 2 for another candidate (a favorite among those who are still candidates). In short, points induce comparisons, not evaluations, and comparisons open the door to Arrow's paradox.

The fourth experiment is the 2007 Orsay experiment. The statistical evidence shows that point-summing methods very strongly favor the centrist candidate (about as much as does Borda's method).[8] Four independent sets of 10,000 random drawings of 201 ballots from the Orsay ballots define different problems. The point summing-winner was determined by giving 5 points for *Excellent*, 4 points for *Very Good*, 3 points for *Good*, 2 for *Acceptable*, 1 for *Poor*, and 0 for *To Reject*. The number of wins in each of four cases are given in tables, 17.4a and 17.4b.

Recall that Bayrou had 25.5% of the first-round votes in the three Orsay precincts and that the 1,733 ballots were so favorable to him that almost any

8. See chapter 6, where the identity of the centrist candidate is deduced from the ballots.

Table 17.4b
Number of Wins of the Centrist Candidate (Bayrou) with a Point-Summing Method, 2007 Orsay Experiment

	Royal		Bayrou		Sarkozy		Tie	
	(3)	(12)	(3)	(12)	(3)	(12)	(3)	(12)
First-past-the-post	641	957	0	0	9,298	9,040	61	3
Majority judgment	631	640	4,326	4,301	5,043	5,059	0	0
Point-summing	155	132	9,546	9,547	251	271	48	50
Borda	51	10	8,750	9,989	1,096	0	103	1

Note: Ten thousand samples of 201 ballots, drawn from a sample of 501 ballots representative of the first-round national vote. (3) indicates the experiment with three candidates; (12) indicates the experiment with all twelve candidates. One of the three candidates, Royal, Bayrou, or Sarkozy, always wins among the twelve candidates.

reasonable method would elect him (see table 15.4), whereas nationally (and in the representative ballots) Bayrou only had 18.6% of the first-round votes. This explains the dramatic difference between tables 17.4a and 17.4b: for the database of table 17.4a, any reasonable method will elect Bayrou most of the time.

17.3 Conclusion

Point-summing methods are not acceptable methods for electing and ranking. The reasons are easily summarized.

· Points mean nothing except that they will be summed. They invite comparisons and not evaluations. This may lead to Arrow's paradox.

· When points are in abundance, for example, 0–100 or 0–10, in a large electorate versus a small jury of experts, the evidence shows that their meanings and uses differ widely among the voters.

· Summing or averaging numbers has absolutely no justification unless they constitute an interval measure. Even when the numbers carry the same meanings and are used by voters in the same way, they almost surely do not constitute an interval measure, so their sums *mean nothing*.

· If by happenstance the numbers did constitute an interval measure for a set of candidates, then almost surely they would not for a subset of them. Table 17.5 shows how very different the distributions of the grades may be depending on the slate of candidates, as is, of course, to be expected.

· The evidence shows that point-summing methods overwhelmingly favor a centrist candidate, which excludes them for political elections. A good method of election should neither overly favor nor overly penalize centrist candidates.

Table 17.5
Average Number of Grades per Majority Judgment Ballot, All Candidates and Four Important Candidates, 2007 Orsay Experiment

	All Candidates	Four Important Candidates
Excellent	0.69	1.57
Very Good	1.25	2.34
Good	1.50	1.94
Acceptable	1.74	1.49
Poor	2.27	0.99
To Reject	4.55	3.68
Total	12	12

Note: The four important candidates are Bayrou, Royal, Sarkozy, and Le Pen.
Totals are normalized to 12.

· Point-summing methods are the most susceptible to manipulation. They are not strategy-proof-in-grading; every judge or voter can both increase and decrease the final score of any candidate or competitor; they are among the methods that maximize manipulability (see chapter 10); their probability of cheating is 1 (see chapter 12); and they are not partially strategy-proof-in-ranking (see chapter 13).

18 Approval Voting

If names be not correct, language is not in accordance with the truth of things. If language be not in accordance with the truth of things, affairs cannot be carried on to success.
—Confucius

A relatively recent novelty in voting mechanisms, approval voting, is championed by Steven J. Brams, Peter C. Fishburn (Brams and Fishburn 1983), and many others. It was first proposed formally by Robert J. Weber (1977).[1] It has been used to elect the officers of several important scientific societies, to elect national representatives in Russia, and in a referendum held in the state of Oregon where one proposition among five was to be chosen. *Approval voting*, which was discussed in the context of left-right spectra in chapter 6, allows each voter to cast as many votes as he wishes, but at most one per candidate, so each voter either approves of a candidate by giving him one vote or disapproves by giving none, and the winner is the candidate with the most votes. The voter is offered the possibility of giving any number of candidates 1 point and the others 0 points, and the order among the candidates is determined by the sums of their points. In its traditional practice and presentation, approval voting is the most restrictive of a large family of point-summing methods (see chapter 17); if presented and practiced completely differently, it may be seen as the most restrictive case of the majority judgment.

18.1 Traditional Arguments

The traditional commonsense arguments for approval voting date back to the 1970s and seduced many (including the authors). First, voters are better able to express their preferences than by naming a single candidate, which makes

1. The idea of "dichotomous voting schemes" was, however, considered earlier by Bartoszyński (1972).

it more attractive for voters to participate in elections. Second, voters who are very positive about a candidate who for some reason is a certain loser (e.g., a one-issue candidate) are able to express this by voting for that candidate yet also voting for other candidates who have a better chance of winning. Third, when there are at least three candidates, a plurality winner often fails to obtain a majority of the voters' approvals, whereas an approval vote winner may well do so and is conferred more legitimacy in any case. Fourth, approval voting helps elect the better candidate: "in general [approval voting] helps elect the Condorcet candidate" (Brams and Fishburn 1983). Fifth, the method is easy to understand and to use in practice, and most voting machines can be adjusted to accommodate it. None of these arguments except the fifth have stood the test of time. The first two pale in comparison with a majority judgment endowed with a rich language of grades (as in the 2007 Orsay experiment); the next two are often false in practice.

Many of the theoretical arguments for and against approval voting, cast in terms of the traditional model[2] are not convincing either.

Approval voting is not strategy-proof (see chapter 13). At first glance, it would seem that an elector should vote *sincerely*: approval of a candidate implies approval of all candidates higher in the elector's ranking (how can this be harmful?). This is true and trivial to prove when there are only three candidates: an elector should always vote for his top and never for his bottom candidate, so in any case the vote must be sincere whether an approval is cast for the other candidate or not. But in the presence of at least four candidates, sincerity is not necessarily the best strategy.[3]

In general, many different profiles of preference-orders will lead to the same approval votes, and two voters with the same preference-orders may cast their approval votes very differently. This opens the door to what seems to be a severely damaging drawback of approval voting: the results—the winner and the order among candidates—are *completely indeterminant* in the following sense. By varying the voters' individual choices of sincere strategies, one and the same profile of individual preference-orders can produce *every* possible collective-order and thus *every* possible winner (Saari and van Newenhizen 1988). An actual vote shows that this can occur: the Social Choice and Welfare

2. Except when the preferences themselves are dichotomous, meaning voters consider candidates as only "good" or "bad" (Barberà, Sonnenschein, and Zhou 1991; Bogomolnaia, Moulin, and Strong 2005).

3. For a detailed analysis of sincerity in approval voting, see Merrill and Nagel (1987); for a fascinating account of the strategies provoked by a type of approval voting in the U.S. presidential election of 1800, leading to thirty-four tied ballots in six days of voting in the House of Representatives before the election of Thomas Jefferson instead of Aaron Burr, see Nagel (2007).

(SCW) Society's 1999 presidential election (Brams and Fishburn 2001; Saari 2001a). In this approval voting election, seventy-one members voted, and the result ranked the candidates $A \succ_S C \succ_S B$ (with 32, 30, and 14 votes, respectively); two voters approved of all three candidates, three approved of two, sixty-four approved of one, and two of none. Voters were asked to indicate their preference-orders to enable the results to be analyzed; fifty-two complied. The approval voting result among the fifty-two was the same, $A \succ_S C \succ_S B$ (with 22, 20, and 9 votes, respectively); of these, one voter approved of two candidates, forty-nine of one, and two of none. The profile of the fifty-two individual preference-orders was

(1) $13 : A \succ B \succ C$ (2) $11 : A \succ C \succ B$ (3) $9 : B \succ C \succ A$
(4) $11 : C \succ A \succ B$ (5) $8 : C \succ B \succ A$.

If we imagine that these fifty-two voters vary their behavior, voting sincerely either for their top candidate or their top two candidates, approval voting produces every possible collective-order:

$A \succ_S B \succ_S C$: type 1 votes top two, others top one;

$A \succ_S C \succ_S B$: all vote top one;

$B \succ_S A \succ_S C$: types 1 and 5 vote top two, others top one;

$B \succ_S C \succ_S A$: types 1, 3, and 5 vote top two, others top one;

$C \succ_S A \succ_S B$: type 2 votes top two, others top one;

$C \succ_S B \succ_S A$: types 1, 2, and 5 vote top two, others top one.

This is an amusing and surprising mathematical result. But it makes little sense as a practical matter to imagine that sincere approval voters place the boundary between approvals and disapprovals arbitrarily; on the contrary, they draw that dividing line with care, strategically or in accord with their evaluations of the candidates. Analyzing approval voting when voters are believed to have rank-orders in their minds suggests irrelevant questions, and answers that have no practical significance.

Complete indeterminacy is a property of any point-summing method. The plurality system and Borda's method are not completely indeterminate, but they are not point-summing methods. With plurality voting, the only possibility is to assign at most one candidate 1 point and the others 0, and with Borda's method the assigned points are necessarily all different, 0, 1, 2 up to the number of candidates less 1.

Approval voting does not and cannot guarantee the election of a Condorcet-winner (not *necessarily* a bad property). This is proven by the SCW Society fifty-two-voter profile presidential election. For this profile the society's

preference-ranking according to the simple majority-rule is $C \succ_S A \succ_S B$ and is transitive, so C is the Condorcet-winner, whereas A is the approval vote winner.

Notice also that the winning candidate A of the SCW Society election, with 32 votes from seventy-one ballots, is not approved of by a majority. A is also the winner with 22 votes from the fifty-two ballots that reported the voters' preference-rankings and thus is not approved of by a majority of those voters either. Strategic voters, as Saari has pointed out, are more than likely to vote only for their favorites. He notes that a politician—the late U.S. Senator and Governor of North Carolina, Terry Sanford, who went on to be President of Duke University—recognized this proclivity, saying of approval voting, "The great weakness, it seems to me, [is] that most voters . . . are inclined to cheat a little and 'single shot' if it suits their purposes, which it generally does" (cited in Saari 2001a). Moreover, experience suggests that familiarity with approval voting abets single-shot votes when there are relatively few candidates (as predicted by Sanford), so outcomes revert to those obtained with plurality voting. The cure induces the identical malady, just as it did in the SCW Society election (whose members are the acknowledged experts of voting theory).

Finally, when strategic behavior is ignored, approval voting may elect a candidate preferred by only one voter of the electorate, as an example discussed earlier but in a different context shows:[4]

	k voters	1 voter	k voters
X :	12,…,12,	12,	4,…,4
Y :	16,…,16,	8,	8,…,8

Imagine that the numbers represent the merit of the candidates in the eyes of the voters (on a scale from a low of 0 to a high of 20); the higher the number, the higher the merit; their comparisons determine the voters' preferences. Suppose that voters give an approval to any candidate whose merit number is 10 or above and otherwise disapprove. Thus X wins with $k+1$ to Y's k approvals, and yet only one voter prefers X to Y.

These are three good reasons to study what happens when approval voting is analyzed as a game.

18.2 The Game of Approval Voting

The game-theoretic point of view analyzes the strategic behavior of voters whose objective is to maximize their utility functions. When a mechanism is

4. Example 16.2. This objection was, in fact, raised against the majority judgment by approval voting enthusiasts.

not strategy-proof, the equilibria of the game of voting are of interest. This problem is first discussed in the context of approval voting in this section. Chapter 20 takes up the problem in a more general context, including majority judgment. Chapter 19 addresses one aspect of the strategic behavior of candidates under majority judgment.

In the game of approval voting, a voter has a utility function that is a function only of who wins (which defines the voter's weak preference-order over the candidates). She chooses a strategy that consists of casting a 0 or 1 for each candidate, knowing that every other voter is faced with the same choice and that the winner is the candidate with the most 1s. It has been shown (Brams and Fishburn 1978) that with approval voting, if there exists a Condorcet-winner C, then there exists a Nash equilibrium—a strategy for every voter with the property that no voter can change her vote to obtain what for her is a better outcome—that elects C in sincere, undominated strategies. A strategy of a voter is *undominated* if there is no other strategy that gives at least as good an outcome and for some profiles a strictly better outcome.

The significance of this result, however, is doubtful. The fact is that Nash equilibria are plentiful (see chapter 20); in particular, approval voting can elect any candidate in sincere, undominated strategies at a Nash equilibrium (Brams and Sanver 2006). (This is related to complete indeterminacy.) Many different refinements introduced by game theorists have been proposed, but most fail. For example, Selten's perfect equilibria, Myerson's proper equilibria, or Merten's stable equilibria applied to approval voting all require that voters have unrealistic amounts of information concerning the behavior of the other voters, ask them to do complex probability computations, and in any case admit many equilibria that sometimes include insincere strategies, sometimes elect a Condorcet-loser, and sometimes give the Condorcet-winner no approval votes whatsoever (De Sinopoli, Dutta, and Laslier 2006).

One of the elections that spurred Weber to propose approval voting in the first place seems to have been the 1970 Senate contest in the state of New York. There were three candidates: James Buckley, a conservative Republican running as the candidate of the Conservative Party, Charles Goodell, a moderate Republican, the candidate of the Republican Party and the Liberal Party, and Richard Ottinger, the candidate of the Democratic Party. Goodell and Ottinger were both viewed as liberal (in the U.S. sense) or center left. The surprising result was

Buckley (B): 39% Goodell (G): 24% Ottinger (O): 37%

Buckley was elected but the voters of New York had clearly voted against the conservative candidate. It has been widely claimed that if approval voting had

been used instead of first-past-the-post, Buckley would have been defeated and Goodell elected. What scientific analysis supports this claim?

For the sake of the argument and in the context of the traditional model, hypothesize that New York's electors had the preference-profile

(1) $39\% : B \succ G \succ O$ (2) $24\% : G \succ O \succ B$
(3) $37\% : O \succ G \succ B$.

Had the right (type 1) voted for its top choice and the left (types 2 and 3) for their top two choices, the argument goes, Goodell and Ottinger would have been tied for first, so one of the two moderate candidates would have been elected. But why expect this? An elector of type 2, sensing this result, would reason that it is better to vote only for Goodell, and for the same reason an elector of type 3 would vote only for Ottinger. If most electors of the left followed this reasoning, Buckley would again emerge the winner.

In practice, electors have some idea of what the results will be from repeated public opinion polls (or other evolving information), so their votes may be imagined as optimal responses in view of the latest information. Given any voting system, imagine a succession of polls, each a virtual election in that system, in which every voter uses the information of the last available poll to cast an optimal vote in the next election or poll. After several rounds do the results converge to a stable winner, or better yet, to a stable order of finish among the candidates, meaning that the winner or the order finally repeats itself ?

This problem may be analyzed in two different albeit related manners. One is to consider a dynamic process where every voter sends his optimal strategic response to the last poll. The question is then to find whether there are reasonable conditions under which the result converges to a stable winner and stable order of finish, together with best response strategies that are simple and easy to use by voters. The other is to consider the more abstract static question: Does there exist a strategy-profile of the voters whose outcome is such that no voter (or more realistically, no coalition of voters) would wish to change his strategy (or their strategies) were he (or they) permitted to do so.

Robert B. Myerson and Robert J. Weber (1993) were the first to obtain significant results in the study of the second question. They imagined a sophisticated model similar to those used to study market equilibria in economics. Voters have numerical utilities for the election of candidates, and in addition probabilities $p = (p_1, \ldots, p_n)$ on the distribution of the total approvals of each candidate that permit them to determine their expected utility gains as a function of their votes. Thus, given any p that is perceived by all, the outcome may be computed when voters cast their ballots to maximize their expected utility gains, whatever the mechanism of voting that is used.

A vote or cast of ballots is a *voting equilibrium* if voting or casting their ballots in the same way had they known the actual distribution of the votes they had cast, also optimizes every voter's expected utility gains: the result is stable. Myerson and Weber proved the existence of equilibria and compared those that arise when different mechanisms of voting are used, notably, first-past-the-post, Borda's, and approval voting. Their main example was a very rough approximation of the 1970 New York Senate election specified by three types of voters having the following utilities for the three candidates (B, G, O)

(1) $40\% : (10, 0, 0)$ (2) $30\% : (0, 10, 9)$ (3) $30\% : (0, 9, 10)$

Thus, for example, voters with type 2 utilities have no use for Buckley, immensely prefer the two others, and have a very slight preference for Goodell over Ottinger.

First-past-the-post admits three voting equilibria. In the first, B is the likely winner: every elector votes for his top choice. In the second, G is the likely winner: voters of type 1 utilities vote for B, the others for G. In the third, O is the likely winner: voters of type 1 utilities vote for B, the others for O.

Approval voting also admits three equilibria. In the first, O is the likely winner: voters of type 1 utilities give a 1 to B; voters of type 2 utilities give a 1 to G and to O; and voters of type 3 utilities give a 1 to O. In the second, G is the likely winner: voters of type 1 utilities give a 1 to B; voters of type 2 utilities give a 1 to G; and voters of type 3 utilities give a 1 to G and to O. In the third equilibrium, all three are likely winners: voters of type 1 utilities give a 1 to B; of the 30% of voters with type 2 utilities, 20% give a 1 to G and 10% give a 1 to G and to O; of the 30% of voters with type 3 utilities, 20% give a 1 to O and 10% a 1 to G and to O.

Borda's method admits an infinity of equilibria, with all three candidates the likely winners in each. Each is determined by α, where $10\% \leq \alpha\% \leq 20\%$, as follows: of the 40% of voters of type 1 utilities, $\alpha\%$ give a 2 to B and a 1 to G and $40 - \alpha\%$ give a 2 to B and a 1 to O; of the 30% of voters of type 2 utilities, $\alpha\%$ give a 2 to G and a 1 to O and $30 - \alpha\%$ give a 2 to G and a 1 to B; of the 30% of voters of type 3 utilities, $\alpha - 10\%$ give a 2 to O and a 1 to B and $40 - \alpha\%$ give a 2 to O and a 1 to G.

Myerson and Weber conclude, "In summary, for the electoral situation studied here we find that under the plurality rule, there is an equilibrium in which the minority (right) candidate is the only likely winner. Under Borda's rule, the minority candidate is always a likely winner at equilibrium. Only under approval voting do we see simultaneously the existence of equilibria in which

the minority candidate is not a likely winner, and the nonexistence of equilibria in which the minority candidate is the only likely winner." Poundstone (2008) reports that Weber believes this is the analysis that "makes the strongest theoretical case for approval voting," but he goes on to say, "No one knows with certainty how masses of real-world, twenty-first-century approval voters will act. Myerson and Weber have shown that approval voting works well when voters are the perfect Machiavelians of game theory . . . However, no one expects all voters to be so calculating, or for all elections to be Nash equilibria" (214, 218). The facts support Poundstone's conclusion: in the French presidential contest of 2007—an election fraught with temptations to vote strategically (*voter utile*)—polls estimated that at most 30% of the votes were deliberately strategic. Myerson and Weber did not study the dynamics.

Myerson investigates extensions of this work in several later papers (1998; 2000; 2002). The mathematical theory of the refinement of Nash equilibria introduces probabilistic perturbations into the model. As Myerson (2002) states, "The analysis of voting games in this paper is based on the assumption that . . . each voter chooses his ballot to maximize the utility that he gets from the election, which is assumed to depend only on which candidate wins the election. This assumption seems natural and realistic, but it implies that each voter cares about his choice of the ballot only in the event that his ballot could pivotally change the outcome of the election. So this theory of rational voting necessarily implies that voters' decisions may depend on the relative probabilities of various ways that one vote may be pivotal in the election, even though these pivotal probabilities may be very small in a large election." These are very small probabilities indeed, and the assumptions are not, we believe, natural or realistic. But the merit of Myerson's approach is that it enables several different methods of voting to be compared in one model. For example, he is able to show that the number of approval voting equilibria is smaller than the number of plurality voting equilibria, so their predictive value is better. He introduces uncertainty by varying the number of voters (or of ballots counted) according to a Poisson distribution. Thus, there is a minute probability that any one voter is the only one to cast a ballot, which clearly makes him pivotal. This model affords two main advantages that simplify the computations: voters share common public information and the actions of voters are independent. Myerson is then able to show, in an admittedly simplified situation in which there are three candidates and two types of voters (so necessarily at least half of the voters have the same preferred candidate), that a Nash equilibrium winner assures the majoritarian outcome, which is not true of plurality voting among other seemingly reasonable methods. The dynamics, however, are not studied. Moreover, the voters'

optimal strategies may not be sincere, and there are situations where there is a Condorcet-winner that is the outcome of no Nash equilibrium.

Three examples put in perspective the qualitative interpretation of these results (Núñez 2008). In the first, there are three candidates and three types of voters (instead of two). One type constitutes a majority so that its preferred candidate is the Condorcet-winner C, but C is not the winner in every equilibrium. In the second, there are four candidates and a Condorcet-winner C who is second in every voter's preference, but at equilibrium the strategies are insincere and C receives zero approval votes. In the last example, there are exactly two equilibria, but neither elects the Condorcet-winner.

Laslier (2009) presents more positive results. Uncertainty is introduced via technical errors: ballots may be miscounted. Miscounts occur independently of the identities of voters and of candidates, and are interpreted as a probability of error. Laslier states, "Rationality implies that a voter can decide her vote by limiting her conjecture to those events in which her vote is pivotal ... [T]his is a very rare event, and it may seem unrealistic that actual voters deduce their choices from implausible premises. One can wish that a positive theory of the voter be more behavioral and less rational."

In this model the strategy that maximizes a voter's expected utility when there are sufficiently many voters—her best response, introduced as the "poll assumption" by Brams and Fishburn (1983, 115) and later called the "leader rule"—is simple to describe. Suppose the two leading contenders of the predicted order are $X \succ_S Y$ (where in case of ties X is taken to be the elector's preferred among the candidates in first place and Y is taken to be the elector's preferred among the candidates in second place). Then what we call the *poll-leader rule* is this:

· If the voter prefers X to Y, then she votes for X and the candidates she prefers to X.

· If the voter does not prefer X to Y, then she votes for the candidates she prefers to X.

This is at once a simple and reasonable rule that yields several important results.

· *Result 1* The poll-leader rule is sincere and undominated, and if there is an equilibrium outcome with no tie, the winner is a Condorcet-winner. This is easy to see. Suppose there are a winner and a runner-up satisfying $X \succ_S Y$. If a majority preferred Y or any other candidate Z to X, then in the next round Y or Z would have more approvals than X, a contradiction. So X is the Condorcet-winner.

· *Result 2* The score of the winner X is his majoritarian score against Y, and the score of any other candidate is his majoritarian score against the winner X.

· *Result 3* Conversely, if the preference-profile has a Condorcet-winner C and a unique strongest contender, then the game has a unique equilibrium that elects C. This is also easy to see. With the poll-leader rule, every candidate's approval score against C must be less than C's.

The poll-leader rule defines a dynamic process (which is not studied in Laslier 2009). At each step voters cast their strategic approval ballots which determine the leaders of the next step. Unfortunately, the process does not necessarily converge, as the following preference-profile shows (where the relative shares of the voter types are given):

$$\begin{array}{ll} 1/3-\epsilon : A \succ C \succ B \succ D & \epsilon : A \succ B \succ D \succ C \\ 1/3-\epsilon : D \succ C \succ A \succ B & \epsilon : D \succ A \succ B \succ C \\ 1/3-\epsilon : B \succ C \succ D \succ A & \epsilon : B \succ D \succ A \succ C. \end{array}$$

If a poll predicts that at time t the leaders are $A \succ_S C$, then the scores at time $t+1$ are $s(A) = 1/3+2\epsilon$, $s(B) = 1/3$, $s(C) = 2/3-2\epsilon$, and $s(D) = 2/3-\epsilon$, so the new leaders are $D \succ_S C$. Continuing, at time $t+2$ the leaders become $B \succ_S C$ and at time $t+3$ they are once again $A \succ_S C$, and so on. So the process does not converge, and the Condorcet-winner C is the perennial runner-up.

The poll-leader rule is based on the two leading candidates. Simpler sincere response strategies may be envisaged. For example, if the elector's preferred candidate is the expected winner X, then he votes for X; otherwise, he votes for the candidates he prefers to X. But the process does not converge. For consider the SCW Society's presidential election (see section 18.1), and suppose the Condorcet-winner C is the predicted winner. Then the approval voting outcome at the next stage is $A \succ_S B \succ_S C$ (with respective scores 24, 22, 19), so A wins. But if A wins and the individuals again use the same strategy, the next outcome is $C \succ_S A \succ_S B$, so C wins, and so on. Indeed, exactly the same phenomenon arises with the preference-profile postulated for the 1970 New York Senate election. The simple majority rule outcome is $G \succ_S O \succ_S B$, so G is the Condorcet-winner. But when G is the expected winner, the strategy yields $B \succ_S O \succ_S G$, whereas when B is the expected winner, the strategy yields $G \succ_S O \succ_S B$.

There is, however, a best response dynamic that does converge to the Condorcet-winner (when he exists) that is based on the poll-leader rule.[5] Given

5. This dynamic is studied in the context of a more general model in chapter 20.

the outcome of a poll at time t, two candidates from among those that have more than half of the electorate's approvals are selected at random and arbitrarily designated winner and runner-up. Otherwise, the leader and runner-up are the candidates with the highest number of votes (chosen at random in case of ties). The voters then use the poll-leader rule, and the next poll is taken. After the first stage, the Condorcet-winner C necessarily has a majority of approvals, but other candidates may also have majorities (as in the last example). C will necessarily conserve a majority of approvals at every subsequent stage (as in the last example). Therefore, at some stage C must be selected as the leader. At that stage, all other candidates will have approval scores equal to his majoritarian (head-to-head) score against C, and C's approval score will be his majoritarian score against the runner-up. C remains the leader in all subsequent stages. If there is no Condorcet-winner, the leaders in subsequent stages will cycle among the top cycle of candidates who defeat all others in face-to-face encounters.

If one accepts as a working hypothesis that these analyses of the game of approval voting show approval voting works well, what makes it work at all in this model must be underlined: the purely strategic behavior of individual electors. However, the very essence of an election is to arrive at a collective decision based on the electors' true opinions; that it should depend only on an elector's strategic considerations concerning winners and nothing else, sometimes implying votes at variance with the voter's true opinions, is not commendable. Moreover, voters are most likely influenced or coordinated by a small number of leaders or parties. The game of majority judgment voting leads to more felicitous conclusions (see chapter 20, where games of voting are studied in greater detail): both behavioral and rational voters are included in the context of one model.

18.3 Approval Judgment

Fundamentally, does it make any sense to sum the number of approvals of the candidates? The standard instructions and explanations accompanying approval voting ballots are deliberately neutral: no question is posed, no meaning is attached to an approval. For example, the SCW Society approval ballot of 2007 stated, "You can vote for any number of candidates by ticking the appropriate boxes." The only rational meaning that can be attached to an approval or "tick" is that a 1 is contributed to the sum of the votes given to a candidate. This induces comparisons of candidates, therefore strategic voting, and thus the possibility of Arrow's paradox.

Table 18.1
Polling Results, March 20 and 22, 2007, French Presidential Election

	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%		

Source: Polling results by BVA.

On the other hand, the words "approval" and "disapproval" carry meanings, and an elector may be influenced by those meanings. A poll conducted several weeks before the first round of the French presidential election of 2007 shows the crucial importance of the meanings attached to the 1s and 0s (table 18.1). The electors gave completely different answers to the two questions. The question on the left in the table poses an absolute question, that on the right a relative one. The first invites an evaluation, the second proffers a contrast. Significantly, the first question elicited a "yes" for the four major candidates, an answer considerably more in keeping with their *Good* or better grades in the 2007 majority judgment experience than the answer to the second question. But which of these questions (or others) does an approval voter have in mind? Undoubtedly, different voters have different conceptions, and so answer different questions. Their votes carry very different meanings: adding them makes no sense.

However, approval voting may be presented, practiced, and analyzed as a special case of the majority judgment when the common language of grades consists of exactly two grades. In this case, the majority-ranking is precisely the approval voting ranking. Suppose (for the sake of the explanation) that the two grades are 1 and 0. A candidate's majority-grade is 1 if she has a majority of 1s and 0 otherwise. A candidate with a majority-grade of 1 is ranked ahead of a candidate with a majority-grade of 0 by both the majority judgment and approval voting. If two candidates have the same majority-grade, the majority-ranking puts the first ahead of the second if the first has more 1s, as does approval voting.

But to be an instance of the majority judgment, a clear and absolute question must be posed, and voters must understand that they are assigning grades to candidates: they are making absolute evaluations of the merits of candidates, *not*

comparing them. This has not been the case in any of the theoretical discussions or applications of approval voting, where the instructions—giving 1s and 0s, adding them, ranking candidates according to their sums—and the analyses of results all suggest the point of view that what is important is comparisons. Had anyone imagined that crosses and no crosses, ticks and no ticks, or 1s and 0s were grades—absolute evaluations—they would (or should) have immediately pointed out that approval voting is a mechanism that excludes Arrow's paradox and thus satisfies IIA (and, of course, satisfies all the other good properties of the majority judgment). This is not true when the implicit or explicit question is relative (like the question at the right in table 18.1, or when crosses or ticks are to be added), for the responses depend on the lists of candidates, leading to Arrow's paradox.

When approval voting is practiced as a majority judgment, a language of two words must be formulated that makes clear the evaluations are absolute grades. To distinguish it from its traditional practice we call it *approval judgment*. The question at the left in table 18.1 (Would each of the following candidates be a good President of France?) is one reasonable possibility of approval judgment. Had it been used in the 2007 Orsay experiment, a candidate might have received a "yes" from every voter who assigned him a grade of at least *Good*. Among the four major candidates this would have given the following result: Bayrou 69.4%, Royal 58.5%, Sarkozy 53.1%, and Le Pen 13.8%. However, had *Very Good* been substituted for *Good*, the result would have been Bayrou 44.3%, Royal 39.4%, Sarkozy 38.9%, and Le Pen 7.6%. It happens that the order among the candidates is the same, but the numbers are very different. And if *Excellent* had been substituted instead, the result would have been a different order: Sarkozy 19.1%, Royal 16.7%, Bayrou 13.6%, and Le Pen 3.0%. To elect a candidate with only 44% of the electorate that affirms he is *Very Good*, the others denying it, or with 80.9% denying that a candidate is *Excellent*, is not a particularly salutary result. The SCW Society election experienced the same difficulty, and other reported experiments do as well.

More probing are analyses that use the 1,733 valid majority judgment ballot of the 2007 Orsay experiment. In the first analysis, 10,000 random samples of 101 ballots were taken to compare how many times each of the candidates was elected by each of four majority judgment methods: (1) the majority judgment as it was presented with a language of six grades, (2) approval judgment where approval means a candidate received a grade of *Good* or better, (3) approval judgment where approval means *Very Good* or better, and (4) approval judgment where approval means *Excellent*. In every case one of the three major candidates, Bayrou, Royal, and Sarkozy, is the winner. The results are given in table 18.2a.

Table 18.2a
Number of Wins among All Candidates, Winner Always Royal, Bayrou, or Sarkozy, 2007 Orsay Experiment

	Royal	Bayrou	Sarkozy	Tie
Approval judgment ($\succeq$*Excellent*)	2,692	602	5,905	801
Approval judgment ($\succeq$*Very Good*)	1,387	6,489	1,446	678
Majority judgment (six grades)	1,271	7,798	919	12
Approval judgment ($\succeq$*Good*)	215	9,632	32	121

Note: Four Majority judgment methods. Ten thousand samples of 101 ballots, which were drawn from 1,733 ballots.

Table 18.2b
Number of Wins among All Candidates, Winner Always Royal, Bayrou, or Sarkozy, 2007 Orsay Experiment

	Royal	Bayrou	Sarkozy	Tie
Approval judgment ($\succeq$*Excellent*)	561	8	9,186	245
Approval judgment ($\succeq$*Very Good*)	1,329	1,339	6,686	646
Majority judgment (six grades)	1,364	4,048	4,579	9
Approval judgment ($\succeq$*Good*)	401	8,743	487	369

Note: Four Majority judgment methods. Ten thousand samples of 101 ballots, which were drawn from a sample of 501 ballots representative of the national vote.

The second analysis is different only in that each of the random samples is drawn from a set of 501 ballots, drawn randomly from the 1,733, that are representative of the first-round national vote (see chapter 6). The results are given in table 18.2b.

The results are concordant. First, they show the dramatic impact of the question that is posed in conducting an approval voting election. Second, when approval means *Good* or better, the candidate of the center, Bayrou, is favored; when approval means *Very Good* or better, and still more so when it means *Excellent*, the candidate of the center is penalized. The problem of exactly which question should be posed and which answer solicited is at once important politically and difficult scientifically. Almost all the literature on approval voting to date has completely ignored the problem of how the question and answer should be formulated, with one notable exception (Koc 1988). Koc shows that the wording of the instructions influences the behavior of the voters. The wording actually does more: it determines winners.

The majority judgment has been criticized because it can elect a candidate preferred by only one voter. As noted, such examples are easily found because only the grades of a candidate count, not who gave them (see example 16.2). But, of course, the same is true of approval voting (or judgment), as that same

example shows, when approval means a grade of at least 10. The response of approval voting enthusiasts: strategic voting will avoid it. When one argument does not work, try another. But, of course, strategic voting will avoid the phenomenon with the majority judgment as well. Comparison is valid when made in the same context.

The practical instances where evaluations are made on the basis of a scale of measurement composed of two levels are rare. Sometimes, in some university courses, students may elect to be graded on a *Pass* or *Fail* scale. There are also situations where sets of objects or competitors are to be classified as acceptable or not. One example is candidates for appointment to a university faculty in France, who must first be judged qualified in order to be allowed to apply for a position. Another example is election to the National Baseball Hall of Fame. A panel of journalists judges candidates either suitable or not for election. Each elector may approve multiple candidates, but to be successful a candidate must be listed on 75% of the ballots (for an analysis of such methods, see Barberà, Sonnenschein, and Zhou 1991).

But when it comes to electing and ranking candidates, a two-level scale of measurement seems so unnecessarily restrictive as to be unnatural. In political elections, or indeed competitions among performers or products, the evaluations of voters or judges are invariably more complex than that allowed by a two-level scale. The 2007 Orsay experiment showed the ease with which voters used a six-level scale, and experience in judging skaters, divers, pianists, students, wines, and other competitors or products shows that considerably finer scales may sensibly be defined and used. So why confine a language of voting to only two grades?

18.4 Practice

On April 21, in the first round of the French presidential election of 2002—still fascinated by the idea of approval voting and before we had any inkling of working on the general problem of electing and ranking—one of the authors initiated an approval voting experiment,[6] conducted under the same general conditions as the Orsay experiment of 2007, in five of Orsay's twelve precincts[7] and the one precinct of Gy-les-Nonains, a small country town in Loiret.

6. The idea of an experiment on approval voting on a large scale in parallel with a French presidential election actually goes back to 1995, when Balinski and Laurent Mann prepared a basic plan but were too late to realize it. For a more detailed account of the 2002 experiment, see Balinski et al. (2003).

7. 1st, 5th, 6th, 7th, and 12th precincts.

Table 18.3a
Number of Ballots with k Crosses ($k = 0, 1, \ldots, 16$), Approval Voting Experiment, Five Precincts of Orsay and Gy-les-Nonains, First Round, French Presidential Election, April 21, 2002

	No. of Crosses									
	0	1	2	3	4	5	6	7	8	9–16
No. of Ballots	36	287	569	783	492	258	94	40	16	12
% of Ballots	1.4	11.1	22.0	30.3	19.0	10.0	3.6	1.5	0.6	0.5

Table 18.3b
Approval Voting Results, Five Precincts of Orsay and Gy-les-Nonains, First Round, French Presidential Election, April 21, 2002

	Percent of Ballots with Crosses	Percent of All Crosses	Official Vote First-Round
Jospin	40.5%	12.9%	19.5%
Chirac	36.5%	11.6%	18.9%
Bayrou	33.5%	10.7%	9.9%
Chevènement	30.3%	9.6%	8.1%
Mamère	28.9%	9.2%	7.9%
Madelin	21.3%	6.8%	5.0%
Taubira	18.9%	6.0%	3.2%
Lepage	17.9%	5.7%	2.8%
Besancenot	17.6%	5.6%	3.1%
Laguiller	15.4%	4.9%	3.7%
Le Pen	14.6%	4.6%	10.0%
Hue	11.5%	3.6%	2.7%
Saint-Josse	7.8%	2.5%	1.7%
Boutin	7.8%	2.5%	1.3%
Mégret	7.7%	2.4%	1.3%
Gluckstein	4.3%	1.4%	0.8%
Total	314.6%	100%	100%

Of the 3,346 voters who voted officially, 2,597 (78%) participated in the experiment and 2,587 ballots were valid. Some of the results of this experiment have already been presented in chapter 6. As explained there, officially voters were confronted with having to give their one vote to one of sixteen candidates. The ballot of the experiment consisted of a list of the candidates together with completely neutral instructions:

Rules of approval voting: The elector votes by placing crosses [in boxes corresponding to candidates]. He may place crosses for as many candidates as he wishes, but not more than one per candidate. The winner is the candidate with the most crosses.

On average, the voters cast 3.15 crosses per ballot (the distribution is given in table 18.3a). The official system offered voters seventeen possible messages; approval voting offered more than 65,000.[8] Of the 2,587 valid ballots, 813 were different. Voters expressed relief at having the possibility of casting crosses for as many candidates as they wished.

The outcomes in the six voting precincts with approval voting and with the official voting are given in table 18.3b. The one significant difference between them is that Le Pen is third in the official vote and eleventh in the approval vote (other differences are that in the official voting, Laguiller moves up three places to follow Madelin, and Besancenot moves up one place to follow Taubira). The four most important candidates, Chirac, Le Pen, Jospin, and Bayrou, all lost relative support in approval voting, whereas every one of the minor candidates gained relative support. If Orsay and Gy-les-Nonains were at all representative of France, the results of the experiment showed that the indecision of the country—the lack of enthusiasm for any one candidate or party—was even more extreme than the usual method of voting indicated. No candidate received anywhere near a majority of the ballots. No legitimacy is added to the first-place candidate, contrary to the claims made for approval voting. Whereas we entered into this experiment persuaded by the usual commonsense arguments that approval voting was a good idea, the results left us with a distinct feeling that it is not a reasonable mechanism. We did not know exactly why. Now we do.[9]

The result of the second round on May 5, 2002, in the five precincts of Orsay and the one of Gy-les-Nonains was Chirac 89.3%, Le Pen 10.77%. In contrast with the majority judgment, the electorate's will expressed by approval voting is not sufficient to predict this outcome (nor therefore to estimate the possible result of any other face-to-face confrontation). Crosses and no crosses do not communicate enough information. The problem is the frequency with which voters assigned crosses to two candidates or no crosses to two candidates (table 18.3c). Do crosses or no crosses for two candidates imply indifference between them, or not?

Three estimates of a face-to-face vote between Chirac and Le Pen were calculated. In each, if a candidate has a cross and the other does not, the first

8. With sixteen candidates there are $2^{16} = 65,536$ possible messages. With the majority judgment there are 6^{16}, or some 2.8 trillion, possible messages.

9. For a different analysis of this experiment, see Laslier and van der Straeten (2004).

Table 18.3c
Percentages of Same Votes to Two Candidates, Approval Voting Results, five precincts of Orsay and Gy-les-Nonains, First-Round, French Presidential Election, April 21, 2002

	Jospin	Chirac	Bayrou	Mamère	Chevènement	Le Pen
Jospin	–	34%	44%	75%	56%	48%
Chirac	34%	–	66%	51%	54%	64%
Bayrou	44%	66%	–	55%	60%	61%
Mamère	75%	51%	55%	–	52%	61%
Chevènement	56%	54%	60%	52%	–	54%
Le Pen	48%	64%	61%	61%	54%	–

Note: Only the more important candidates are shown. "Same votes" means both crosses or both no crosses.

is given 1 vote, the second 0. The first estimate gives $\frac{1}{2}$ vote to each candidate if both have crosses or neither do; this interprets crosses and no crosses as expressing a voter's indifference between them. This yields the estimate Chirac 61%, Le Pen 39%.

The second estimate gives $\frac{1}{2}$ vote to each if both have crosses, otherwise 0; this interprets crosses as expressing a voter's indifference between them and no crosses as saying nothing. This yields the estimate Chirac 79%, Le Pen 21%.

The last estimate gives no vote to each if both have crosses or both do not; no indifference is expressed. This yields the estimate Chirac 80%, Le Pen 20%.

None of these estimates comes close to the actual result in the six voting precincts. Several crosses on a voter's approval ballot—and even more so, several no crosses—does not mean the voter is indifferent among the corresponding candidates. This shows that the approval voting mechanism does not induce the voters to correctly express their preferences or their indifferences.

Again, the problem is that crosses mean different things to different people. In this experiment, and more generally wherever approval voting has been used, it appears to be a mechanism that simply adds crosses: implicitly the vote is relative; it asks voters to make pair-by-pair comparisons. The evidence confirms that this invites strategic voting.

The crosses, it turns out, were used in the same way by the voters. Not only are the average numbers of crosses per ballot about the same across all six precincts but so also are the distributions of the number of crosses per ballot (table 18.3d).

This does not, however, imply that the two options constituted a common language of absolute grades because use includes strategic behavior, and perhaps what is in common is the strategic behavior. The point is, If voters assign

Table 18.3d
Percentages of Ballots with k Crosses ($k = 0, 1, \ldots, 16$), Approval Voting Experiment, Five Precincts of Orsay and Gy-les-Nonains, First Round, French Presidential Election, April 21, 2002

	No. of Crosses										
	0	1	2	3	4	5	6	7	8	9–16	Average
All precincts	1.4	11.1	22.0	30.3	19.0	10.0	3.6	1.5	0.6	0.5	3.15
Gy-les-Nonains	3.3	13.2	25.3	28.0	17.6	8.0	1.9	1.6	0.8	0.3	2.90
Orsay 1st	0.5	10.6	21.5	31.1	19.1	9.3	4.9	1.5	0.7	0.7	3.25
Orsay 5th	1.7	9.9	22.5	30.7	20.4	8.8	4.6	0.8	0.6	0.0	3.11
Orsay 6th	1.1	11.4	20.4	29.8	17.9	12.3	4.4	1.8	0.7	0.4	3.22
Orsay 7th	0.6	10.7	20.9	30.1	21.8	9.8	2.4	2.4	0.4	0.9	3.22
Orsay 12th	1.5	11.1	22.0	31.7	16.7	11.4	3.4	1.2	0.5	0.5	3.13

crosses because of absolute evaluations of the merits of candidates, then the language is common; otherwise, the language is not common. If the behavior is absolute, Arrow's paradox cannot arise; if it is relative, the paradox can arise. Another experiment that was conducted in 2007 in parallel with the first round of the French presidential election provides data that allow analysis of this issue.

The Illkirch–Louvigny–Cigné (ILC) experiment (Baujard and Igersheim 2007) tested two mechanisms at once—approval voting and a point-summing mechanism with points 0, 1, or 2 (see chapter 17)—in six different voting precincts: three in Illkirch (Alsace), two in Louvigny (Basse-Normandie), and one in Cigné (Mayenne). There were 2,836 participants (62% of those who voted officially). The approval voting ballot stated,

Instructions: You indicate, among the 12 candidates, those that you support. To do so encircle the name of that or those candidates whom you support. You may encircle one name, several names or no name . . . The candidate elected with [this] method is the one who receives the highest number of supports.

The question posed was again neutral. The outcomes over the six precincts are given in table 18.4a. Again no candidate had circles in a majority of the ballots; again the four major candidates all lost relative support in approval voting, whereas every one of the others gained; again the mechanism failed in that the winner's majority-grade was "not support."

However, the experience shows that populations make a remarkably homogeneous use of the means they are offered to express themselves. On average, voters cast 2.33 circles per ballot, and close to the same was true in each of the three towns. Moreover, the distributions of the number of circles per ballot in each of the three towns were very similar as well, so the circles were used in about the same way by all voters (table 18.4b).

Table 18.4a
Approval Voting Experiment, Illkirch–Louvigny–Cigné, French Presidential Election, April 22, 2007

	Percent of Ballots with Circles	Percent of All Circles	Official Vote, First Round
Bayrou	49.7%	21.4%	23.0%
Sarkozy	45.2%	19.4%	34.1%
Royal	43.7%	18.8%	23.6%
Besancenot	23.7%	10.2%	4.1%
Voynet	16.9%	7.3%	2.1%
Le Pen	11.6%	5.0%	7.6%
Bové	11.5%	4.9%	1.1%
Laguiller	9.3%	4.0%	1.0%
de Villiers	9.0%	3.9%	1.7%
Buffet	7.4%	3.2%	0.8%
Nihous	3.4%	1.5%	0.6%
Schivardi	1.4%	0.6%	0.3%

Table 18.4b
Percentage of Ballots with k Crosses ($k = 1, 2, \ldots, 12$), Approval Voting Experiment, Illkirch–Louvigny–Cigné, French Presidential Election, April 22, 2007

	No. of Crosses								
	1	2	3	4	5	6	7	8–12	Average
All	27.3	33.6	25.1	9.8	2.8	0.9	0.5	0.1	2.33
Cigné	30.7	29.3	22.3	13.0	2.8	2.3	0.5	0.0	2.34
Louvigny	23.6	35.1	26.3	10.9	2.7	0.7	0.4	0.2	2.39
Illkirch	29.3	33.2	24.5	8.6	2.8	0.9	0.5	0.1	2.28

The analysis of the absolute versus relative vote issue is based on the considerable information found in the majority judgment ballots of the 2007 Orsay experiment. Since the language is common to random samples of one hundred, fifty, and even twenty voters from the three precincts in Orsay, it is reasonable to hypothesize that the distribution of grades in the 2007 French presidential election is common to all voters anywhere in France (note that the language is common, not the evaluations of the candidates). In the approval voting experiment there were 2.33 circles per ballot. If voting behavior was based on an absolute scale only, then voters would cast circles either for the candidates deemed *Excellent*, or those deemed *Very Good* or better, or *Good* or better, and so on. But (see table 6.1) there are on average 0.69 *Excellent*s, 1.94 *Very Good*s or better, and 3.44 *Good*s or better: none of these comes close to 2.33, suggesting that the behavior is not purely absolute.

Table 18.5
Average Number of highest, second highest, and third highest grades, Three Precincts of Orsay, April 22, 2007

Average No. of Grades	Three Precincts	1st Precinct	6th Precinct	12th Precinct
Highest	1.64	1.51	1.62	1.80
Second highest	2.19	2.08	2.16	2.34
Third highest	2.76	2.73	2.78	2.76

Each majority judgment ballot assigns a grade to every candidate. The highest grade is given to one or more candidates; the second highest to one or more candidates; and so on down the list. Their averages may be computed (table 18.5); they are common to all three precincts as well. If voting behavior was based on a relative scale—assuming these averages are roughly common to all of France—then 2.33 should be about equal to 1.64, or $3.83 = 1.64 + 2.19$, or greater. It is not, suggesting that the behavior is not purely relative.

Behavior in the 2007 approval voting experiment is better explained as a mixture of absolute and relative behavior:

· A voter casts circles for every candidate deemed above a *Good*.

· If the voter deems no candidate above a *Good*, he casts circles for every candidate receiving his highest grade.

This behavior implies an average of 2.26 circles per approval ballot in the three Orsay precincts, an average of 2.09 in the 1st, 2.27 in the 6th, and 2.43 in the 12th. This is in substantial agreement with the 2.33 observed in the 2007 approval voting experiment. Applying this behavior to the majority judgment ballots of the Orsay experiment to simulate an approval vote gives the following percentages of ballots with circles: Bayrou 51.1%, Royal 44.8%, Sarkozy 44.1%, Besancenot 16.8%, Voynet 14.5%, Buffet 11.6%, de Villiers 9.9%, Bové 9.0%, Laguiller 9.0%, Le Pen 8.7%, Nihous 3.2%, Schivardi 2.6%. These percentages are close to those reported in table 18.4a. Of course, the results could also be due to a complex mélange of varied strategic and nonstrategic behavior.

The 2002 Sciences Po experiment (see chapter 17) adds to the evidence that approval votes are not absolute evaluations. Recall that there were fifteen candidates, and one ballot included a point-summing method with points between 0 and 10 and approval voting. The average number of approvals per ballot was 2.76. The ballots make it possible to see when approvals and disapprovals of candidates—crosses and no crosses—correspond to each point

Table 18.6
Distributions: Crosses and No Crosses, and Probability of a Cross, for each point $k = 0, 1, \ldots, 10$ Assigned, Science Po, January 2002

	Points $k =$					
	0	1	2	3	4	5
Percent of Crosses	3.1	1.2	2.7	5.1	5.4	16.2
Percent of No Crosses	58.6	10.6	8.4	7.1	5.8	6.0
$\text{Prob}_k\{\text{Cross}\}$	.05	.10	.24	.42	.48	.73
	6	7	8	9	10	
Percent of Crosses	16.4	17.4	14.9	5.0	12.7	
Percent of No Crosses	2.2	0.6	0.4	0.1	0.2	
$\text{Prob}_k\{\text{Cross}\}$	.88	.97	.97	.98	.98	

assigned to the candidates. And knowing the number of crosses c_k and no crosses nc_k given to candidates who are assigned k points makes it possible to estimate $\text{Prob}_k\{\text{Cross}\} = \frac{c_k}{c_k + nc_k}$, the probability that a cross is given to a candidate assigned k points. The data are shown in table 18.6.

The distribution of the crosses is very widely scattered and clearly shows that they carry no common, absolute meaning. At least the probability of a cross grows regularly as the assigned point k increases. When an approval or cross is assigned to every point $k \geq 5$, the average number of approvals per ballot is 3.44; when to every point $k \geq 6$, the average number of approvals per ballot is 2.26; showing again the crosses' lack of meaning.

A last, quite surprising observation reinforces the idea that voters express relative opinions in approval voting. The 2.33 on average approvals of twelve candidates in the 2007 ILC experiment is an approval rate of 19.4%. The 3.15 on average approvals of sixteen candidates in the 2002 Orsay experiment is an approval rate of 19.7%. The 2.76 on average approvals of fifteen candidates in the 2002 Sciences Po experiment is an approval rate of 18.4%. This 18%–20% rate of approval is incredibly stable. It cannot be that slightly less than one-fifth of the candidates are always *Good* or better independent of their identities.[10] Behavior that sees voters consistently approving of 18%–20% of the candidates suggests that voters are making relative evaluations just as they are asked to do, not absolute evaluations.

10. Table 17.5 gives data showing that the distributions of evaluations change as the lists of candidates vary.

We conclude that the approval voting experiments exhibit behavior in the assignment of crosses or circles that is not absolute. There are two implications. Arrow's paradox cannot be excluded. The experimental evidence supports the theoretical claims. Approval voting as practiced to date is not an instance of the majority judgment with two grades. Rather it is an instance of range-voting—a point-summing method whose scores are given no meaning—with only two scores.

But, of course, it makes little sense to restrict a language of evaluation to two grades.

19 Comparisons of Voting Methods

That each party endeavors to get into the administration of the government, and exclude the other from power, is true, and may be stated as a motive of action: but this is only secondary; the primary motive being a real and radical difference of political principle.
—Thomas Jefferson

The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, an adoption of two strongly contrasted positions between which the voter may choose. Instead, each party strives to make its platform as much like the other's as possible.
—Harold Hotelling

Previous chapters have compared various methods of voting with the majority judgment on the basis of the theoretical properties the methods satisfy or fail to satisfy. The experimental evidence given in this chapter depends entirely on the 2007 Orsay experiment. We believe that the participants by and large expressed their true opinions, for they at once had no incentive not to do so—participation itself in a nonbinding vote was an indication of a will to cooperate—and their input messages were consistent with the expressions of their official votes (as well as other ancillary evidence). Thus the fact that strategic considerations played no role makes it possible to compare the various methods on the basis of practical, realistic and honest expressions of opinion. This could not be done with the results of a real election because in that case some voters will send strategic messages rather than honest ones.

Beginning with a back-of-the-envelope model to contrast the appeal to the center of first-past-the-post and majority judgment, the manipulability and bias in favor of or against centrist candidates is studied experimentally. The results confirm what common sense suggests: first-past-the-post, point-summing, and Borda are the most manipulable; first- and two-past-the-post tend systematically to deny the election of a centrist candidate, whereas Borda and point-summing tend systematically to elect a centrist candidate. A centrist is a candidate situated at the center of the political spectrum, between the leading left and right

parties (and is not necessarily a Condorcet-winner, as the experimental evidence shows). We believe that these results are robust, but more experimentation should be done.

19.1 Bias for the Center

It has often been observed that with the electoral systems commonly practiced, and in particular, first-past-the-post, candidates or parties vie for the center: they wish to express the opinions of the median-voter. The primitive idea is that there is a left-to-right political spectrum, a line on which voters and candidates are situated. A Democratic candidate wishes to occupy a position or point on the line as far as feasible to the right in order to obtain the votes of all those whose opinions are at that point or to the left, and a Republican candidate wishes to occupy a point on the line as far as feasible to the left in order to obtain the votes of all those whose opinions are at that point or to the right. In terms of the French election of 2007, Bayrou was the Condorcet-winner because he occupied a position on the line close to the median-voter's: to the right of the socialist Royal, he would against her obtain all the votes on the right; to the left of the U.M.P. Sarkozy, he would against him obtain all the votes on the left. And in the runoff between the leftist Royal and the rightist Sarkozy, Royal tried to inch toward the right to obtain some moderate rightist votes and all those to the left of them, whereas Sarkozy made overtures to the left to obtain some moderate leftist votes and all those to the right of them.

Beginning with an issue of economic equilibrium when two firms compete and costs of transportation from their production's facilities to the buying public are significant, Hotelling (1929) developed a simple but ingenious back-of-the-envelope model to analyze the situation, which was subsequently developed in detail by Downs (1957) in his famous book. In the language of elections the model's essentials may be described in the following terms. The unit interval $[0, 1]$ is the line that represents the political spectrum; a point is a political position. Voters are uniformly distributed on it. The election game is played by two candidates. Each chooses some point on the line. A voter casts one vote for the candidate nearest to her point, and the candidate with the most votes wins. More generally, voters may be distributed according to some density function f or the corresponding cumulative distribution function F. Downs analyzed a normal distribution.

If a candidate announces a point x that is not that of the median-voter—in the uniform case the point $\frac{1}{2}$, so that one-half of the voters are to the right, the other half to the left (more generally, the point $F^{-1}(\frac{1}{2})$)—she is sure to lose the election. For if $x < \frac{1}{2}$ and the opponent chooses any point y in the

interval $(x, \frac{1}{2})$, the opponent obtains a majority; and symmetrically, if $x > \frac{1}{2}$, the opponent chooses $y \in (\frac{1}{2}, x)$. Thus the unique equilibrium is for both players to announce the political position of the median opinion, or, as Hotelling claimed, "each party strives to make its platform as much like the other's as possible." With the standard majority mechanism, the set of political offers reduces to one position, to the detriment of the voters, many of whose ideas and hopes are never defended, discussed, or implemented.

Pursue, now, the same simplicity of model, but suppose that voters give grades to the candidates. A voter at point x gives to a candidate at point y the grade $1 - |x - y|$. Thus if $x = y$, the grade is the highest possible, namely 1, and if $x = 0$ and $y = 1$ (or $x = 1$ and $y = 0$), the grade is the lowest possible, namely 0. Two different games will be played where the candidates choose points on the line. In one the winner is the candidate with the highest majority-grade; in the other it is the candidate with the highest average grade (the point-summing or range-voting winner). In either case, the optimal strategy of a candidate is to choose a position that maximizes the possibility of winning.

With the average grade mechanism, a candidate chooses y to maximize

$$\int_0^1 (1 - |x - y|)dx = 1 - \int_0^y (y - x)dx - \int_y^1 (x - y)dx = 1 - \frac{y^2}{2} - \frac{(1 - y)^2}{2},$$

so $y = \frac{1}{2}$. Once again, the electoral competition is reduced to the point of the median-voter. This is true with the uniform distribution; more generally, y should maximize $\int_0^1 (1 - |x - y|) f(x)dx$ but is again the median of the distribution $F^{-1}(\frac{1}{2})$.

With the majority-grade mechanism the result is entirely different. Suppose a candidate chooses a point $y \in \left(0, \frac{1}{4}\right)$. The higher half of the grades are given by voters at points $x \leq \frac{1}{2}$, and the lowest grade of that half by the voter at $x = \frac{1}{2}$. The lowest is the majority-grade, which is $1 - \left(\frac{1}{2} - y\right) < \frac{3}{4}$. Similarly, when the position of the candidate is $y \in \left(\frac{3}{4}, 1\right)$, his majority-grade is also less than $\frac{3}{4}$. If, however, $y \in \left[\frac{1}{4}, \frac{3}{4}\right]$, then all of the voters in the interval $\left[y - \frac{1}{4}, y + \frac{1}{4}\right]$ (comprising half of the voters) give the candidate at least the grade $\frac{3}{4}$, and at least one gives exactly that grade, so it is the candidate's majority-grade. Thus, here is another Hotelling-type back-of-the-envelope model. It suggests that the majority judgment opens the political spectrum to a diversity of political positions as long as they are not too extreme; this cannot but be better for democracy. The analysis may be pursued with another distribution $F(x)$. The optimal solutions are necessarily in $\left[F^{-1}\left(\frac{1}{4}\right), F^{-1}\left(\frac{3}{4}\right)\right]$, so candidates of the left, the center, or the right may be elected as opposed to the Borda, Condorcet, and point-summing methods, where the optimal solution is always the centrist

candidate $F^{-1}\left(\frac{1}{2}\right)$. However, it may be difficult to compute the majority-grade: for some F's the solution is unique, for others not.

This suggestion is borne out by the empirical evidence. Table 19.1 gives the distributions of the winners in 10,000 random samples from 101 ballots drawn from the 501 representative ballots of the 2007 Orsay experiment according to each of nine different methods. The same qualitative results are obtained when samples are of 201 ballots or when they are drawn from all 1,733 ballots of the Orsay experiment. However, when all the ballots of the experiment are used, the results are very much more favorable to the centrist candidate Bayrou (since he did much better in Orsay than nationally), so whereas the ordering of the methods is the same, the numbers of Bayrou wins are considerably higher and thus give less insight into the differences between methods. Approval judgment when approval means *Excellent* (AJ $\succeq$ *Excellent*), first- and two-past-the-post, and approval judgment when approval means at least *Very Good* (AJ $\succeq$ *Very Good*) are biased against the centrist candidate (in order, decreasingly). Condorcet, point-summing, approval voting when approval means at least *Good* (AJ $\succeq$ *Good*), and Borda are biased in favor of the centrist candidate (in order, increasingly). The majority judgment stands somewhere in the middle: although Sarkozy wins most often, each of the three main political forces has a reasonable chance of winning.

One method's bias in favor of the centrist candidate is highly dependent on the number of candidates, namely, Borda's. It is already biased with three candidates, but with twelve it is overwhelmingly so. Observe also that the only two methods that are chaotic in that they are sensitive to the number of candidates are the two sum-scoring methods, first-past-the-post and Borda's. Moreover, Saari's claim that Borda's method is the least chaotic of the sum-scoring methods is put into doubt by the experimental evidence. The explanation is that Saari's analysis assumes "impartial culture," meaning that all preference-orders are equally likely, which is simply not true in practice. What is observed in practice is a statistical left-right spectrum. All the other methods are stable with regard to the number of candidates. Approval judgment when approval means at least *Acceptable* or at least *Poor* is not recommended: imagine a candidate elected because he has the largest number of at least *Acceptable* or at least *Poor*! The Australian alternative vote would have given the same results as two-past-the-post with three candidates (Bayrou, Royal, Sarkozy); it is biased against the center but is less chaotic than first-past-the-post or Borda.

Table 19.2 only concerns the number of times the centrist candidate (Bayrou) wins. It confirms that the methods are ordered from least to most in favor of the centrist candidate. Take any row, for example, the majority judgment row. Bayrou wins 4,037 times; of those, the number of his wins with the preceding

Table 19.1
Number of Wins among Royal, Bayrou, and Sarkozy, 2007 Orsay Experiment

	3 Candidates			
	Left ⟵ Royal	Bayrou	⟶*Right* Sarkozy	Tie
AJ ≥ *Excellent*	557	7	9,190	246
First-past-the-post	1,713	46	8,092	149
Two-past-the-post	2,137	883	6,555	425
AJ ≥ *Very Good*	1,303	1,304	6,703	690
Majority judgment	1,317	3,982	4,690	11
Condorcet	679	6,519	1,975	473
Point-summing	854	7,859	1,093	194
AJ ≥ *Good*	390	8,828	454	328
Borda	483	7,201	2,154	162
	12 Candidates			
	Left ⟵ Royal	Bayrou	⟶*Right* Sarkozy	Tie
AJ ≥ *Excellent*	508	3	9,238	251
First-past-the-post	2,112	48	7,824	16
Two-past-the-post	2,174	764	6,675	387
AJ ≥ *Very Good*	1,277	1,316	6,753	654
Majority judgment	1,321	4,037	4,631	11
Condorcet	663	6,552	1,972	436
Point-summing	859	7,875	1,090	176
AJ ≥ *Good*	380	8,801	463	356
Borda	377	9,592	21	10

Note: Ten thousand samples of 101 ballots, which were drawn from a sample of 501 ballots representative of the national vote. The percentage estimates of the first-round results on the basis of these 501 representative ballots come close to those of the official national vote. In these 501 ballots Sarkozy had 30.7%, Royal had 25.5%, Bayrou had 18.7% and Le Pen had 9.3%.

Recall that in every sample drawn and when all twelve candidates are present, one of Royal, Bayrou, or Sarkozy is always the winner.

The Condorcet paradox occurred in 354 cases among three candidates and in 377 cases among twelve candidates.

methods are all substantially lower for they are less favorable to the centrist; of those, the number of his wins with the succeeding methods are all quite close to 4,037 (they cannot, of course, be higher) for they are more favorable to the centrist. This is true of every method, every row of the table.

19.2 Manipulability

Every method is manipulable; there is no escape from the possibility that some voters (or judges) will try to game the vote. The operational question is, What

Table 19.2
Number of Wins of the Centrist Candidate (Bayrou), 2007 Orsay Experiment

	FPP	TPP	AJ $\succeq VG$	MJ	Condorcet	PS	AJ $\succeq G$	Borda
First-past-the-post	**48**	48	41	45	48	48	48	46
Two-past-the-post	48	**764**	276	429	764	741	741	762
AJ $\succeq$ *Very Good*	48	741	**1,316**	1,313	1,270	1,300	1,287	1,305
Majority judgment	45	429	1,313	**4,037**	3,655	3,879	4,001	3,994
Condorcet	48	764	1,270	3,655	**6,552**	6,223	6,310	6,523
Point-summing	48	741	1,300	3,879	6,223	**7,875**	7,604	7,862
AJ $\succeq$ *Good*	48	741	1,287	4,001	6,310	7,604	**8,801**	8,640
Borda	48	762	1,305	3,994	6,523	7,862	8,460	**9,592**

Note: FPP = first-past-post; TPP = two-past-post; AJ = approval judgment; *VG* = *Very Good*; MJ = majority judgment; PS = point-summing; *G* = *Good*. AJ $\succeq$ *Excellent* is not included because Bayrou only won three times (see table 19.1).

Ten thousand samples of 101 ballots, which were drawn from a sample of 501 ballots representative of the national vote.

A diagonal entry (boldface) in the row of a method M is the number of Bayrou wins with method M; an off-diagonal entry is the number among those wins that he wins with another method (e.g., with Condorcet's method Bayrou won **6,552** times; in 764 of those, he also won with two-past-post).

method or methods best resist manipulation? The theoretical reasons that the majority judgment dominates all other known methods on this ground have already been proven. So, what may happen in practice?

In the 2007 Orsay experiment, Bayrou was the winner and Royal the runner-up, with respective majority-gauges

$$\text{Bayrou} : (44.3\%, \textit{Good}+,\ 30.6) \qquad \text{Royal} : (39.4\%, \textit{Good}-,\ 41.5\%).$$

Could those voters who gave a higher grade to Royal than to Bayrou have manipulated the outcome to make Royal the winner by raising Royal's grades *or* lowering Bayrou's (for the majority judgment is partially strategy-proof-in-ranking, so a voter cannot contribute to lowering Bayrou's majority-gauge *and* to raising Royal's majority-gauge)? The answer is yes, they could have, if all or a great many of them had exaggerated the grades they had assigned. The evidence shows, however, that this is unrealistic. The polls suggest that in the official election at most some 30% of the voters voted strategically not in accord with their true beliefs.

Consider the data given in table 19.3, where B stands for Bayrou and R for Royal. The table shows the percentages of all the ballots that graded Royal above Bayrou by types of grades.

· *Type VII* 9.2% of all ballots gave to Royal an *Excellent* or a *Very Good* and to Bayrou an *Acceptable* or worse; they are unable to change either candidate's majority-gauge. However, with a point-summing method, those

Table 19.3
Types of Ballots That Grade Royal above Bayrou, 2007 Orsay Experiment

Type of Ballot	Percent of All Ballots	*Excellent*	*Very Good*	*Good*	*Acceptable*	*Poor*	*To Reject*	Strategy
I	2.8%				R	B	B	0%
II	6.3%	R	B→	→	→ |			33%
III	6.9%		R	B→	→ |			33%
IV	2.4%		| ←	←R	B			33%
V	3.2%	R		B→	→ |			67%
VI	2.1%		| ←	←R		B	B	67%
VII	9.2%	R	R		B	B	B	Cannot

who gave Royal a *Very Good* could have increased her point total, and those who gave Bayrou an *Acceptable* or a *Poor* could have decreased his point total.

· *Type III* 6.9% of all ballots gave to Royal a *Very Good* and to Bayrou a *Good*; they are unable to increase Royal's majority-grade but can decrease Bayrou's. Yet those who are able to decrease Bayrou's have no incentive to decrease it below *Acceptable* (indicated by the vertical line) because doing so will not further decrease his majority-gauge. However, with a point-summing method, they could increase Royal's point total, and they would have the incentive to give Bayrou the lowest possible grade.

· *Type IV* 2.4% of all ballots gave to Royal a *Good* and to Bayrou an *Acceptable*; they are able to increase Royal's majority-gauge but are unable to decrease Bayrou's. Yet those who are able to increase Royal's have no incentive to increase her grade above *Very Good* (indicated by the vertical line) because doing so will not further increase her majority-gauge. However, with a point-summing method, they would have the incentive to give Royal the highest possible grade, and they could decrease Bayrou's point total.

Consider a scenario where voters who rate Royal merely *Acceptable* do nothing, one-third change who give Royal one grade above Bayrou, and two-thirds change who give Royal two grades above Bayrou.

· Type I voters change nothing.

· 33% of types II, III, and IV voters change (as indicated by the arrows in table 19.3).

· 67% of types V and VI voters change (as indicated by the arrows).

Then the two respective candidates' majority-gauges become

Bayrou : (42.2%, *Good*+, 37.1) Royal : (41.6%, *Good*+, 41.5%),

and the manipulation fails to change the outcome, Bayrou remains ahead of Royal. Assigning 5 points to *Excellent* down to 0 points for *To Reject* and applying the same scenario, the point-summing method makes Royal the winner (and note that type VII voters are able to manipulate).

More evidence concerning the manipulability of methods may be found using the database of the 1,733 ballots of the 2007 Orsay experiment. Sample problems of 201 ballots were drawn at random, and two different scenarios of manipulation were tested (samples of 101, 201, and 301 were taken in this and the other simulations; they consistently give concordant qualitative results). To test any one method, restrict the sample problems to those that have an unambiguous winner A and runner-up B. Then change the messages of voters in accordance with two scenarios:

- *Scenario 1* 30% of those voters who gave a higher grade to B than A (chosen at random) change to give B the highest and A the lowest possible grades.
- *Scenario 2* All those voters who gave a grade to B two levels above A change to give B the highest and A the lowest possible grades.

The manipulation is successful if after the change A is no longer the winner.

These two types of manipulation will have qualitative impacts that differ depending on the method of voting. With the majority judgment, or its specialization to any approval judgment, the majority-gauge of any candidate C other than A or B remains the same; and with point-summing, the final point total of such a candidate does not change either. Thus the only possible winners after the change are A or B.

With Borda's method, A is placed at the bottom and B at the top of the list of every voter who manipulates. Thus A's Borda-score decreases, B's increases, as may also the Borda-score of another candidate C; however, C's score cannot increase more than B's, so once again the only possible winners after the change are A or B.

With first-past-the-post, those voters who gave a higher grade to B than to A gave no vote to A, so A's total vote cannot be reduced and stays the same. If B was the one candidate evaluated *Excellent* by the voter who wishes to manipulate, he cannot; if B was among those evaluated highest by the voter who wishes to manipulate, he can only do so if that highest grade is not *Excellent*. Any other candidate C's total vote can only stay the same or decrease. So the only possible winners after the change are A and B. Two-past-the-post cannot be manipulated with these scenarios; the argument just given means that A and B necessarily survive the first round, so A can only remain the unambiguous winner.

With Condorcet the situation is different. Since A and B are unambiguously the top two candidates, there cannot be a Condorcet-cycle among the top

Table 19.4
Number of Successful Manipulations in Ten Thousand samples of 201 Ballots, 2007 Orsay Experiment

Grades	AJ $\succeq E$	PS	Borda	FPP	AJ $\succeq VG$	Condorcet	AJ $\succeq G$	MJ
			Scenario 1, Drawn from 501 Representative Ballots					
Six	9,154	9,962	9,522	6,000	8,601	7,075	8,763	6,820
Six±	9,166	9,999	9,949	7,941	8,656	7,957	8,778	6,958
			Scenario 2, Drawn from 501 Representative Ballots					
Six	9,779	9,282	7,162	6,607	7,373	5,501	4,489	7,072
Six±	9,750	9,963	9,355	8,769	7,309	7,200	4,358	7,112
			Scenario 1, Drawn from All (1,733) Ballots					
Six	9,935	9,463	8,699	7,411	7,826	4,392	6,166	3,857
Six±	9,940	9,972	9,721	8,748	7,860	7,669	6,167	3,878
			Scenario 2, Drawn from All (1,733) Ballots					
Six	9,924	6,504	5,389	7,236	4,050	2,048	1,538	2,591
Six±	9,924	9,368	8,074	8,579	4,050	5,426	1,538	2,591

Note: AJ = approval judgment; E = *Excellent*; PS = point-summing; FPP = first-past-post; VG = *Very Good*; G = *Good*; MJ = majority judgment.

finishers (they are always among Bayrou, Royal, and Sarkozy).[1] A necessarily retains a majority against B after the change; and B necessarily retains a majority against any other candidate C after the change. However, it may be that lowering A's grades implies that some other candidate C has a majority over A after the change. In this case a Condorcet-cycle (namely, $A \succ_S B \succ_S C \succeq_S A$) has been produced by the change, and there is no winner. This is a successful manipulation.

The 2007 Orsay ballots have six grades, but there are twelve candidates, so often candidates have identical grades. This artificially limits the possible manipulation of the methods that depend on comparisons of candidates (Borda, first-past-the-post, and Condorcet). To lift this constraint a second experiment was conducted in which extra highest and lowest grades, *super-Excellent* and *worse-than-To-Reject*, are added to the usual set of six grades (referred to as "Six" in the tables) for the purposes of unequivocal manipulation (the eight grades are referred to as "Six±" in the tables).

Table 19.4 shows that majority judgment and approval judgment when approval means at least *Good* (AJ $\succeq G$) consistently best combat manipulation.

1. Condorcet-cycles among the nine other candidates are ignored because they are not germane to this analysis.

Table 19.5
Successful Manipulations, Majority Judgment Compared with Other Methods, 2007 Orsay Experiment

	Scenario 1			
	501 Ballots		All (1,733) Ballots	
	Six Grades	Six ± Grades	Six Grades	Six ± Grades
Majority judgment	77%	77%	39%	55%
AJ > *Excellent*	89%	90%	99%	99%
Majority judgment	51%	50%	33%	33%
Point-summing	100%	100%	96%	100%
Majority judgment	48%	50%	31%	40%
Borda	96%	100%	89%	98%
Majority judgment	78%	77%	40%	40%
First-past-the-post	51%	77%	58%	77%
Majority judgment	75%	77%	33%	40%
AJ > *Very Good*	88%	89%	76%	76%
Majority judgment	56%	59%	34%	35%
Condorcet	47%	59%	33%	72%
Majority judgment	53%	55%	54%	33%
AJ > *Good*	83%	84%	59%	59%

Note: AJ= approval judgment.
When winner and runner-up are the same candidates with both methods, the percentages shown are for successful manipulations of each.

The serious drawback of approval judgment ($\succeq$ *Good*) is that it is very biased in favoring centrist candidates (see tables 19.1 and 19.2). When approval judgment changes and means setting the threshold for approval at higher grades, it becomes more manipulable and less biased in favor of (or more biased against) the centrist candidate. It should not be surprising that Condorcet is difficult to manipulate because the only possible successes result from the creation of Condorcet-cycles. There are more successful manipulations when the samples are drawn from the 501 representative ballots than from all 1,733 ballots because in the first case the race among the three principal candidates was much closer than in the second case. Indeed, in the second case Bayrou's majority judgment ballots almost stochastically dominate those of Royal and Sarkozy: Bayrou leads in all cumulative grades except *Excellent* (see table 15.4).

Table 19.5 compares majority judgment with each of the other methods when both methods agree on the winner and the runner-up. For instance, in scenario 1, with samples drawn from all ballots, the comparisons made with the grades Six±, majority judgment admitted successful manipulation in 33% of

Table 19.5
(cont.)

	Scenario 2			
	501 Ballots		All (1,733) Ballots	
	Six Grades	Six ± Grades	Six Grades	Six ± Grades
Majority judgment	82%	81%	45%	45%
AJ > *Excellent*	95%	96%	95%	95%
Majority judgment	50%	49%	18%	18%
Point-summing	92%	100%	63%	93%
Majority judgment	41%	42%	17%	17%
Borda	73%	95%	56%	82%
Majority judgment	85%	84%	32%	32%
First-past-the-post	57%	87%	47%	71%
Majority judgment	48%	48%	25%	25%
AJ > *Very Good*	76%	74%	32%	32%
Majority judgment	62%	61%	19%	19%
Condorcet	29%	48%	11%	46%
Majority judgment	85%	85%	18%	18%
AJ > *Good*	38%	38%	13%	13%

the cases, and point-summing admitted successful manipulation in 100% of the cases.

19.3 Conclusion

The evidence of this chapter helps to clarify fundamental differences among methods of voting.

· The methods most championed by reformists—Borda's and point-summing—are extremely biased in favor of centrist candidates and are also among the most manipulable. Moreover, Borda's is the most chaotic.

· The widely used methods—notably, first- and two-past-the-post—are extremely biased against centrist candidates. Moreover, first-past-the-post is very manipulable (the manipulability of two-past-the-post could not be compared with the others in this experiment).

· The alternative vote with twelve candidates elects in this experiment the same candidate as two-past-the-post with the three major candidates; as a consequence, it is extremely biased against centrist candidates.

· Approval judgment AJ $\succeq$ *Excellent* and AJ $\succeq$ *Very Good* are very biased against the centrist candidate and are highly manipulable.

The least manipulable methods are majoritarian: Condorcet, approval judgment AJ $\succeq$ *Good*, and the majority judgment.

· AJ $\succeq$ *Good* is hard to manipulate in this election because the winning candidates won with a majority-grade *Good*. When the highest majority-grade in an election is, say, *Very Good*, AJ $\succeq$ *Good* or AJ $\succeq$ *Excellent* would have been manipulable, whereas AJ $\succeq$ *Very Good* would be less manipulable.

· However, AJ $\succeq$ *Good* is very biased in favor of the centrist candidate.

· Condorcet's method is far more biased in favor of the centrist than the majority judgment; when it is manipulable, there is no winner, so some other ancillary rule must be invoked; and, as discussed in chapter 20, when voters are strategic, the final result does not reflect the true opinions of voters.

· The majority judgment is one of the least manipulable methods and seems to be the most balanced with regard to the left-right spectrum.

When all or many of the voters are strategic, and their utilities depend *only* on the winner, *all* methods are manipulable. Chapter 20 addresses this hypothesis.

20 The Game of Voting

Rationality is only one of several factors affecting human behavior; no theory based on this one factor alone can be expected to yield reliable predictions.
—Robert J. Aumann

The analysis and comparison of methods, traditional and new, and thus the evaluation of which are best, depend on *context*. Different contexts are encountered in theory and in practice. Often, in debates or arguments, when in one context a method satisfies the criteria that are sought, willy-nilly it is attacked in terms of another context. To be coherent, arguments should compare methods in each of the several possible contexts separately. In the various contexts discussed so far, the majority judgment has been shown, we believe, to dominate the other methods in that it comes closest to meeting the important desirable properties, and avoiding the undesirable ones, that have been identified across the years. Nothing has been said, however, concerning the context raised in chapter 18, which began with the pioneering article of Myerson and Weber (1993), where voting is viewed as a game and outcomes as equilibria, preferably reached via a sequence of best responses by the voters in view of their utilities and the information at hand. In the words of Myerson and Weber, "Voting equilibria based on polls are somewhat analogous to competitive equilibria based on prices in economic models."

The aim of this chapter is to show that the majority judgment does at least as well if not better than the other methods in this context as well. It must be recognized at the outset, however, that the standard assumptions in the literature on this context leave much to be desired. As the economists George Akerlof and Robert Shiller stated (recalling John Maynard Keynes's famous expression), "Insofar as animal spirits exist in the everyday economy, a description of how the economy really works must consider those animal spirits" (2009, 5). The animal spirits of voters are often ignored; voters are viewed as purely rational agents who seek to maximize their expected utilities, and rationality is assumed

to mean that their utilities depend *only* on the identity of the elected candidate. In this case the Condorcet-winner looms large in the equilibrium analysis. But this transposition of economic theorizing into elections is patently false. Voters have much more complicated (unknown) utilities that for many undoubtedly include an honest expression of their opinions. We know firsthand of many sophisticated voters—economists and social choice theorists—who, in the 2007 French presidential election, preferred Bayrou to Sarkozy, were sure Sarkozy would be in the second round and that in the second round he would defeat Royal and be defeated by Bayrou, and yet voted for Royal in the first round.

Fundamentally, a method should not be considered good just because purely *strategic* behavior—perhaps in total violation of some voters' most cherished hopes and wishes—yields equilibria that satisfy certain properties. It is the contrary that should be sought. A method should elicit the honest expression of voters' opinions as inputs, for the aim of an election is to produce outputs that represent as best as possible the true wishes of societies and of juries.

20.1 Equilibria

Throughout it is assumed that there are at least $n \geq 3$ voters or players. A voter of the game chooses a *strategy* t in a set Θ (which is the same for all players). The strategies depend on the underlying model, which may be a rank-order of candidates, grades of a common language assigned to candidates, points assigned to candidates, or indeed any abstract set. A *strategy-profile* is a $(t_1, \ldots, t_n) \in \Theta^n$, where t_i is the strategy of player i. A *mechanism* is a function that associates to each strategy-profile a winning candidate. Each voter or player i has a utility function u_i that depends exclusively on the identity of the winner.

A candidate X is a *Nash-equilibrium winner* for a given method of election if there exists a strategy-profile for which X is the winner, and for any other candidate Y and any voter i with utility $u_i(Y) > u_i(X)$, i cannot make Y the winner by unilaterally changing his strategy. A method of election admits *no veto power* (Maskin 1999) if for any candidate X there exists a strategy $t \in \Theta$ that when used by $n - 1$ voters assures the election of X.

Theorem 20.1 (Folk Theorem) *For any method that admits no veto power, every candidate is a Nash-equilibrium winner.*

Proof Take any candidate X, and suppose all n voters use the strategy t that permits any $n - 1$ of them to elect X. If any voter deviates, the remaining $n - 1$ still elect X. ∎

Thus, the concept of a Nash-equilibrium winner is, with no further refinement, of little use. This result has given rise to a large literature that seeks refined concepts of equilibria. For the most part the results are negative, many equilibria remain, and the different concepts make it difficult if not impossible to compare the relative merits of the different results. A focus of recent interest is the question, When is there a strong-equilibrium winner? (Aumann 1959), meaning that no coalition of voters can deviate from their strategies and thereby elect a preferred candidate (e.g., Sertel and Sanver 2004). This concept is particularly germane to elections because voters talk among themselves (orally and via the Internet), belong to political parties (which sometimes give careful instructions for exactly how to vote, as in Australia), and have access to large amounts of information (much of it common), including repeated opinion polls. Thus sets of voters may often adopt the same strategies.

The central fact is that when any reasonable method of election of the traditional or the new model is formulated as a game—with the inputs the strategies and voters' utilities devoid of animal spirits—*only* the Condorcet-winner can be a strong-equilibrium outcome.

A method of election is *majoritarian* if for any candidate X and any strict majority of voters, there exists a common strategy $t \in \Theta$ for each of them such that whatever the strategies of the others, X is elected. A candidate X is a *strong-equilibrium winner* for a given method of election if there exists a strategy-profile for which X is the winner, and for any other candidate Y and any coalition of voters $i \in \mathcal{S}$ with utilities satisfying $u_i(Y) > u_i(X)$, the voters of $\mathcal{S}$ cannot make Y the winner by together changing their strategies. A candidate C is a *Condorcet-winner* if there is no candidate X strictly preferred to C by a majority of the voters (i.e., their utilities satisfy $u_i(X) > u_i(C)$).

Theorem 20.2 *For any majoritarian method, a candidate is a strong-equilibrium winner if and only if the candidate is a Condorcet-winner.*

Proof Let C be a Condorcet-winner, and suppose all voters use the strategy t that permits a majority of voters to elect C. Then any coalition of voters who strictly prefer another candidate X to C do not constitute a majority and so cannot elect X. Conversely, let X be a strong-equilibrium winner, and suppose she is not a Condorcet-winner. Then there is another candidate Y strictly preferred by a majority of voters to X. Together they can deviate and elect Y, implying X is not a strong-equilibrium winner, a contradiction. This slightly strengthens a statement by Sertel and Sanver (2004). ■

A method of election is *weakly majoritarian* if for any candidate X and any strict majority of voters $\mathcal{K}$ there exists a strategy-profile $t_{\mathcal{K}} \in \Theta^{\mathcal{K}}$ for the players $\mathcal{K}$ such that whatever the strategies of the others, X is elected.

Theorem 20.3 *Only a weakly majoritarian method can always implement the Condorcet-winner as a strong-equilibrium winner.*

Proof Suppose the contrary. Then there is a candidate X and a majority of voters $\mathcal{K}$ that cannot guarantee the election of X.

X is clearly the Condorcet-winner when all the $\mathcal{K}$ voters prefer X to all other candidates, and all voters not in $\mathcal{K}$ prefer all other candidates to X. Suppose X is a strong-equilibrium winner. By hypothesis, there must be a deviation of voters outside of $\mathcal{K}$ that can elect another candidate. Since all those voters least prefer X, this change is beneficial to each of them, showing that X cannot be a strong-equilibrium winner. ∎

The methods of Condorcet, first- and two-past-the-post, approval voting, single transferable vote, majority judgment, and point-summing are all majoritarian methods and thus weakly majoritarian. But Borda's is not, as the following theorem shows. Recall that a monotonic sum-scoring method is a sum-scoring method whose scores $s = (s_1, \ldots, s_m)$ from highest to lowest are strictly monotonic decreasing $(s_1 > s_2 > \cdots > s_m)$.

Theorem 20.4 *Monotonic sum-scoring methods are not weakly majoritarian.*

Proof For simplicity, consider only three candidates, and $s = (1, a, 0)$, where $0 < a < 1$. Consider the profile

$$(1)\ n + n' : A \succ B \succ C \qquad (2)\ n : B \succ C \succ A,$$

where n, n' are chosen to be positive and to satisfy $an > (2-a)n'$. The Condorcet-winner is A. However, type 1 voters are unable to enforce the election of A. For if they were able to, then they must be able to by always placing A in the first position. Thus they would be able to by choosing the strategy-profile

$$\lambda(n+n') : A \succ B \succ C \qquad (1-\lambda)(n+n') : A \succ C \succ B,$$

for an $\lambda \in [0, 1]$. If $\lambda \geq \frac{1}{2}$ and type 2 voters counter with

$$n : B \succ C \succ A,$$

A's score is $n + n'$ and B's is at least $\frac{a}{2}(n+n') + n$, so B's is higher. If, on the other hand, $\lambda < \frac{1}{2}$ and type 2 voters counter with

$$n : C \succ B \succ A,$$

A's score is $n + n'$ and C's is at least $\frac{a}{2}(n + n') + n$, so C's is higher. Thus, in any case, type 1 voters cannot make A the winner. ■

Thus Borda's rule and the other monotonic sum-scoring methods cannot always implement a Condorcet-winner as a strong-equilibrium winner.

More is true. A method of election is *best response majoritarian* [1] if for any candidate X and any strategy of a minority, the majority always has a best response strategy that elects X. Any reasonable method, including Borda's and sum-scoring methods, is a member of this class.

Theorem 20.5 *If a best response majoritarian method has a strong-equilibrium winner, that candidate must be the Condorcet-winner.*

Proof Let X be a strong-equilibrium winner of a best response majoritarian method, and suppose X is not a Condorcet-winner. Then a majority prefers some other candidate Y to X. Each of its voters could respond by announcing the strategy-profile that elects Y. But then X is not the strong-equilibrium winner, a contradiction. ■

Thus any reasonable method can only elect the Condorcet-winner as a strong-equilibrium winner. But for the Condorcet-winner to be elected, the method must be weakly majoritarian. This excludes Borda's method and all monotonic sum-scoring methods.

20.2 Honest Equilibria

Inherent in the proofs of the last section is that for most methods when the utilities of voters depend only on the identity of the winner, there are a huge number of strong-equilibria strategy-profiles that elect the Condorcet-winner. In practice, the most likely of the equilibria comprise cases when the voters express themselves as honestly as possible. Are there equilibria in which all or most of the voters express themselves honestly? This is important because the outcomes of elections should come as close as possible to reflecting the true opinions of voters. It will be seen that there is no method, in either the traditional or the new model, that can guarantee the election of the Condorcet-winner with the honest opinions of the voters. On the other hand, the middlemost mechanisms, notably, the majority judgment, admit equilibria where at least a majority of the grades given each candidate are honest.

1. See Sertel and Sanver (2004), who give a similar definition and note that Borda is not majoritarian but is best response majoritarian.

Theorem 20.6 *No single-valued mechanism in the traditional model guarantees the election of the Condorcet-winner as a strong equilibrium when voters' rank-orders are honest.*

Proof This theorem and its proof are due to Brams and Sanver (2006). Take any Condorcet-consistent method in the traditional model and consider the five-voter "strategy-profile"

(1) $2{:}A \succ D \succ B \succ C$ (2) $2{:}B \succ D \succ C \succ A$
(3) $1{:}C \succ A \succ B \succ D$.

It has no Condorcet-winner. It must nevertheless designate a single winning candidate. Whichever candidate it designates as winner, the implication is that a Condorcet-winner is not a strong-equilibrium winner, as the following argument shows.

Suppose that the winner of the "strategy-profile" is A but the true profile is the same except that the two type 1 voters have the preferences $A \succ C \succ D \succ B$. Then C is the Condorcet-winner. But if these two voters switch to their "strategy-profile," A is elected, which is better for them. Suppose, then, that the winner of the "strategy-profile" is either B or D but the true profile is the same except that the two type 2 voters have the preferences $B \succ D \succ A \succ C$. Then A is the Condorcet-winner. But if these two voters switch to their "strategy-profile," B or D is elected, which is better for them. Suppose, finally, that the winner of the "strategy-profile" is C but the true profile is the same except that the type 3 voter has the preference $C \succ D \succ A \succ B$. Then D is the Condorcet-winner. But if this voter switches to her "strategy-profile," C is elected, which is better for her. ■

The same is true in the new model.

Theorem 20.7 *No social ranking function* (SRF) *in the new model (with at least four grades) guarantees the election of a Condorcet-winner when voters' grades are honest.*

Proof For simplicity, take four grades 3, 2, 1, and 0 (from best to worst). It suffices to compare the two profiles:

	k judges	1 judge	k judges
X:	$1, \ldots, 1,$	$0,$	$3, \ldots, 3$
Y:	$2, \ldots, 2,$	$3,$	$0, \ldots, 0$

and

	k judges	1 judge	k judges
X :	$1, \ldots, 1,$	$0,$	$3, \ldots, 3$
Y :	$0, \ldots, 0,$	$3,$	$2, \ldots, 2$

In the first the Condorcet-winner is Y with a majority of $k+1$; in the second, the Condorcet-winner is X with a majority of $2k$. But any SRF ranks X and Y identically in both cases because each has precisely the same set of grades, so Condorcet-consistency is impossible in the new model. ■

Therefore, there is no method of election in either model that elects the Condorcet-winner (when she exists) as a strong-equilibrium winner without cheating. But in the new model the Condorcet-winner may be elected with every candidate receiving a majority of honest grades.

Theorem 20.8 *For any middlemost aggregation mechanism there exists a strong equilibrium that elects the Condorcet-winner C with his true middlemost grade α when voters have no indifferences. Moreover, every candidate is assigned a majority of honest grades.*

Proof Let α be the true middlemost grade of a Condorcet-winner C. To begin, note that α is necessarily above the minimum grade α_{min}. For if not, then a majority of C's grades are α_{min}. The "no indifference" assumption implies that any other candidate B must be preferred by a majority to C, contradicting the fact that C is the Condorcet-candidate.

Construct a strategy-profile as follows:

· All voters who assign C at least α give their honest grades; all others give C the grade α;

· Any voter who assigns a grade α or above to another candidate B and who prefers C to B, gives B a lower grade than α (barely below will do); otherwise voters give candidates other than C their honest grades.

With this strategy-profile, C's middlemost grade is α, the true one. Take B to be any other candidate, and let $\mathcal{S}$ be the coalition of voters who prefer B to C. $\mathcal{S}$ is necessarily a strict minority of voters. The strategy-profile implies that all voters outside $\mathcal{S}$—a strict majority—give to B grades below α, so B's middlemost grade must be below α. Thus C is the winner.

The coalition $\mathcal{S}$—the only voters who wish to make B the winner instead of C—cannot increase B's middlemost grade because the majority outside $\mathcal{S}$ (which prefers C to B) gives grades below α to B; nor can $\mathcal{S}$ decrease C's

Table 20.1
Hypothetical Honest Grades, 1970 New York Senate Election

	Type			
	1a 10%	1b 29%	2a 16%	2b 8%
B	*Excellent*	*Very Good*	*Acceptable*	*Poor*
G	*Poor*	*Good*	*Very Good*	*Very Good*
O	*Acceptable*	*To Reject*	*Good*	*Acceptable*

Note: Total percentages of types 1, 2 and 3 correspond to actual 39% for Buckley, 24% for Goodell, and 37% for Ottinger.

Table 20.2
Strategy Profile That Elects Condorcet-Winner as a Strong Equilibrium, with True Middlemost Grade, 1970 New York Senate Election

	Type			
	1a 10%	1b 29%	2a 16%	2b 8%
B	*Excellent*	*Very Good*	*Acceptable*	*Poor*
G	*Good*	*Good*	*Very Good*	*Very Good*
O	*Acceptable*	*To Reject*	*Good*	*Acceptable*

middlemost grade because the majority outside $\mathcal{S}$ (which prefers C to B) gives at least α to C. So C is a strong-equilibrium winner.[2]

The outcome of this equilibrium enjoys very desirable properties. First, the winner C has his true middlemost grade α. Second, only voters who grade C below α—a strict minority—cheat by increasing the grades they give him. Third, voters who cheat by decreasing another candidate B's grades are only those who grade C strictly above α—a strict minority. Thus, the majority of the grades given any candidate are the true grades of voters. ■

Observe that this strategy-profile depends only on the true middlemost grade of a Condorcet-winner, so dropping or adding a minor candidate does not change the outcome. Also, when there is no indifference (which is likely if the language is rich enough), the true middlemost grade of a Condorcet-winner is necessarily above the minimum grade (contrary to approval judgment).

2. If the language is not rich enough to eliminate indifferences (in the traditional model there are no indifferences by fiat), the equilibrium may be proven by assuming $\alpha \succ \alpha_{min}$, and altering the second part of the strategy-profile by replacing "who prefers C to B" by "who prefers C to B or is indifferent between them."

Table 20.1
(cont.)

	Type		
	3a 22%	3b 15%	Majority-Gauge
B	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)
G	*Acceptable*	*Very Good*	(39%, *Good*, 32%)
O	*Very Good*	*Excellent*	(37%, *Good*, 47%)

Table 20.2
(cont.)

	Type		
	3a 22%	3b 15%	Majority-Gauge
B	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)
G	*Good*	*Very Good*	(39%, *Good*, 0%)
O	*Acceptable*	*Acceptable*	(16%, *Acceptable*, 29%)

Notice also that if instead of using the strategy just described, some voters deviate by giving their honest grades, the Condorcet-winner's majority-gauge decreases and all the other candidates' majority-gauges increase. This implies that if the winner changes, then the new winner must be ranked higher than the Condorcet-winner according to the honest grades of voters. This is a positive property.

To illustrate the idea, consider an example modeled on the 1970 New York Senate election among James Buckley (B), Charles Goodell (G) and Richard Ottinger (O) (see section 18.2), and suppose there are six types of voters and the honest profile (with the usual six grades) given in table 20.1.

Goodell is the Condorcet-winner, and the majority judgment order of finish is Goodell $\succ_S$ Ottinger $\succ_S$ Buckley. The strategy-profile of the proof is given in table 20.2 together with the corresponding majority-gauges (a middlemost aggregation mechanism). All of B's strategic grades are honest, and only 32% of G's and 37% of O's are not honest.

The strategy-profiles of the proof last theorem 20.8 may be interpreted as follows. First, voters are motivated by the wish to elect a certain candidate; second, they are motivated by the wish to express themselves as honestly as

possible or to give candidates final grades as close as possible to their assessments. It will be seen in the last section of this chapter that when utilities depend not only on who is the winner but also on the winner's final grade, the winner in all strong-equilibria strategy-profiles is elected with his true majority-grade.

Some theorists do not believe in using the concept of a strong equilibrium because this implies considerable coordination among the voters. In any case, if the vote is secret and since a voter is never certain about the strategies of others, he may respond to some probabilistic belief about others' behavior. The way such probabilities are formed induces a best response correspondence. The next two sections, which explore the fixed points of such correspondences, are fairly technical. It turns out that the Condorcet-winner emerges as the unique possible equilibrium outcome; the number of possible equilibria is very small, sometimes unique; and the output of the election is determined by many honest votes. Thus a quite different set of arguments conclude with results very similar to those developed in previous sections.

20.3 Best Response Equilibria

"[The] voting behavior of at least some persons is a function of their expectations of the election outcome; published poll data are assumed to influence these expectations, hence to affect the voting behavior of these persons" (Simon 1954, 245–246).

Imagine a large electorate, approaching elections, a plethora of polls and other public information, and a method of voting where voters assign grades to candidates. An individual voter has a very small chance of changing the winner by changing the grades he assigns in any case, and he knows it. But it is reasonable for a voter to believe that if the grade he assigns to one candidate changes the outcome of the election, then all the other grades he assigns cannot change the outcome. For example, if the final grade is the median, for a voter to change the medians of two candidates is less likely by orders of magnitude than changing the median of one candidate (a rare event to begin with). In the jargon of game theory, if the probability of being pivotal for one candidate is some $\epsilon(n)$ (where n is the number of voters and $\epsilon(n) \to 0$), being twice pivotal has probability of, say, $\epsilon(n)^2$.

Assume that a voter is *sophisticated*—he reasons that the probability of being able to affect the result with more than one of the grades he assigns is so minuscule as to be negligible—or a voter behaves with *limited rationality*—the calculation to decide how to correlate the grades he assigns is too complex a task to perform, so he assigns each grade independently.

As a consequence, there are as many possible significant events as there are candidates, and to maximize his utility a voter considers the candidates one at a time and closes his eyes to secondary effects.

Best Response Strategies for Any Mechanism That Depends on the Majority-Gauge Suppose the voters only have or only use the following information. X is the likely winner with final grade α, and Y is the runner-up with final grade β (where $\alpha \succeq \beta$). Then when voters are sophisticated or behave with limited rationality,[3] they respond as follows:

- If the voter prefers a candidate Z to X, he assigns Z a grade higher than α (if α is the highest grade, he assigns it).
- If the voter prefers X to a candidate Z, he assigns Z a grade lower than α (if α is the lowest grade, he assigns it).
- If the voter prefers X to Y, he assigns to X a grade higher than β (if β is the highest grade, he assigns it).
- If the voter prefers Y to X, he assigns X a grade lower than β (if β is the lowest grade, he assigns it).

In each of the four cases the voter maximizes the chance for the grade he assigns to elect the candidate he prefers, forgetting about possible interactions among other candidates, given the information at his disposal. "Higher" means strictly higher and "lower" means strictly lower. However, among the many possible best responses it is reasonable to suppose that the voters choose to be as honest as possible and thus give grades as close as possible to the honest grades. This may be justified by ascribing to voters lexicographic utilities where the winner is what counts first and honesty or the final grade second (see section 20.6).

To clarify the idea consider again the 1970 New York Senate election example (see table 20.1), where Goodell is the Condorcet-winner. Suppose Goodell is the likely winner with majority-grade *Good*, and Ottinger is the runner-up with majority-grade *Acceptable*. Then the best response strategies are as shown in table 20.3 (assuming voters are sophisticated or behave with limited rationality, as honestly as possible). The outcome is Goodell the winner with majority-grade *Good* and Ottinger the runner-up with majority-grade *Acceptable*, so this is an equilibrium (or fixed point) with Goodell the winner and Ottinger the

3. When there are only two grades—approval judgment—these rules specialize to the poll-leader rule (see chapter 18). This strategy may be deduced with arguments similar to those given by Laslier (2009).

Table 20.3
Best Response Strategies of Sophisticated or Limited-Rationality Voters of Table 20.1

	Type			
	1a 10%	1b 29%	2a 16%	2b 8%
B	*Excellent*	*Very Good*	*Acceptable*	*Poor*
G	*Poor*	*Good*	*Very Good*	*Very Good*
O	*Very Good*	*To Reject*	*Acceptable*	*Acceptable*

Note: These voters are sophisticated, or they behave with limited rationality to *G*, the winner, with majority-grade *Good* and to *O*, the runner-up, with majority-grade *Acceptable*.

Total percentages of types 1, 2, and 3 correspond to actual 39% for Buckley, 24% for Goodell, and 37% for Ottinger.

runner-up. The reader may check that for the most part voters assign honest grades.

On the other hand, repeated best responses do not always converge from any starting point. For example, if Ottinger is the winner and Goodell the runner-up, both with majority-grade *Good*, the response is Goodell the winner with majority-grade *Very Good* and Ottinger the runner-up with majority-grade *Acceptable*. Then the next response is Ottinger winner and Goodell runner-up, both with majority-grade *Good*, so the process cycles (just as it did in the approval voting example).

A unique leading candidate X with majority-grade α and unique runner-up Y with majority-grade β (where $\alpha \succeq \beta$) is a *fixed-point equilibrium*, written $(X, \alpha; Y, \beta)$, if the voters' response to its announcement produces the same outcome. Note that given $(X, \alpha; Y, \beta)$ the best response rule determines the voters' strategies uniquely. In this section and the next, which also concerns best responses, some statements or proofs hold generically, e.g., in any face-to-face confrontation there is an absolute majority in favor of one candidate. Otherwise disagreeable details must be spelled out.

Theorem 20.9 *Suppose a best response strategy of voters is used when the voting mechanism is the majority-gauge. (1) There exists a fixed-point equilibrium $(X, \alpha; Y, \beta)$ only if X is the Condorcet-winner. (2) If C is a Condorcet-winner, then either there exists a fixed-point equilibrium or there exists a sequence of best responses in which once C is the winner, C remains the winner in every succeeding round.*

It is implicitly assumed here that there are no ties among strategic majority-gauges (or that they are resolved by some additional rule). As will be seen anon, the profiles for which a fixed-point equilibrium is not guaranteed are rare. These

Table 20.3
(cont.)

	Type		
	3a 22%	3b 15%	Majority-Gauge
B	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)
G	*Poor*	*Poor*	(24%, *Good*, 47%)
O	*Very Good*	*Excellent*	(47%, *Acceptable*, 29%)

equilibria are close to being strong equilibria: if X is the winner and a coalition of voters prefer Z to X, they cannot increase the majority-gauge of Z; and those who prefer Y to X cannot decrease the majority-gauge of X.

Proof of First Statement Suppose that $(X, \alpha; Y, \beta)$ is a fixed-point equilibrium. Note, first, that the best response strategy implies that X's majority-grade is strictly above Y's majority-grade, $\alpha \succ \beta$.

Assume that X is not a Condorcet-winner. Then there is some candidate Z (perhaps Y) who is preferred to X by a majority of voters. If α is not the highest grade, Z must have a higher majority-grade than X at the next round, contradicting the fact that $(X, \alpha; Y, \beta)$ is an equilibrium. So suppose α is the highest grade. If $Z = Y$, then Y has a higher majority-grade than X at the next round, a contradiction. So suppose $Z \neq Y$. Then Z's majority-grade at the next round is the highest grade, and Z must replace either X or Y as the leader or runner-up, again a contradiction. ■

Proof of Second Statement Let C be the Condorcet-winner. Let δ be the higher of C's true majority-grade or one above the minimum grade. In response to C the leader with majority-grade δ, the runner-up Y has a strategic majority-grade β that is at most $\delta - 1$ (meaning just below δ).

Suppose $\beta = \delta - 1$. In response to the runner-up candidate Y with strategic majority-grade $\delta - 1$, the majority of the voters who prefer C to Y assign at least δ to C (and a minority assign a grade less than $\delta - 1$), so C's strategic majority-grade must be at least δ. But the only grades that are increased over the true grades are increased to δ, so by definition of δ, C's strategic majority-grade must be exactly δ. Since the best responses yield these outcomes, $(C, \delta; Y, \delta - 1)$ is a fixed-point equilibrium.

Suppose $\beta \prec \delta - 1$, and consider the response of the runner-up Y or indeed of any candidate $Z \neq C$. Among the majority of voters who prefer C to Z, some lower the grades of Z to $\delta - 1$; among those who prefer Z to C, some increase the grades of Z; and yet Z's strategic majority-grade is at most β (since Y's the

runner-up is β). This implies that the true grades given to Z (and Y) by a majority of voters (all of whom prefer C to Z) are at most β. Therefore, the strategic majority-grades of the best responses of any Z against (C, α) where $\alpha \succeq \beta + 1$ will be identical to Z's strategic majority-grade against (C, δ); in particular, Y's strategic majority-grade will be exactly β and any other candidate's will be at most β, though the runner-up may change. However, the strategic majority-gauge of the best responses of any Z against (C, α) where $\alpha \succeq \beta + 2$ will be identical to Z's strategic majority-gauge against C with grade δ; so in particular, Y will remain the runner-up.

In response to runner-up Y with strategic majority-grade β, C's strategic majority-grade is some $\alpha_1 \succ \beta$. If $\alpha_1 \succ \beta + 1$, then $(C, \alpha_1; Y, \beta)$ is a fixed-point equilibrium. If $\alpha_1 = \beta + 1$ and it happens the runner-up is again Y, then $(C, \alpha_1; Y, \beta)$ is again a fixed-point equilibrium.

The last remaining possibility is $\alpha_1 = \beta + 1$ and the runner-up changes to some candidate Z with strategic majority-grade β. Let α_2 be C's best-response grade to the runner-up (Z, β). If $\alpha_2 = \beta + 1$ then $(C, \alpha_2; Z, \beta)$ is a fixed-point equilibrium. If not, the best responses yield the sequence

$$(C, \alpha_2; Z, \beta) \to (C, \alpha_2; Y, \beta) \to (C, \alpha_1; Y, \beta) \to (C, \alpha_1; Z, \beta) \to (C, \alpha_2; Z, \beta),$$

so C is the winner in every succeeding round. ■

Comments on Equilibria First, there are few fixed-point equilibria. If there is one where the grade of the Condorcet-winner is α, then the best responses almost surely determine a unique runner-up Y with a unique strategic majority-grade β. Since $\alpha \succ \beta$, there can be none where α is the minimum grade. Thus, there are at most as many fixed-point equilibria as the number of grades in the language less 1. In fact, the situation is even better, for the number of type $(C, \alpha; Y, \alpha - 1)$ is likely to be small: if α is low compared to the true majority-grade of C, best responses will drive up the strategic majority-grade of C; if high, they will drive down the strategic majority-grade of Y.

Finally, if $(C, \alpha; Y, \beta)$ is an equilibrium with $\alpha \succeq \beta + 2$, then there can be no other equilibrium with Y the runner-up. For suppose otherwise, namely, $(C, \alpha'; Y, \beta')$ is also an equilibrium. If $\alpha' \succeq \beta + 1$, then the best response to (C, α') gives Y the majority-grade β (by the arguments given in the preceding proof), so $\beta' = \beta$ and hence $\alpha' = \alpha$. If $\alpha' \preceq \beta$, then $\beta' \prec \beta$, and since $\alpha \succ \beta + 1$ was a best response to C against the runner-up (Y, β), it remains a best response to C against (Y, β'), so $\alpha' = \alpha$, a contradiction.

Second, for most profiles, fixed-point equilibria exist. Let C be the Condorcet-winner, and A be the true majority-gauge-winner with majority-grade δ. The following are true:

$(C, \delta + 1; Y, \delta)$ is a fixed-point equilibrium when either

- $A \neq C$ and δ is not the maximum grade, or
- $A = C$ and δ is the minimum grade;

$(C, \delta; Y, \delta - 1)$ is a fixed-point equilibrium when either

- $A \neq C$ and δ is the maximum grade, or
- $A = C$, δ is not the minimum grade and the true runner-up B has a true majority-grade of at least $\delta - 1$.

The only case when there may be none is when $A = C$ and the true runner-up B has a true majority-grade of less than $\delta - 1$. But then, as was seen, either there exists an equilibrium or there is a sequence of best responses where the Condorcet-winner always remains the winner.

Proof This proves only the first of these assertions, that $(C, \delta + 1; Y, \delta)$ is a fixed-point equilibrium when $A \neq C$ and δ is not the maximum grade. The other proofs use similar arguments. Set $\alpha = \delta + 1$ and $\beta = \delta$, and let Z be any candidate other than C. Z's true majority-grade is at most δ. A majority of voters prefer C to Z. They will give their true grade to Z if it is less than α and give $\delta = \alpha - 1$ if their true grade is at least α, which cannot change Z's majority-grade since those that decreased did not give a grade below Z's true majority-grade. Among the remaining voters (a minority) who prefer Z to C some may have increased Z's grade, so the strategic majority-grade of Z—which can be at most $\delta = \alpha - 1$, since a majority gave at most that grade—is at least Z's true majority-grade but below α.

Therefore, every candidate Z other than C has, in response to the Condorcet-winner C with grade α, a strategic majority-grade equal to or above his true majority-grade but at most δ, including, of course, A. This implies that there must be a runner-up candidate Y with strategic majority-grade $\delta = \alpha - 1$ (who is almost surely unique because it is determined by the majority-gauge, and may well be A).

The true majority-grade of C is at most δ. In response to the runner-up candidate Y with majority-grade δ, a majority of the voters will assign at least $\delta + 1$ to C and a minority will assign a grade less than δ, so C's strategic majority-grade must be at least $\delta + 1$. But the only grades that are increased over the true grades are increased to $\delta + 1$, so since the true majority-grade is at most δ, the strategic majority-grade cannot be above $\delta + 1$ and thus must be exactly $\delta + 1$. Since the best responses yield these outcomes, $(C, \delta + 1; Y, \delta)$ is a fixed-point equilibrium. ■

When the election mechanism is point-summing, the best response strategies of voters who are sophisticated or who behave with limited rationality depend on the leader and runner-up and their grades as before, except that voters only use extreme grades since every little bit helps. Thus whenever the prescription is to give a higher grade, the voter gives the highest possible grade; whenever it is to give a lower grade, the voter gives the lowest possible grade. As a consequence, rational voters are encouraged to forget their true grades and to use two grades, the highest and the lowest, so the analysis becomes exactly that developed for approval voting in chapter 18. This means, in particular, that final grades—scores when a point-summing method is used, approval or disapproval when approval voting is used—convey little or nothing concerning society's opinion of the candidates.

20.4 Best Response Dynamics

A question remains: Does there exist a best response dynamic that converges to the Condorcet-winner (when he exists), as there was for approval voting? When all voters simultaneously respond with the grades of all candidates, the best response strategies for mechanisms that depend on the majority-gauge do not necessarily converge.

Imagine that a poll announces that the leading candidate is X_0 with grade α_0 and the runner-up is Y_0 with grade β_0 (implying $\alpha_0 \succeq \beta_0$). This determines the first *outcome* $(X_0, \alpha_0; Y_0, \beta_0)$ in a sequence of polls ending in the actual election. In a real situation, the outcome of a next round does not come about because every candidate's majority-gauge is computed simultaneously but rather because some one or several of them are in the news and the voters give their best responses concerning only them. To model this idea the dynamic of successive polls studied here assumes that voters give their best response grades to some one candidate Z at each round—who may be the leader, the runner-up, or some other candidate—in view of the current outcome. This may or may not change the outcome.

Table 20.4
Best Response Strategies of Sophisticated or Limited-Rationality Voters of Table 20.1

	10%	29%	16%	8%
B	*Excellent*	*Excellent*	*Acceptable*	*Poor*
G	*Poor*	*Very Good*	*Very Good*	*Very Good*
O	*Excellent*	*To Reject*	*Good*	*Acceptable*

Note: These voters are sophisticated, or they behave with limited rationality to (*G, Very Good*; *O, Good*), a fixed-point equilibrium.

To understand the idea, consider the 1970 New York Senate example of table 20.1. Suppose it happens that the candidates' respective majority-gauges are

O : (47%, *Good*, 24%) B : (39%, *Acceptable*, 30%)
G : (24%, *Acceptable*, 10%),

so that the outcome (O,*Good*; B, *Acceptable*) is announced. The best responses to candidate B yield

	10%	29%	16%	8%	22%	15%	Majority-Gauge
B :	*Excellent*	*Very Good*	*Acceptable*	*Poor*	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)

so the outcome is the same. The best responses to candidate O now yield

	10%	29%	16%	8%	22%	15%	Majority-Gauge
O :	*Poor*	*To Reject*	*Good*	*Good*	*Very Good*	*Excellent*	(37%,*Good*, 39%)

so the outcome (O,*Good*; B, *Acceptable*) is announced. The best responses to candidate G become

	10%	29%	16%	8%	22%	15%	Majority-Gauge
G :	*Poor*	*Very Good*	*Very Good*	*Very Good*	*Acceptable*	*Acceptable*	(0%,*Very Good*, 47%)

and the two leading candidates are now (G , *Very Good*; O, *Good*). Taking the three candidates in any order elicits the best responses given in table 20.4.

When, instead, the initial announcement is (G, *Good*; O,*Good*), and the voters again begin by responding with B's grades, nothing changes. If they then give O's grades, with outcome (G, *Good*; O, *Acceptable*), followed by G's grades, the result is as shown in table 20.5. If the voters now respond by giving grades to B or O or G, the outcome remains (G,*Good*; O, *Acceptable*): there is convergence and the winner is the Condorcet-winner.

In this example there are exactly two fixed-point equilibria: (G, *Very Good*; O, *Good*) and (G, *Good*; O, *Acceptable*). There can be no equilibrium with G the winner and O the runner-up where their majority-grades differ by more than

Table 20.4
(cont.)

	22%	15%	Majority-Gauge
B	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)
G	*Acceptable*	*Acceptable*	(0%, *Very Good*, 47%)
O	*Excellent*	*Excellent*	(47%, *Good*, 37%)

Table 20.5
Best Response Strategies of Sophisticated or Limited-Rationality Voters of Table 20.1

	10%	29%	16%	8%
B	*Excellent*	*Very Good*	*Acceptable*	*Poor*
G	*Poor*	*Good*	*Very Good*	*Very Good*
O	*Very Good*	*To Reject*	*Acceptable*	*Acceptable*

Note: These voters are sophisticated, or they behave with limited rationality to (*G, Good*; *O, Acceptable*), a fixed-point equilibrium.

one level (it was noted in section 20.3 that if there exists an equilibrium where Y the runner-up's majority-grade is at least two levels below the winner's, then it is the unique equilibrium in which Y is runner-up), and it is simple to check that no others exist where G's majority-grade is one above O's. Note that in the first case G's grade is above his true majority-grade and O's is his true majority-grade, whereas in the second case G's grade is his true majority-grade and O's is one below. B's majority-grade and majority-gauge are in both cases the true ones. The strategic majority-grades are close to the true ones.

Specialized to approval voting (or approval judgment), the same dynamic uses the leader and runner-up (the grades add nothing). If, for example, the voters initially give approvals only to those candidates they evaluate to be *Excellent*, the respective approval scores of the candidates are B, 10; G, 0; and O, 15; and the initial outcome is (O, B). The successive columns of table 20.6 give the respective approval scores of each of the candidates and the associated outcomes. For example, from the scores (39,0,61) and outcome (O, B) of the third column, the best responses of the voters to candidate G are the approvals and disapprovals (0,1,1,1,0,0) for each of the six types, changing the approval score of G to 53 and the outcome to (O, G) in the fourth column. At equilibrium the approval score of the Condorcet-winner (here G, score 53) is the number of voters that prefer him to the runner-up (here O); the approval score of every other candidate X is the number of voters who prefer X to the Condorcet-winner. These scores depend only on the relative preferences. They could result from a host of different evaluations of the candidates (e.g., in the true preferences of table 20.1, types 2a, 2b, 3a, and 3b are indistinguishable to approval voting strategies).

The *best response dynamic procedure* (BRDP) is as follows.

· *Time* t. A strategic profile of majority voting grades and two leading candidates, the leader and the runner-up, are given.

· *Time* $t+1$. Some one candidate Z is chosen at random (Z may be any candidate). Voters give their best response strategy grades to Z. This defines

Table 20.5
(cont.)

	22%	15%	Majority-Gauge
B	*Poor*	*Acceptable*	(39%, *Acceptable*, 30%)
G	*Acceptable*	*Acceptable*	(24%, *Good*, 47%)
O	*Very Good*	*Excellent*	(47%, *Acceptable*, 29%)

Table 20.6
Approval Scores of Best Response Strategies of Sophisticated or Limited-Rationality Voters of Table 20.1

		B	*O*	*G*	*B*	*O*	*G*
B	10	39	39	39	39	39	39
G	0	0	0	53	53	53	53
O	15	15	61	61	61	47	47
Outcomes	*(O, B)*	*(B, O)*	*(O, B)*	*(O, G)*	*(O, G)*	*(G, O)*	*(G, O)*

Note: These voters are sophisticated, or they behave with limited rationality using approval voting. The last column is the fixed-point equilibrium (*G, O*).

the strategic profile at time $t + 1$. A new leader and runner-up are designated. If there are two or more candidates with the highest majority-grade, the leader and the runner-up are chosen at random among them; otherwise the candidate with the highest majority-grade is the leader, and the runner-up is chosen at random among candidates with the next highest grade.

The random choice of a leader is justified because candidates with highest majority-grades are necessarily important candidates, even though their majority-gauge may not be the highest, so polls and the media may well at some time take one of them to be the leader.

Theorem 20.10 *Beginning with any initial profile of grades of a finite language, the BRDP converges (almost surely) to an outcome whose winner is the Condorcet-winner (if he exists); otherwise the outcomes cycle, the leader always one of the majority top cycle of candidates.*

Proof The BRDP defines a Markov process. A state of the process is the set of candidates with the highest majority-grade.

Suppose first that there exists a Condorcet-winner C. The state in which C is the unique candidate with the highest majority-grade is clearly absorbing, i.e., once reached, never left.

There is a positive probability that from any initial state, the BRDP generates a sequence of states that leads to C the unique leader. Take any state. Either C

belongs or not. If not, the BRDP selects C with a positive probability, and C's majority-grade necessarily increases and either becomes the unique candidate with the highest majority-grade—in which case the absorbing state has been reached—or is among those with the highest majority-grade, which must be the highest possible grade α_{max}. There is a positive probability that C is selected as leader, and voters are asked to respond to another candidate Z having the maximum majority-grade. In that case Z is eliminated from the set of candidates having the maximum grade. Continuing, candidates are ejected (with positive probability) one by one from the set of those with the majority-grade α_{max}.[4]

Suppose next that there is no Condorcet-winner. Then there is a majority top cycle of candidates $C_1 \succ_{maj} C_2 \succ_{maj} \cdots \succ_{maj} C_k \succ_{maj} C_1$: a majority prefers C_1 to C_2, C_2 to C_3, ... , and each of these candidates is preferred by a majority to any candidate who is not in the top cycle. If a state is reached that includes only candidates of the top cycle, then succeeding states will only include candidates of the top cycle (because only candidates of the top cycle can defeat a candidate of the top cycle).

There is a positive probability that from any initial state, the BRDP generates a sequence of states that leads to a state that contains only top cycle candidates. Take any state. If it includes no candidate of the top cycle, then with positive probability the new state will include such a candidate. If it does include a candidate of the top cycle, then candidates not of the top cycle are eliminated (with positive probability) one by one.[5] ■

The preceding theorems described situations where a Condorcet-winner exists. The proof of the last theorem shows that when there is no Condorcet-winner, the winner will necessarily belong to the majority top cycle of candidates.

20.5 Strategic Majority Judgment Winner

What can be said about the equilibria of games of voting when animal spirits are part of the analysis? This may, of course, be modeled with utility functions that are not restricted to the clearly unrealistic assumption that a voter cares *only* about who is elected. Hypothesize instead that in the new model there are three types of voters with different utilities:

· *Type I voters i* have utilities u_i that depend only on the winner of the election (as in the usual analyses). In the new model these u_i are assumed to be

4. The procedure continues to evolve to an equilibrium.

5. In fact, the proof immediately extends to an altered and more realistic BRDP where only one voter or a subset of voters responds at each time period t to one candidate.

compatible with the true grades of the voters. When the language of grades is sufficiently rich, $u_i(X) > u_i(Y)$ if and only if i's true grade for X is above i's true grade for Y; otherwise $u_i(X) > u_i(Y)$ implies i's true grade for X is at least as high as i's true grade for Y. Type I voters are winner optimizers.

· *Type II voters* i have utilities u_i that are single-peaked-in-grading and thus depend only on the final grades of candidates. The closer the final grade of a candidate X to i's true grade for X, the greater is i's utility. Type II voters are final-grade optimizers.

· *Type III voters* i have utilities that depend only on honesty: the further the deviation of the grades such voters i assign to a candidate X from their true grade for X, the lower are their utilities. These voters will simply always assign their true grades. Type III voters are honesty optimizers.

Type I voters constitute the class of voters that has traditionally been studied. In fact, it seems that only a relatively small percentage of voters in national elections are of this type. (Even after the debacle of the 2002 French presidential election, at most 30% of French electors were of this type in the next presidential election, 2007.) Type II voters were analyzed previously, and it was shown that when the mechanism is an order function, and in particular, the majority judgment, their optimal strategy is to assign to candidates the grades they honestly believe are warranted. Type II and III voters are those who really wish to send messages declaring to the public at large, as well as to the candidates themselves, the esteem with which they regard politicians. They are, of course, frustrated by the traditional model and the methods actually used to conduct elections.

To begin, assume there are only voters of types I and III, and consider the game of voting with an arbitrary mechanism F. A candidate X is the *strategic winner of F against Y* if X wins when all voters of type I who prefer X to Y give to X the highest possible grade and to Y the lowest possible grade, those who prefer Y to X do the opposite, and all other voters assign both X and Y the grades they honestly believe are merited. A candidate X is the *strategic winner of F* if no other candidate is the strategic winner against him.

Theorem 20.11 *Suppose all voters are of types I or III. C is a strong-equilibrium winner of a mechanism F if and only if C is a strategic winner of F.*

Proof Suppose C is a strong-equilibrium winner of F, and let B be any other candidate. If B is the strategic winner of F against C, then since type III voters give honest grades, a coalition of type I voters that strictly prefer B to C can manipulate by giving C the lowest possible grade and B the highest possible grade and thus elect B, a contradiction.

Suppose, then, that C is a strategic winner of F. Consider the strategy-profile where every voter of type I gives to C the highest grade and to every other candidate the lowest grade, and all other voters give honest grades. Any coalition that prefers another candidate B to C is unable to elect B by changing its strategy because C is a strategic winner, so C is a strong-equilibrium winner. ■

A candidate X is the *strategic majority winner against* Y if X is the winner against Y with the majority judgment when all voters of type I who prefer X to Y give to X the highest possible grade and to Y the lowest possible grade, those who prefer Y to X do the opposite, and all other voters (of types II and III) assign both X and Y the grades they honestly believe are merited. A candidate X is the *strategic majority winner* if X is the strategic majority winner against every other candidate. A candidate X is a *coalitional-equilibrium winner* if no coalition of voters of the same type can change their strategies and elect a candidate Y preferred by all of them (Laraki 2009).

Theorem 20.12 *A strategic majority winner C is a coalitional-equilibrium winner.*

Proof All type I voters give to C the highest possible grade and to any other candidate the lowest possible grade. All type II and III voters assign grades honestly. By definition, no set of type I voters who prefer a candidate B to C can manipulate. Moreover, no set of type II voters who would like to increase or decrease the majority-gauge of any candidate can do so because the majority-gauge is strategy-proof-in-grading. ■

The theorem cannot be strengthened to deviating coalitions formed by voters of types I and II. However, this would require that the participating type II voters assign grades that are not their unique undominated strategies and would make such coalitions both rare and unstable.

What does this theorem say? Under the generally accepted assumption that a rational player opts for her undominated strategy when it exists, the theorem says that the strategic majority winner is a strong-equilibrium winner. When type I voters are many in comparison with voters of types II and III, the strategic majority winner will be a Condorcet-winner; when type I voters are few in comparison with voters of types II and III, the strategic majority winner will be the majority judgment winner (see chapter 19).

On the other hand, if the mechanism assigns to each candidate her average grade, then all voters of type II who assigned a grade above the winner's average have an interest in increasing it to the highest possible grade, and those who

assigned a grade below the winner's average have an interest in decreasing it to the lowest possible grade.

20.6 Condorcet-Judgment-Winner

More realistically, assume now that the preferences of voters are over the pairs (A, α), where A is the winner and α is the winner's final grade. A voter's preferences $\succeq$ over these pairs are assumed to be complete, transitive, and when the winner A is the same, the preferences over the pairs (A, α) is single-peaked in the majority-grades α, that is, a voter prefers the winner's final grade to be as close as possible to his ideal grade. The ideal grade given a winning candidate maximizes the voter's utility. In the *honest profile* of the electorate a voter assigns his ideal grade to each candidate.

A is called a *Condorcet-judgment-winner* if there is no candidate B such that a majority of voters strictly prefer (B, β) to (A, α), where β is any grade and α is the honest majority-grade of A.

A voter's preference over pairs is *lexicographic* if $(A, \alpha) \succ (B, \beta)$, then $(A, \alpha') \succ (B, \beta')$ for all α', β'. Consequently, when the preference is lexicographic, a Condorcet-winner coincides with a Condorcet-judgment-winner. In general the concepts may differ.

Theorem 20.13 *With the majority-gauge, (A, α) is a strong-equilibrium outcome if and only if A is a Condorcet-judgment-winner and α is his honest majority-grade.*

Proof Assume there are $2n$ or $2n + 1$ voters and that (A, α) is a strong-equilibrium outcome. If α is not the true majority-grade of A, then a majority of voters will prefer A to win with his honest majority-grade (say, β) rather than for him to win with the majority-grade α.

If β is not the minimal grade, a majority can assign the grade β to A and the minimal grade to all other candidates, implying that the new output is (A, β). If, on the other hand, β is the minimal grade, n voters of the majority can assign the highest possible grade to A and the others the minimal grade to A; and assign to all other candidates the minimal grade. In that case, the highest majority-gauge is for A, who is elected with his honest majority-grade. Consequently, if (A, α) is a strong-equilibrium outcome, α is his honest majority-grade.

Assume now that some pair (B, β) is preferred to (A, α) by a majority of voters. This majority can elect (B, β) as described. Thus A is a Condorcet-judgment-winner.

Conversely, if A is a Condorcet-judgment-winner with the honest majority-grade α, the described strategy elects A with his honest majority-grade. ■

Voters may have preferences over pairs consisting of a ranking of the candidates and their associated majority-grades. If a pair of a ranking and its majority-grades is the output of a strong equilibrium, then no other pair is preferred to it by a majority of voters (otherwise the majority could impose its will).

20.7 Conclusion

The utilities that judges and voters maximize are unknown. They are undoubtedly more complex than over pairs of rank-orders and associated majority-grades. How, then, *should* utilities be modeled or approximated? We believe they should be made to depend as much as possible on the results or the outputs of the system that is studied. When first-past-the-post is analyzed, the output is the number of votes received by each candidate, together with the winner and the rank-order that is induced. When Borda's method is used, the output is the Borda-score of each candidate, together with the winner and the rank-order that is induced. When majority judgment is used, the output is each candidate's distribution of grades, together with each candidate's resulting majority-grade and majority-gauge, the winner and the majority-ranking. The difficulty is that realistic utilities resist qualitative analyses. This is why utilities have been given relatively simple formulations such as depending only on the identity of the winner or on the pair of a winner and a final grade.

One thing is clear. Extending the concept of utilities to all the outputs of an election—those that are announced publicly—helps to explain many obscure phenomena that have received considerable attention. Examples abound. Take first-past-the-post. Abstention: perhaps a voter does not wish a winner to win by much. Participation: perhaps the contrary (e.g., Chirac versus Le Pen, which brought out a record number of voters). Votes given small party candidates: voters know they cannot win and may wish they do not win, and yet they vote for them.

It would be more realistic with the majority judgment to assume that voters have utilities that depend on the entire distribution of the grades assigned by the electorate and the attendant majority-gauges and majority-ranking, but this could lead to much more complex mathematics.

The simple analysis given here suffices to give additional insight as to why the majority judgment is a more honesty-inducing mechanism than others in the context of the game of voting.

21 Multicriteria Ranking

Each chamber, transept, coins some squint,
Remorseless line, minting their separate wills—
—Hart Crane

Each voter distinguishes one candidate for political office from another by some ill-defined mix of criteria that may touch upon party affiliation and party platform; honesty and moral outlook; voice, appearance, and charisma; foreign, economic, and social policies; and a host of other considerations. But there is no agreement among voters on which of these aspects are more or less important: each voter is left to integrate all the criteria he believes of importance to reach a final judgment on the merit of each candidate.

Skaters, gymnasts, countries, pianists, wines, . . . , however, are routinely evaluated on the basis of separate criteria, attributes, or characteristics, and the evaluations of the parts are aggregated into an evaluation of the whole. For skaters and gymnasts, the merit of separate parts of performances are measured, then aggregated into a measure of the whole. For countries, the quality-of-life index is a weighted sum of indicators that concern health, political stability, security, community life, political freedom, and so on. For pianists, often a sequence of increasingly demanding performances weeds out competitors to end up with a handful who are ranked on the basis of their final recitals. For wines, each of a set of well-defined criteria depending on odor, taste, aspect, and "total impact" is evaluated in terms of a common language; the evaluations are transformed into numbers; and the numbers are added to determine final number-grades. The goal of this chapter is to extend the majority judgment to multicriteria problems.

21.1 Aggregating Criteria

Proverbs that warn against combining incomparables abound in all cultures: "comparing apples and oranges," "comparer des pommes et des poires," "sumar peras con manzanas," "comparing grandmothers and toads" (Serbian), and the somewhat less elegant "confundir la mierda con la pomada" (Colombian). What can be said mathematically about aggregating different characteristics, attributes, or criteria?

Indeed, that is a question that may be posed for only *one* judge. So, to begin, suppose there are a set of l criteria $K = \{1, \ldots, k, \ldots, l\}$ but only one judge. Let the common language of the jth criterion be Λ_j. How is the one judge to aggregate the grades given to the different criteria so as to determine an order of finish?

The judge gives a vector of grades $\alpha = (\alpha_1, \ldots, \alpha_l)$ to each competitor A, where $\alpha_j \in \Lambda_j$. The judge's profile of grades is

$$\Phi = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_{l-1} & \alpha_l \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \beta_1 & \beta_l & \cdots & \beta_{l-1} & \beta_l \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{pmatrix},$$

each row corresponding to some competitor.

A *ranking rule* $\succeq_R$ is needed to determine the order of finish among any number of competitors. Take $A \approx_R B$ to mean A and B are tied, and $A \succ_R B$ to mean $A \succeq_R B$ and not $A \approx_R B$. The ranking rule should satisfy the following:

· *Neutrality* $A \succeq_R B$ for the profile Φ implies $A \succeq_R B$ for the profile $\sigma\Phi$, for σ any permutation of the competitors (or rows).

· *Transitivity* $A \succeq_R B$ and $B \succeq_R C$ implies $A \succeq_R C$.

· *Monotonicity* Suppose A's grades are $\alpha = (\alpha_1, \ldots, \alpha_l)$ and B's grades are $\beta = (\beta_1, \ldots, \beta_l)$. If $\alpha_k \succ \beta_k$ for all $k \in K$, then $A \succ_R B$; and if $\alpha_k \succeq \beta_k$ for all $k \in K$, then $A \succeq_R B$.

· *Independence of irrelevant alternatives in ranking* (IIAR) $A \succeq_R B$ for some profile Φ implies $A \succeq_R B$ for any profile Φ' obtained from Φ by adjoining or eliminating some other competitor (or row).

Moreover, if no pair of languages Λ_j are common, then the ranking rule should also be

· *Preference-consistent* $A \succeq_R B$ for some profile Φ implies $A \succeq_R B$ for any profile Φ' obtained from Φ by a monotonic transformation ϕ_j of the language of grades Λ_j of each criterion.

The desirable properties are strangely reminiscent: they ask what Arrow's theorem asked when the role of voters or judges is played by the criteria (see chapter 11). As a consequence, the only rule $\succeq_R$ that satisfies the properties is the dictatorial rule: to decide on the relative standing of two competitors, there is a well-defined sequence of criteria. The first criterion decides; if it declares the competitors tied, the second criterion decides; if the second criterion declares a tie, the third criterion decides; . . . ; if all the criteria of the sequence declare a tie, the result is a tie. A very similar result is given by Plott, Little, and Parks (1975).

More generally, if some of the languages Λ_j are common, a set of criteria with a common language decides; if there are ties among competitors, a second set of criteria with a common language decides; and so on.

What is to be done? Practice, once again, gives ideas. Several methods have been used, some of which have already been described. Two are recalled here.

Lexicographic Multicriteria One criterion or a set of criteria with a common language decides; if there are ties among competitors, a second criterion or set of criteria decides; if ties remain, a third criterion or set of criteria decides; and so on. This may be said to be (in part) the procedure used in the Chopin International Piano Competition, where successive stages (or criteria) eliminate competitors, with a final stage (or criterion) to determine the order of finish among six competitors.

Multicriteria Weighted Point-Summing The language of each criterion is translated into points. The different criteria are assigned weights that correspond to their relative importance; or the points used in each criterion already reflect their relative importance (as they do in wine competitions; see chapter 7 and section 21.2). This is the procedure used in figure skating, gymnastics, and wine competitions, among others. But in such instances, the points (adjusted, when appropriate, by the weights) are routinely added.

Take A to be any competitor. There are two ways to find A's total score when addition is the underlying approach:

· Find each judge's aggregate total score for competitor A, then calculate A's sum (or trimmed sum), or average (or trimmed average) over all judges. This is the usual way.

· Find the sum (or trimmed sum)—or average (or trimmed average) of A's kth criterion points over all judges, then aggregate them.

When the procedure takes sums or averages, the two calculations give the same results. When the procedure takes trimmed sums or averages (most often meaning one or two of the highest and lowest grades are eliminated), they may not give the same results.

These traditional weighted point-summing procedures have their analogues in the majority judgment. Better, when possible, theory and practice suggest that a common language should be used by all criteria, which leads to what seems to be the most reasonable application of the majority judgment to multicriteria evaluation. Before describing these methods, the results of a wine competition are examined and analyzed to develop intuition and insight.

21.2 Common Language: Wine Competitions

Les Citadelles du Vin is an annual wine competition held in the Bordeaux area in June, organized by Jacques Blouin, a well-known French œnologist.[1] Since 2006 it has used two methods for classifying wines: the official method, a traditional weighted point-summing method, approved by the International Organization of Vine and Wine (OIV); and on experimental method, the majority judgment with a single global criterion. The analysis given here is based on the results of the 2006 competition.

In 2006, 1,247 wines were classified by some sixty judges organized into twelve juries of five judges (occasionally, because of temporary absences, fewer judges). The evaluation forms that judges were asked to fill out for each wine included the ten criteria for the weighted point-summing method (table 21.1a) and the five grades for the majority judgment (table 21.1b) as well as a detailed set of descriptors concerning characteristics of the wine that are used for other purposes. In table 21.1a the sum of ten *Excellent*s is 100, of ten *Very Good*s 86, of ten *Good*s 72, of ten *Fair*s 56 and of ten *Insufficient*s 40, giving a rough sense of how numbers correspond to word evaluations.

In 2006 the majority judgment experiment used five grades (see table 21.1b). This was not the optimal choice; too often, judges waffled by giving the middle grade, *Good*. In subsequent years, six grades were used—*Excellent, Very Good, Good, Average, Passable, Mediocre*—which gave more satisfactory results.

Table 21.2 gives the five-person jury no. 2's grades for each of three white wines A, B, and C. The usual weighted point-summing method approved by

1. We are indebted to J. Blouin for giving us the official results of the 2006 competition.

Table 21.1a
Form to Record the Inputs of One Judge for One Wine, Weighted Point-Summing Method, Still Wines, *Les Citadelles du Vin,* 2006

		Excellent	*Very Good*	*Good*	*Fair*	*Insufficient*
View						
Limpidity	[VL]	5	4	3	2	1
Aspect	[VA]	10	8	6	4	2
Nose						
Genuineness	[NG]	8	7	6	4	2
Intensity	[NI]	6	5	4	3	2
Quality	[NQ]	16	14	12	10	8
Taste						
Genuineness	[TG]	8	7	6	4	2
Intensity	[TI]	6	5	4	3	2
Persistence	[TP]	8	7	6	5	4
Quality	[TQ]	22	19	16	13	10
Harmony						
Overall judgment	[HO]	11	10	9	8	7

Note: On the actual forms, there were boxes to check instead of points. The number points that translate the word-grades in each criterion were absent though known to the judges.

The abbreviations for the criteria were added by the authors.

Table 21.1b
Form to Record the Inputs of One Judge for One Wine, Majority Judgment Experiment, *Les Citadelles du Vin,* 2006

FOR YOU, THIS WINE IS:				
Excellent ☐	*Very Good* ☐	*Good* ☐	*Average* ☐	*Mediocre* ☐

the OIV determines the final grade of a wine as the average of the point totals given it by the judges. This, in turn, determines the medal a wine is awarded (or not). In this example wine A's final grade is 87.2, B's 82, and C's 83.6, so that the order of finish is $A \succ_S C \succ_S B$. A's majority-gauge is (1, *Very Good*, 2), B's is (2, *Good*, 1), and C's is (2, *Good*, –), so that the order of finish is $A \succ_S C \succ_S B$. The orders happen to agree.

The grades given these three wines (among well over a thousand) are roughly representative of the competition: they are actual inputs that by and large resemble those of other wines and the other eleven juries. There are some disagreements between the inputs to the two methods, for instance, judge 3's evaluation of wine B is above that of judge 1's according to the weighted point-summing method, but below that of judge 1's according to the majority-grade. These disagreements are rather rare and appear to have no important impact on

Table 21.2
A jury's Grades for Three White Wines, *Les Citadelles du Vin,* 2006

Judge	Weighted Point-Summing											Majority-Grade
	VL	VA	NG	NI	NQ	TG	TI	TP	TQ	HO	Sum	
White Wine *A*												
1	5	8	7	5	14	7	5	7	19	10	87	*Very Good*
2	5	10	6	5	12	6	5	7	16	9	81	*Good*
3	5	10	8	5	14	7	5	8	16	10	88	*Very Good*
4	5	10	8	6	16	8	6	8	22	11	100	*Excellent*
5	5	8	7	5	14	6	4	6	16	9	80	*Average*
White Wine *B*												
1	5	8	7	5	14	7	5	7	16	9	83	*Very Good*
2	5	10	6	5	12	6	5	7	16	9	81	*Good*
3	5	10	7	5	14	7	4	7	16	9	84	*Good*
4	5	10	7	5	12	7	5	6	19	10	86	*Very Good*
5	5	10	7	5	12	6	4	6	13	8	76	*Average*
White Wine *C*												
1	5	10	7	5	14	7	4	7	16	10	85	*Very Good*
2	5	8	7	5	14	7	5	6	16	9	82	*Good*
3	5	10	7	4	12	7	5	7	16	10	83	*Good*
4	5	10	7	4	12	7	4	6	19	10	84	*Very Good*
5	5	8	7	5	14	7	5	7	16	10	84	*Good*

the medals that are awarded. In analyzing the results of the 2006 competition (and later competitions, when the majority judgment used six grades), Blouin (2008) remarked that the two methods gave homogeneous results, and that "the differences were essentially due to the presence of one extreme grade, usually a very low one, which could be considered abnormal."

Individual judges' grades will naturally differ, for some judges may be relatively generous, others relatively harsh. This is why, of course, using the majority-grade is important: it effectively avoids the impact of grades that are too high or too low. The less expert the judges, the more it is judicious to obtain grades for each characteristic. Top-notch experts may render better judgments by integrating for themselves the various characteristics.

Blouin has argued that wines are not for experts but for consumers who drink them, and for them a single global criterion makes the most sense, for it corresponds to how they appreciate wines. In that case, many consumers, not just five, should evaluate wines, and it is reasonable to expect that, as with many voters, many consumers would use the language of grades in much the

same manner, making single global evaluations via the majority judgment a good method to use. For inevitably several extreme grades in a small jury such as five will have greater impact than in a larger jury even with the majority judgment. One of the reasons that practical people have developed systems to evaluate competitors by asking for grades (usually numbers) on many different attributes and characteristics of the competitors, or on many separate parts of performances, must surely be to assure that there are many grades, thereby dampening the impact of extreme or cranky grades. This is particularly important when points are summed. Two cranky grades in a jury of five—both high or both low—will have an enormous impact when points are summed, but they cannot impose their will on the jury with the majority judgment.

Did the judges of the 2007 *Citadelles du Vin* competition use the grades—the grades for each of the ten characteristics in the weighted point-summing method and the grades of the majority judgment—in the same manner, that is, did all judges use each of the various grades with about the same frequencies? Individual judges vary in the pattern or distribution of the grades they use, and they have different tastes and different biases, just as voters of the left, the center, and the right have different tastes for candidates who span a wide political spectrum from left to right. Thus comparing two judges or two juries of five judges may be misleading, for they will naturally differ from judge to judge and jury to jury. More fundamentally, the question is to show that statistically there is a distribution of the grades used by all the judges—a common language of grades—that is a reasonable representation of those used by each judge in the following sense: when sets of several judges' grades are aggregated over several differing wines, to eliminate the individuals' biases in one direction or another, the distributions obtained are close to one another.

The elaborate and detailed analysis of the grades used by voters could be applied to the grades used by the judges in this competition, but only the raw data similar to that in table 15.14 are given here. It is, we believe, convincing in showing that, yes, the judges used the grades in the same way, for the distributions obtained are very close to one another.

Table 21.3a gives the frequencies with which each of the grades were used by the first six and last six juries, who evaluated, respectively, 351 and 295 wines. Six juries means thirty judges. The usage rates are very similar, although no wine tasted by one set of juries was tasted by the other set of juries. The average grades, based on the numbers corresponding to the grades given in table 21.1a, are almost identical. Table 21.3b gives the frequencies with which each of the five majority judgment grades were used by the two sets of juries. The usage rates are again very similar, and the average grades (attributing 5 to *Excellent*, 4 to *Very Good*, . . . , and 1 to *Mediocre*) are also almost identical.

Table 21.3a
Frequencies of Grades Given to Characteristics of Wines, *Les Citadelles du Vin,* 2006

Grade	Juries	VL	VA	NG	NI	NQ
Excellent	1–6	52%	40%	3%	3%	2%
	7–12	53%	46%	4%	2%	3%
Very Good	1–6	39%	42%	51%	44%	35%
	7–12	32%	35%	38%	35%	30%
Good	1–6	8%	16%	33%	44%	48%
	7–12	13%	15%	41%	47%	47%
Fair	1–6	1%	2%	9%	9%	11%
	7–12	2%	4%	13%	14%	16%
Insufficient	1–6	0%	0%	3%	1%	4%
	7–12	0%	1%	4%	2%	5%
Average grade	1–6	4.4	8.4	6.3	4.4	12.4
	7–12	4.4	8.4	6.0	4.2	12.2

Note: Juries 1–6 tasted 351 wines, juries 7–12 tasted 295 wines.
"Average grade" means the average of corresponding numbers (see Table 21.1a).

If—as has been the practice in evaluating wines, ranking figure skaters and gymnasts, or evaluating the quality of life in nations—several criteria must be accounted for, what method should be used?

21.3 Multicriteria Majority Judgment

Both of the traditional procedures for multicriteria inputs have analogues in the spirit of the majority judgment when the aggregation of grades across criteria is a well-defined function (such as a weighted sum or average, or a majority-grade). Which procedure to use will depend upon the particular application.

· *Judge-based majority judgment* Find each judge's aggregate grade across the criteria for competitor A, then calculate A's majority-value (or majority-gauge) over all judges to determine the majority-ranking.

· *Criterion-based majority judgment* Find the majority-grade of A's kth criterion over all judges, then aggregate the majority-grades. If ties need to be resolved to determine the majority-ranking, then the second majority-grade of A's kth criterion over all judges is found and then aggregated; if ties persist, then the third-majority-grades are calculated; and so on.

Using the same function to aggregate grades across criteria in both procedures may well lead to different results. Thus, to each such function there correspond two procedures.

Table 21.3a
(cont.)

Grade	TG	TI	TP	TQ	HO
Excellent	3%	3%	3%	3%	2%
	2%	3%	2%	2%	1%
Very Good	49%	38%	33%	27%	28%
	38%	32%	26%	21%	23%
Good	37%	46%	48%	53%	54%
	43%	46%	51%	48%	52%
Fair	9%	11%	14%	14%	13%
	13%	17%	18%	23%	18%
Insufficient	2%	1%	2%	3%	3%
	3%	3%	3%	6%	5%
Average grade	6.3	4.3	6.2	16.3	9.1
	6.0	4.1	6.0	15.7	9.0

Table 21.3b
Frequencies of Grades Given to Characteristics of Wines, Majority Judgment Method, *Les Citadelles du Vin,* 2006

Juries	*Excellent*	*Very Good*	*Good*	*Average*	*Mediocre*	Average Grade
1–6	1%	29%	49%	21%	0%	2.9
7–12	2%	23%	50%	25%	0%	2.8

Note: Juries 1–6 tasted 351 wines, and juries 7–12 tasted 295 wines.

To see how each of these methods work, they are applied to finding the three white wines' majority-grades and rankings for the judge-based procedure and the sequence of majority-grades for the criterion-based procedure, assuming that the function that aggregates the grades is the sum of their corresponding numbers.

The judge-based majority judgment calculates the majority-grades of the sum of the points: A's majority-grade is 87, B's is 83, and C's is 84, which is sufficient to determine their order, $A \succ_S C \succ_S B$. The procedure is simple and direct, but it reveals nothing about the jury's verdict on each of the characteristics.

The criterion-based majority judgment first calculates the first majority-grade of each criterion's grades (and the second, third, . . . if necessary). Then it aggregates them; in this case it adds them. The results are shown in table 21.4. A, with first and second majority-grades (86,82), is first; C with (86,79) is second; and B with 83 is last, $A \succ_S C \succ_S B$. The two procedures happen to give the same result. The procedure is somewhat more involved, but it renders

Table 21.4
Criterion-Based Majority Judgment Calculation for Inputs of Table 21.2, *Les Citadelles du Vin,* 2006

	Criterion										
	VL	VA	NG	NI	NQ	TG	TI	TP	TQ	HO	Sum
	White Wine *A*										
1st majority-grade	5	10	7	5	14	7	5	7	16	10	86
2d majority-grade	5	8	7	5	14	6	5	7	16	9	82
	White Wine *B*										
1st majority-grade:	5	10	7	5	12	7	5	7	16	9	83
	White Wine *C*										
1st majority-grade	5	10	7	5	14	7	5	7	16	10	86
2d majority-grade	5	8	7	4	12	7	4	6	16	10	79

grades to each characteristic, which is useful information. In practice, wines are not usually rank-ordered, so it would not be necessary to compute the second majority-grades.

In many if not most practical instances where several criteria are used, each criterion is evaluated in the same language of grades: for instance, in the *Citadelles du Vin* competition, the words are *Excellent*, *Very Good*, *Good*, *Fair*, and *Insufficient*. They are translated into numbers, which are then added.

The cleanest, most straightforward generalization of the majority judgment to such problems is multicriteria majority judgment. Weights w_i are attached to each criterion i in accord with its importance. For example, given the number scores for judging wines in table 21.1a, one way to approximate the weights is to simply add up the numbers used for each criterion, so that $w_{VL} = 15$, $w_{VA} = 20, \ldots, w_{HO} = 45$.

· *Multicriteria majority judgment* Obtain the grades for each criterion as usual. Replicate the grades of criterion i according to its weight—w_i times the number of each grade given—and then use the majority-gauge (or majority-value) to assign grades and determine the majority-ranking.

The idea is simple. The justification is straightforward. Each criterion implies looking at a wine from a different perspective. So assigning grades for the different aspects of a wine means measuring it from independent points of view. It is as if ten different judges were evaluating the wine (though in reality one judge gives grades to ten characteristics), each with a weight equal to its importance. In elections the voters must themselves determine the relative

Table 21.5
Multicriteria Majority Judgment, *Les Citadelles du Vin*, 2006

	Weights									
	15 VL	30 VA	27 NG	20 NI	60 NQ	27 TG	20 TI	30 TP	80 TQ	45 HO
	White Wine *A*									
Judge 1	*Exc*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*
Judge 2	*Exc*	*Exc*	*G*	*VG*	*G*	*G*	*VG*	*VG*	*G*	*G*
Judge 3	*Exc*	*Exc*	*Exc*	*VG*	*VG*	*VG*	*VG*	*Exc*	*G*	*VG*
Judge 4	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*	*Exc*
Judge 5	*Exc*	*VG*	*VG*	*VG*	*VG*	*G*	*G*	*G*	*G*	*G*
Majority-grade	*Exc*	*Exc*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*	*G*	*VG*
	White Wine *B*									
Majority-grade	*Exc*	*Exc*	*VG*	*VG*	*G*	*VG*	*VG*	*VG*	*G*	*G*
	White Wine *C*									
Majority-grade	*Exc*	*Exc*	*VG*	*VG*	*VG*	*VG*	*VG*	*VG*	*G*	*VG*

importance of the different characteristics of a candidate, whereas in judging wines, figure skaters, and other competitions, there is often a general consensus about the relative importance of the different characteristics.

To show how this method works, it is applied in table 21.5 to find the three white wines' majority-gauges, assuming the weights are as suggested. The details are given for wine *A* only. The original grades have been inserted instead of their corresponding numbers for wine *A*. Every grade given to *A* is *Excellent, Very Good* or *Good*. Letting (x_{Exc}, x_{VG}, x_G) be the count of *Excellent*s, *Very Good*s and *Good*s attached to each criterion, the assignment of weighted grades is

VL (75,0,0) NQ (60,180,60) TP (60,60,30)
VA (90,60,0) TG (27,54,54) TQ (80,80,240)
NG (54,54,27) TI (20,60,20) HO: (45,90,90)
NI (20,80,0)

This allows the majority-grades of each criterion to be calculated, which is useful information.[2]

2. Nicolas, the chain of wine sellers in France, used to give grades going from *1* to *10* to each of ten descriptive characteristics of the wines they sold. They were color, bouquet, balance, frankness, finesse, body, personality, acidity, harshness, maturity.

Note that the weighted majority-grade of a criterion is equal to its majority-grade because grades assigned to a criterion are only replicated. The total count of grades for A is (531,778,521), the sum of the separate counts attached to the criteria. So A's multicriteria majority-gauge is (531, *Very Good*+, 521). Wine B is also assigned a grade of *Fair*, so its count of grades contains four numbers, (195,714,736,125) (the 125 is the number of *Fairs* it received), giving a multicriteria majority-gauge of (195,*Very Good*−, 861); the result for wine C is a total count of (165,980,625), so a multicriteria majority-gauge of (165,*Very Good*−, 625). Thus the majority-ranking is $A \succ_S C \succ_S B$.

Given weights that determine the relative importance of the criteria, judge-based and criterion-based majority judgment procedures are immediately defined.

· *Judge-based procedure* Find each judge's weighted majority-grade for competitor A, then calculate A's majority-value over all judges (and in case of ties, find each judge's weighted second majority-grade for competitor A, then calculate A's second majority-value over all judges, ...).

· *Criterion-based procedure* Find the majority-grade of A's kth criterion over all judges, then aggregate the weighted majority-grades (and in case of ties, go to the second majority-grades).

Both of these procedures are different from multicriteria majority judgment.

Which procedure to use must be dictated, however, by the practical application. Arriving at a global evaluation as a function of a series of local evaluations is not necessarily valid. Peynaud and Blouin (2006) asked, "Is the value of a wine the simple sum of individual values, or are some more important, even decisive?" They answered, "It is obvious that a sum of grades is not a good system for evaluating quality. One weakness may be damning, and sometimes one exceptional quality makes a great wine."

22 A Summing Up

It is hard necessity and not speculation or a desire for novelty which forces us to change the old classical view . . . Changes of view are continually forced upon us by our attempts to understand reality. But it always remains for the future to decide whether or not a better solution of our difficulties could have been found.
—Albert Einstein and Leopold Infeld

Paradoxes and impossibility theorems have dominated the *theory* and *analysis* of social choice and voting from Condorcet's to Arrow's, and on to all the many others that continue to be found down to the present day. Paradoxes and anomalies—most notably and most importantly, Condorcet's and Arrow's—have plagued the *reality* and *practice* of voting and judging across the years. Today the world shows signs of a growing awareness that perhaps the mechanisms used to elect and to rank—pure inventions of the human mind—are not electing the candidates the voters want nor designating the order of finish the juries want (e.g., Poundstone 2008).

To model problems of the real world in the social sciences is no different than in any other science. Einstein and Infeld's description of the endeavor is wonderful and cannot be bettered. It is at once deep in its perspective, light in its spirit, and full of delightfully telling analogies, so we hazard to use their words to bring this book to an end.

They liken the role of the research scientist to that of the detective "who, after gathering the scientific facts, finds the right solution by pure thinking." They then retract in one particular: "The detective must look for letters, fingerprints, bullets, guns, but at least he knows that a murder has been committed . . . For the detective the crime is given, the problem formulated: who killed Cock Robin? The scientist must, at least in part, commit his own crime, as well as carry out the investigation. Moreover, his task is not to explain just one case, but all phenomena which have happened or may still happen" (1938, 78).

Their retraction may be (or may have been) correct for physics, but in the theory of voting or of social choice, not only has the murder been committed,

it was committed centuries ago. Yet, despite more and more clues, its solution has remained an enigma. For, if Condorcet's and Arrow's paradoxes are to be avoided, then which voter or judge gave what grade must be forgotten; only the grades assigned to the competitors are relevant; and the scale of grades must be absolute for each individual judge or voter. And if the results are to be meaningful, then the scale of grades must be common to all judges or voters. These statements are proven (theorem 9.2 and Arrow's theorems 11.6a and 11.6b) in the context of the new model, but the new model encompasses the old, when a voter's or judge's input messages to the traditional model—rank-orders—are determined by their input grades, namely, a higher grade means ranks higher, equal grades means ranks equally. In fact, all methods based on the traditional model are meaningless. Hence, the only possible meaningful methods of election must be based on a new model. Who committed the murder is thus perfectly clear: the traditional model's basic paradigm that judges or voters have in mind and give as inputs rank-orders of the competitors. And yet, the detectives in the matter have steadfastly accepted that paradigm.

That concerns, however, only half of the problem.

What is, after all, the job of the theory and practice of social choice, of aggregating the opinions of voters in elections and aggregating the evaluations of judges in competitions? It is to gather, as precisely as possible, the true opinions and evaluations of individuals, and to determine, as precisely as possible, the true aggregate wills of electorates and juries. Mere rank-order inputs falsify the true opinions of judges and voters. This has been recognized implicitly by the practitioners in skating, piano, wine, and other competitions (including ranking students in schools and universities), who have increasingly used grades. It has largely been ignored in elections. Two voters who place a candidate first—or second, or anywhere in their lists—may in fact evaluate the candidates completely differently (as ample evidence given in this book has shown). But with the exception of Australia and Ireland, very few nations even ask voters to input as much information as a rank-order. The inputs to the most used systems—notably, first-past-the-post and two-past-the-post—are single candidates, which conveys very little information concerning the opinions of voters. When the inputs are approvals and disapprovals, or 0s and 1s, only a pinch of extra information is elicited.

Common languages of grades are commonplace: they exist in myriad applications, though not heretofore in voting. Wines are a perfect example. "While some would suggest that scoring [grading] is not well suited to a beverage that has been romantically extolled for centuries, wine is no different from any other consumer product. There are specific standards of quality that full-time wine

professionals recognize, and *there are benchmark wines against which all others can be judged*" (Parker 2002, 3, our emphasis). This affirms the existence of evaluations that have developed over time and have become absolute evaluations. This ideal is true in many other competitions as well, and we believe it can also be realized in voting.

The evidence of the Orsay experiment shows that a common language exists—or may be created—for voting. Ideally, voters know exactly what each grade of the language means: two voters who assign a *Good* to a candidate have identical meanings in mind. But that, of course, cannot be true in practice. It is not true of any language: when one person announces to another that her dress is blue, the other has only an approximate idea of what that means, for it may be a pale blue or a deep blue or a blue-green (a "grue"), and so on. In fact, the words used to describe colors in different cultures have been the subject of intense study and debate. From the World Color Survey's study of 110 unwritten languages, Berlin and Kay (1969) formulated these hypotheses: "(1) the existence of universal constraints on cross-language color naming, and (2) the existence of a partially fixed evolutionary progression according to which languages gain color terms over time" (Kuehni 2007, 151). The paths of the development of new words for finer distinctions among colors differ, but the six colors identified as the fundamental colors by Ewald Hering in the nineteenth century—black, white, yellow, blue, red, green—occur more frequently than others, though not consistently. There is no question that within a culture, the words for colors carry the same meanings. These studies suggest that words for colors carry universal meanings. There is a linear spectrum of an infinity of colors. Yet people have used a scale of the same six words having the same meanings to differentiate among them. Why should the existence of a scale of grades to evaluate merit having the same meaning in one or another competition be rejected ?

It is incontestable that inputs given in the six-word language of grades *Excellent*, *Very Good*, *Good*, *Acceptable*, *Poor*, and *To Reject* convey much more precise common meanings within a culture than the input of a rank-order, the name of one candidate, or the names of several candidates. Common sense, the observation of practice (in skating, diving, wines), and the fact that human beings can and do communicate even very sophisticated ideas with the words of a language, together suggest that the most reasonable hypothesis is that a population that shares a common culture and language does understand words in essentially the same way. In any case their definitions are clearly given in dictionaries.

Indeed, why, a priori, should such understandings be denied across cultures rather than be universal? Reasonably accurate translations from one language to another may be found in dictionaries. Miller's analysis concludes that the capacity of men and women to make absolute judgments on a unidimensional

scale is limited to some 7 levels ± 2. At its end he reminds readers, "What about the seven wonders of the world, the seven seas, the seven deadly sins, the seven daughters of Atlas in the Pleiades, the seven ages of man, the seven levels of hell, the seven primary colors, the seven notes of the musical scale, and the seven days of the week?" (1956, 97). We believe that there *is* an optimal number for each application, there *is* a right way of defining a common language of grades for every application. The choice is not arbitrary. Six or seven seems to be a particularly apt number for human cognition in a scale of merit.

The majority judgment was used officially by the Nieman Foundation of Harvard University to discern the *Louis Lyons Award for Conscience and Integrity in Journalism* in November 2009. The jury was composed of nineteen Nieman Fellows. Five journalists (or groups of journalists) were the nominees. All of them were very highly considered. As a consequence, the Nieman Fellows chose the following common language of seven grades: *Absolutely Outstanding, Outstanding, Excellent, Very Strong, Strong, Commendable*, and *Neutral*. The winner's majority-grade was *Absolutely Outstanding*, two nominees' majority-grades were *Outstanding*, and two were *Excellent*. Five of the nineteen judges gave their highest grade to more than one nominee; three gave no *Absolutely Outstanding*; *Outstanding* was the lowest grade assigned by five; and exactly one judge gave different grades to all candidates (so only one rank-ordered them). This confirms the qualitative behavior of the voters in the Orsay experiment. Once again, the traditional inputs are inadequate; they do not model reality.

Common languages of grades used by professionals to judge competitions among sportsmen or products can be much richer, so long as they are well defined and well understood. It is important to realize that while the mathematical model developed in this book sometimes assumes an infinite language, all of the *properties* carry over to finite languages. Technically, an infinite language is sometimes, but not always, necessary to be able to give complete *characterizations*.

Imagining an ideal common language of grades with which every voter can perfectly express her opinions is like imagining a frictionless world in physics. "We have seen that this law of inertia [1] cannot be derived directly from experiment, but only by speculative thinking consistent with observation. The idealized experiment can never be actually performed, although it leads to a profound understanding of real experiments" (Einstein and Infeld 1938, 8–9). Common languages of grades are observed in practice. In voting, the 2007

1. "Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon" (Einstein and Infeld 1938, 8).

Orsay experiment (as well as other experiments) shows that given options with which to express themselves, voters with a shared cultural background use those options in the same way. When sets of several grades are aggregated over many candidates to eliminate individual voters' political biases, the distributions of grades obtained are close to one another. This is true of the grades in the Orsay voting experiment as it is of the grades used in the *Citadelles du Vin* competition. Mathematically, that seems to be the best one can do to show that the language is common. In practice, every audience to whom the distributions of the grades of the candidates in the Orsay experiment were presented anonymously—that is, with no names attached to the distributions—were able to identify the four major ones (Sarkozy, Royal, Bayrou, and Le Pen), suggesting that the grades are meaningful, they make sense, they do constitute a common language.

It has been contended by some that the majority judgment ballot is too complex, that it asks too much of voters. We do not believe so. On the contrary, the inputs are arguably the simplest for voters. First, because the ballot has greater cognitive simplicity: it is very natural to evaluate the merits of candidates in a scale defined by words. Second, because every electoral experiment to date has shown that voters were able to fill out the ballots quickly, and moreover appreciate being able to better express their opinions. It is clear that for voters it is easier to evaluate candidates in a natural scale of six grades than to rank-order candidates, and that it permits them to express their opinions more accurately. Some contend that the first- and two-past-the-post methods are by far the simplest for voters to understand. We do not believe so. Although it is easy enough to name one candidate, the subtleties of Arrow's and Condorcet's paradoxes engendered by these methods are not well understood by voters. In addition, voters are frustrated because they are unable to express their opinions concerning all the candidates. Instead they are faced with a stark strategic choice that is more difficult to resolve than it is to give one's honest evaluations of the several candidates.

Given a common language of grades for voters or judges to declare inputs, how are the inputs to be aggregated? "Physical concepts are free creations of the human mind, and are not, however it may seem, uniquely determined by the external world. In our endeavor to understand reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations. He will never be able to compare his picture with the real mechanism and he cannot even imagine the possibility or the meaning of such a comparison. But he

certainly believes that, as his knowledge increases, his picture of reality will become simpler and simpler and will explain a wider and wider range of his sensuous impressions. He may also believe in the existence of the ideal limit of knowledge and that it is approached by the human mind. He may call this limit the objective truth" (Einstein and Infeld 1938, 33).

Concepts to determine the properties of good mechanisms for aggregating opinions are also free creations of the human mind. They are determined by common sense, ethics, realities of human behavior, the limits of meaningful measurement, and the qualitative properties of actual results. But, as in physics, "To draw quantitative conclusions we must use the language of mathematics. Most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in language comprehensible to everyone. To follow up these ideas demands the knowledge of a highly refined technique of investigation. Mathematics as a tool of reasoning is necessary if we wish to draw conclusions which may be compared with experiment" (Einstein and Infeld 1938, 29).

Ethics imposes that a mechanism must be neutral and anonymous (or impartial) and obey the will of majorities. Common sense imposes that a mechanism must satisfy unanimity. These are the essential principles, the rock-bottom necessities. Four major arguments single out the majority judgment as the one method to be used:

Human Behavior Voters and judges may manipulate by sending input messages not in keeping with their true opinions to try to tilt the outcomes to their advantage. This imposes the need for mechanisms that doom manipulation to failure, or if that ideal is impossible, that best resist the attempt to manipulate. This idea has different formulations—strategy-proof-in-grading, strategy-proof-in-ranking (unattainable by any mechanism), partially strategy-proof-in-ranking, the manipulability of mechanisms—yet the mathematics shows that each, together with the essential principles, singles out the majority judgment.

Gaming the Vote Some believe that elections should be viewed strictly as pure games: every voter seeks to maximize his utility only. Some even suggest that voters should be encouraged to view elections as pure games. What does this imply? The only analyses that have been carried through to date postulate winner optimizers, voters whose utilities depend only on the winner (which in reality is quite simply false). Essentially all methods—with the exception of Borda's and more generally sum-scoring methods—elect a Condorcet-winner as a strong-equilibrium solution in this context. Thus essentially all methods (though not Borda's) are Condorcet-consistent when elections are pure games. However, the voters' inputs are those that maximize their utilities (measured

only by who wins), not their true evaluations; indeed, their inputs may be quite different from the true expression of their opinions, clearly a bad consequence. On the other hand, the majority judgment elects a Condorcet-winner in a strong equilibrium as well, but with his true majority-grade; moreover, every candidate receives a majority of true grades. But no method elects a Condorcet-winner with the voters' true opinions unless it heeds who gave what grade (theorems 20.6 and 20.7), in which case Arrow's and Condorcet's paradoxes rear their ugly heads.[2]

In the context of a game closer to reality, voters' utilities depend on pairs consisting of the winner and his final grade. The majority judgment elects the Condorcet-judgment-winner with his honest majority-grade.

Meaningfulness A claim measured by some representation of a scale of evaluation is meaningful if the same claim is true measured by any other representation of that scale. A common language of grades is an ordinal scale, so language-consistency and order-consistency (together with the basic ethical and commonsense properties) single out the majority judgment as the one possible method; and when there is no common language, Arrow's theorem shows there can be no preference-consistent method, no meaningful method.

Practice The 2007 Orsay field experiment constitutes an exceptionally interesting database for the study of different voting methods. It shows the existence of a statistical left-right spectrum. It shows that when a centrist candidate may be clearly identified, the well-known methods of voting may be ordered according to their biases against or for the centrist candidate. The majority judgment and Condorcet's are least biased for or against the centrist; Borda's, point-summing, and approval judgment (approval meaning *Good* or better) are very biased for the centrist; first- and two-past-the-post, approval judgment (approval meaning either *Very Good* or better, or *Excellent*) are very biased against the centrist. The data also permit the findings concerning manipulability to be confirmed experimentally. We believe that these conclusions are robust—though more experimentation should be pursued—and that they could not have been obtained with laboratory experiments where it is impossible to capture the rich complexity of the motivations of real voters.

"Fundamental ideas play the most essential role in forming a physical theory. Books on physics are full of complicated mathematical formulae. But thought

2. Thus when Condorcet-consistency is claimed for a method, beware! The statement may not mean what you think it means.

and ideas, not formulae, are the beginning of every physical theory. The ideas must later take the mathematical form of a quantitative theory, to make possible the comparison with experiment" (Einstein and Infeld 1938, 291). So, too, in the social sciences: substitute the words "social choice" for the words "physics" and "physical" and the statements maintain all their meanings.

There is, in essence, one new idea, one change in view: a simple model of how voters and judges express their opinions by evaluating the merits of candidates or competitors in a common language of grades rather than comparing them. Once the model is in hand, the rest follows from the blueprint provided by the classical theory of social choice.

Einstein and Infeld recall a fundamental truth:

> Science is not and will never be a closed book. Every important advance brings new questions. Every development reveals, in the long run, new and deeper difficulties. (1938, 308)

There is no doubt in our minds, however, that the change in view immensely improves the representation of reality, permits deeper understanding and analysis, and so leads to a vastly better mechanism for juries and electorates to rank and to elect.

References

Adams, C. 2008. Using a visual analog pain scale. <http://ergonomics.about.com/od/ergonomicbasics/ss/painscale.htm>.

Akerlof, G. A., and R. J. Shiller. 2009. *Animal spirits: How human psychology drives the economy and why it matters for global capitalism.* Princeton, N.J.: Princeton University Press.

American Federation of Mineralogical Societies. 2008. Mohs scale of mineral hardness. <http://www.amfed.org/t_mohs.htm>.

Arrow, K. J. 1951. *Social choice and individual values.* New Haven, Conn.: Yale University Press. 2d ed. New York: Wiley, 1963.

Aumann, R. J. 1959. Acceptable points in general cooperative *n*-person games. In *Contributions to the theory of games,* vol. 4, ed. R. D. Luce and A. W. Tucker, 287–324. Annals of Mathematics Studies 40. Princeton, N.J.: Princeton University Press.

Balinski, M. L. 1991. Gaspard Monge: pour la patrie, les sciences et la gloire. In *Mathématiques appliquées aux sciences de l'ingénieur,* ed. C. Carasso, C. Conca, R. Correa, and J.-P. Puel, 21–37. Toulouse: Cépaduès-Éditions.

———. 2004. *Le suffrage universel inachevé.* Paris: Éditions Belin.

———. 2008. Fair majority voting (or how to eliminate gerrymandering). *American Mathematical Monthly* 115: 97–113.

Balinski, M. L., and G. Demange. 1989a. Algorithms for proportional matrices in reals and integers. *Mathematical Programming* 45: 193–210.

———. 1989b. An axiomatic approach to proportionality between matrices. *Mathematics of Operations Research* 14: 700–719.

Balinski, M. L., A. Jennings, and R. Laraki. 2009. Monotonic incompatibility between electing and ranking. *Economics Letters* 105: 145–147.

Balinski, M. L., and R. Laraki. 2007a. A theory of measuring, electing and ranking. *Proceedings of the National Academy of Sciences* 104: 8720–8725. Issued in a more extensive version as Cahier 2006-11, Laboratoire d'Économétrie, École Polytechnique. November 26, 2006.

———. 2010. Election by majority judgment: Experimental evidence. In *In situ and laboratory experiments on electoral law reform: French presidential elections*, ed. B. Dolez, B. Grofman, and A. Laurent, to appear. Berlin-Heidelburg-New York: Springer. First issued as Cahier 2007–28, Laboratoire d'Économétrie, École Polytechnique, December 17, 2007.

Balinski, M. L., R. Laraki, J.-F. Laslier, and K. van der Straeten. 2003. Le vote par assentiment: une expérience. Cahier 2003-013. Laboratoire d'Économétrie, École Polytechnique.

Balinski, M. L., J.-F. Laslier, and K. van der Straeten. 2002. Compte-rendu d'une expérience de vote à l'Institut d'Études Politiques de Paris. Working paper. March 20.

Balinski, M. L., and F. Pukelsheim. 2006. Matrices and politics. In *Festschrift for Tarmo Pukkila,* ed. E. P. Liski, J. Isotalo, J. Niemel a, S. Puntanen, and G.P.H. Styan, 233–242. Tampere, Finland: University of Tampere.

Balinski, M. L., and S. T. Rachev. 1997. Rounding proportions: Methods of rounding. *The Mathematical Scientist* 22: 1–26.

Balinski, M. L., and V. Ramirez. 1999. Mexico's 1997 apportionment defies its electoral law. *Electoral Studies* 18: 117–124.

Balinski, M. L., and H. P. Young. 1982. *Fair representation: Meeting the ideal of one-man, one-vote.* New Haven, Conn.: Yale University Press. 2d ed. Washington, D.C.: Brookings Institution Press, 2001.

Barberà, S., and M. Jackson. 1994. A characterization of strategy-proof social choice functions for economies with pure public goods. *Social Choice and Welfare* 11: 241–252.

Barberà, S., H. Sonnenschein, and L. Zhou. 1991. Voting by committees. *Econometrica* 59: 595–609.

Bartoszyński, R. 1972. Power structure in dichotomous voting. *Econometrica* 40: 1003–1019.

Basset, G. W., and J. Persky. 1999. Robust voting. *Public Choice* 99: 299–310.

Baujard, A., and H. Igersheim. 2007. Expérimentation du vote par note et du vote par approbation lors de l'élection présidentielle française du 22 avril 2007 (rapport final). Paris: Centre d'analyse stratégique.

Berlin, B., and P. Kay. 1969. *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

Black, D. 1958. *The theory of committees and elections.* Cambridge: Cambridge University Press.

Blackorby, C., W. Bossert, and D. Donaldson. 2005. *Population issues in social choice theory, welfare economics, and ethics.* Cambridge: Cambridge University Press.

Blackorby, C., D. Donaldson, and J. A. Weymark. 1984. Social choice with interpersonal utility comparisons: A diagrammatic introduction. *International Economic Review* 25: 327–356.

Black's Law Dictionary. 2001. 2d pocket ed. St. Paul, Minn.: West Group.

Blouin, J. 2008. Citadelles du vin: une dégustation plus près du consommateur. Commentary on 2008 competition. Private communication.

Bogomolnaia, A., H. Moulin, and R. Strong. 2005. Collective choice under dichotomous preferences. *Journal of Economic Theory* 122: 165–184.

Borda, J.-C., Le Chevalier de. 1784. Mémoire sur les élections au scrutin. In *Histoire de l'Académie royale des sciences: Année 1781,* 657–665. A footnote states that the ideas were presented before the Academy on June 16, 1770.

Bossert, W., and J. A. Weymark. 2004. Utility in social choice. In *Handbook of utility theory,* vol. 2: *Extensions,* ed. S. Barberà, P. J. Hammond, and C. Seidl, 1099–1177. Boston: Kluwer.

Brams, S. J., and P. C. Fishburn. 1978. Approval voting. *American Political Science Review* 72: 831–847.

———. 1983. *Approval voting.* Boston: Birkhauser.

———. 2001. A nail-biting election. *Social Choice and Welfare* 18: 409–414.

Brams, S. J., and M. R. Sanver. 2006. Critical strategies under approval voting: Who gets ruled in and ruled out. *Electoral Studies* 25: 287–305.

Carroll, L. 1871. *Through the looking glass.* London: Macmillan.

———. 1916. *Alice's adventures in wonderland.* New York: Sam'l Gabriel Sons. Originally published London: Macmillan, 1865.

Chernoff, H. 1954. Rational selection of decision functions. *Econometrica* 22: 422–443.

Colegrove v. Green. 1946. 328 U.S. 549.

Condorcet, Le Marquis de. 1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Paris: l'Imprimerie royale.

———. 1789. Sur la forme des élections. Pamphlet, sec. 12, 25–26. Also in *Oeuvres de Condorcet* (1847), ed. A. Condorcet O'Connor and M. F. Arago. Paris: Firmin Didot frères.

Conseil constitutionnel. 1986. Decision no. 86-208 DC. July 2.

Copeland, A. H. 1951. A "reasonable" social welfare function. Seminar on Applications of Mathematics to the Social Sciences. University of Michigan, Ann Arbor.

Court of Arbitration for Sport. 2004. Arbitral award. Lausanne. October 21.

Dantzig, G. B. 1963. *Linear programming and extensions.* Princeton, N.J.: Princeton University Press.

Darwin, C. 1958. The autobiography of Charles Darwin, 1809–1882, with the original omissions restored. Edited and with appendix and notes by his granddaughter Nora Barlow. London: Collins. Paperback ed. New York: W.W. Norton, 1969.

Dasgupta, P., and E. Maskin. 2004. The fairest vote of all. *Scientific American* 290 (March): 92–97.

———. 2008. On the robustness of majority rule. *Journal of the European Economics Association* 6: 949–973.

d'Aspremont, C., and L. Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44: 199–209.

De Sinopoli, F., B. Dutta, and J.-F. Laslier. 2006. Approval voting: three examples. *International Journal of Game Theory* 35: 27–38.

Debuigne, G. 1970. *Larousse des vins.* Paris: Librairie Larousse.

Déclaration des droits de l'homme et du citoyen. 1789.

Décret. 1813. Projet de décret relatif à l'organisation particulière du commerce des vins à Paris. December 14. <http://www.napoleonica.org/gerando/GER03194.html>.

Dodgson, C. L. 1873. A discussion of the various methods of procedure in conducting elections. In *The theory of committees and elections,* by D. Black. Cambridge: Cambridge University Press, 1958.

———. 1874. Suggestions as to the best method of taking votes, where more than two issues are to be voted on. In *The theory of committees and elections,* by D. Black. Cambridge: Cambridge University Press, 1958.

———. 1876. A method of taking votes on more than two issues. In *The theory of committees and elections,* by D. Black. Cambridge: Cambridge University Press, 1958.

———. 1884. *The principles of parliamentary representation.* London: Harrison and Sons. *Supplement*. Oxford: E. Baxter, 1885.

Downs, A. 1957. *An economic theory of democracy.* New York: Harper.

Dubey, P., and J. Geanakoplos. 2006. Grading in games of status: Marking exams and setting wages. Cowles Foundation Discussion Paper No. 1544R, January.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics* 27: 642–669.

Economist Intelligence Unit (EIU). 2005. The Economist Intelligence Unit's quality-of-life index. London: EIU.

Einstein, A., and L. Infeld. 1938. *The evolution of physics.* Cambridge: Cambridge University Press.

Elman, B. A. 2000. *A cultural history of civil examinations in Late Imperial China.* Berkeley: University of California Press.

Emerson, J. 2006. Analyses. <http://www.stat.yale.edu/ jay/>.

Enelow, J. M., and M. J. Hinich. 1984. *The spatial theory of voting: An introduction.* Cambridge: Cambridge University Press.

Estlund, D. M. 2008. *Democratic authority: A philosophical framework.* Princeton, N.J.: Princeton University Press.

European Skating Championships. 2006. Lyon, France. January. Results. <http://www.isufs.org/events/fsevent00008670.htm>.

Faces Pain Scale—Revised. 2010. <http://www.usask.ca/childpain/fpsr/>.

Farquharson, R. 1969. *The theory of voting.* New Haven, Conn.: Yale University Press.

Farvaque, E., H. Jayet, and L. Ragot. 2007. Quel mode de scrutin pour quel "vainqueur"? Une expérience sur le vote préférentiel transférable. Working paper. Laboratoire Équippe, Université de Lille. May.

Fédération Internationale de Natation (FINA). 2005. Diving rules 2005–2009. <http://www.fina.org/>.

Felsenthal, D., and M. Machover. 2008. The majority judgement voting procedure: A critical evaluation. *Homo oeconomicus* 25: 319–334.

Fishburn, P. 1982. Monotonicity paradoxes in the theory of elections. *Discrete Applied Mathematics* 4: 119–134.

———. 1984. Discrete mathematics in voting and group choice. *SIAM Journal of Algebraic and Discrete Methods* 5: 263–275.

Fleurbaey, M., and F. Maniquet. 2008. Utilitarianism versus fairness in welfare economics. In *Justice, political liberalism, and utilitarianism,* ed. M. Fleurbaey, M. Salles, and J. Weymark, 263–280. Cambridge: Cambridge University Press.

Frederick Chopin International Piano Competition. 2006. <http://www.chopin.pl/>.

Galton, F. 1907a. One vote, one value. *Nature* 75: (February 28): 414.

———. 1907b. Vox populi. *Nature* 75 (March 7): 450–451.

Geanakoplos, J. 2005. Three brief proofs of Arrow's impossibility theorem. *Economic Theory* 26: 211–215.

General Electric Company. 2000. Annual Report. <http://www.ge.com/annual00/letter/index.html>.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41: 587–601.

Goodin, R., and K. Roberts. 1975. The ethical voter. *American Political Science Review* 69: 926–928.

Grofman, B., and S. L. Feld. 2004. If you like the alternative vote (a.k.a. the instant runoff) then you ought to know about the Coombs rule. *Electoral Studies* 23: 641–659.

Grote, D. 2005. *Forced ranking: Making performance management work.* Boston: Harvard Business School Press.

Hägele, G., and F. Pukelsheim. 2001. Llull's writings on electoral systems. *Studia Lulliana* 41: 3–38.

———. 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. G. Christianson, T. M. Izbicki, and C. M. Bellitto, 229–249. Washington, D.C.: Catholic University of America Press.

Hammond, P. 1976. Equity, Arrow's conditions, and Rawls' difference principle. *Econometrica* 44: 793–804.

Herberger College. 2006. International piano competition rules and regulations. <http://herbergercollege.asu.edu/pianocompetition/rules.php>.

Hicks, C. L., C. L. von Baeyer, P. Spafford, I. van Korlaar, and B. Goodenough. 2001. The faces pain scale revised: Toward a common metric in pediatric pain measurement. *Pain* 93: 173–183.

Hilgevoord, J., and J. Uffink. 2008. The uncertainty principle. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta. <http://plato.stanford.edu/archives/fall2008/entries/qt-uncertainty/>.

Hillinger, C. 2004. Utilitarian collective choice and voting. Discussion paper 2004-25. Department of Economics, University of Munich, Germany.

Holcombe, R. G., and L. W. Kenny. 2007. Evidence on voter preferences from unrestricted choice referendums. *Public Choice* 131: 197–215.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Inada, K. 1969. The simple majority decision rule. *Econometrica* 37: 490–506.

International Skating Union (ISU). 1998. New ISU figure skating results system. Communication No. 997. October 30.

———. 2002. Calculations for Olympic winter games pairs skating. <http://www.icecalc.com/events/owg2002/results/>.

———. 2004. Special regulations: Single and pair skating, as accepted by the 50th ordinary congress, June. <http://www.isu.org/>.

International Wine and Spirit Competition (IWSC). 2006. IWSC promotional leaflet. <http://www.iwsc.net/>.

Jennings, A. 2009. Weakly monotonic aggregation functions. Working paper. Mathematics Department, Arizona State University, Tempe. September 4.

Julia, D. 1990. Gaspard Monge, examinateur. *Histoire de l'éducation,* no. 46, 111–133.

Kemeny, J. 1959. Mathematics without numbers. *Daedalus* 88: 571–591.

———. 1962. Preference rankings: An axiomatic approach. In *Mathematical models in the social sciences,* ed. J. G. Kemeny and J. L. Snell, 9–23. Boston: Ginn and Co.

Kim, S.-R. 1990. On the possible scientific laws. *Mathematical Social Sciences* 20: 19–36.

Kintsch, W., and J. T. Cacioppo. 1994. Introduction to the 100th anniversary issue of the *Psychological Review. Psychological Review* 101: 195–199.

Kirkpatrick v. Preisler. 1969. 394 U.S. 526.

Koc, E. W. 1988. An experimental examination of approval voting under alternative ballot conditions. *Polity* 20: 688–704.

Krantz, D. H., R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement.* Vol. 1. New York: Academic Press.

Kuehni, R. G. 2007. Nature and culture: An analysis of individual focal color choices in World Color Survey languages. *Journal of Cognition and Culture* 7: 151–172

Kuhn, T. S. 1961. The function of measurement in modern physical science. *Isis* 52: 161–190. Reprinted in *The Essential Tension,* 178–224. Chicago: University of Chicago Press, 1977.

———. 1970. *The structure of scientific revolutions.* 2d ed. enlarged. Chicago: University of Chicago Press. Originally published in 1962.

Kurrild-Klitgaard, P. 1999. An empirical example of the Condorcet paradox of voting in a large electorate. *Public Choice* 107: 1231–1244.

Laplace, P.-S., Marquis de. 1820. *Théorie analytique des probabilités.* 3d ed. Paris: Courcier. Originally published 1812.

Laraki, R. 2009. Coalitional equilibria of strategic games. Cahier 2009-42. Laboratoire d'Économétrie, École Polytechnique. October.

Laslier, J.-F. 2009. The leader rule: A model of strategic approval voting in a large electorate. J*ournal of Theoretical Politics* 21: 113–136.

Laslier, J.-F., and K. van der Straeten. 2004. Vote par assentiment pendant la présidentielle 2002: Analyse d'une expérience. *Revue Française de Science Politique* 54: 99–130.

Le vote de valeur: Pour renforcer l'acte démocratique. 2007. <http://www.votedevaleur.info/co/pres.html>.

London, J., and I. McLean. 1990. The Borda and Condorcet principles: Three medieval applications. *Social Choice and Welfare* 7: 99–108.

Loosemore, S. 1997. If it ain't broke, don't fix it: An analysis of the figure skating scoring system. <http://www.frogsonice.com/skateweb/obo/score-tech.shtml>.

Markham, D. Jr. 1997. *1855, Histoire d'un classement des vins de Bordeaux.* Bordeaux: Editions Féret.

Martin, J. 2002. Aux origines de la "science des examens" (1920–1940). In *L'examen: Évaluer, sélectionner, certifier XVIe–XXe siècles,* ed. B. Belhoste, 177–199. Paris: Institut national de recherche pédagogique.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. *Microeconomic theory.* Oxford: Oxford University Press.

Maskin, E. 1999. Nash implementation and strong Nash equilibria. *Review of Economic Studies* 66: 23–38.

Mason, W. 2006. Judging brief for 2006 competition's jury. <http://www.top100wines.com/main/default.asp>.

May, K. O. 1952. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica* 20: 680–684.

Merrill, S. III. 1988. *Making multicandidate elections more democratic.* Princeton, N.J.: Princeton University Press.

Merrill, S. III, and J. Nagel. 1987. The effect of approval balloting on strategic voting under alternative decision rules. *American Political Science Review* 8: 509–524.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81–97.

Mondovino. 2004. Documentary film. Dir. J. Nossiter.

Mosteller, F., and J. W. Tukey. 1977. Data analysis and regression. Reading, Mass.: Addison-Wesley.

Moulin, H. 1980. On strategy-proofness and single peakedness. *Public Choice* 35: 437–455.

———. 1988. *Axioms of cooperative decision making.* Cambridge: Cambridge University Press.

Moulin, L. 1953. Les origines religieuses des techniques électorales et délibératives modernes. *Revue internationale d'histoire politique et constitutionnelle* 3 (N.S.): 106–148.

Muller, E., and M. Satterthwaite. 1977. The equivalence of strong positive association and strategy-proofness. *Journal of Economic Theory* 14: 412–418.

Myerson, R. B. 1998. Population uncertainty and Poisson games. *International Journal of Game Theory* 27: 375–392.

———. 2000. Large Poisson games. *Journal of Economic Theory* 94: 7–45.

———. 2002. Comparison of scoring rules in Poisson voting games. *Journal of Economic Theory* 103: 219–251.

Myerson, R. B., and R. J. Weber. 1993. A theory of voting equilibria. *American Political Science Review* 87: 102–114.

Nagel, J. H. 2006. A strategic problem in approval voting. In *Mathematics and democracy: Recent advances in voting systems and collective choice,* ed. F. Pukelsheim and B. Simeone, 133–150. New York: Springer.

———. 2007. The Burr dilemma in approval voting. *Journal of Politics* 69: 43–58.

Nanson, E. J. 1882. Methods of election. *Transactions and Proceedings of the Royal Society of Victoria* 18: 197–240.

Narens, L., and R. D. Luce. 2008. Meaningfulness and invariance. In *New Palgrave Dictionary of Economics*, 2d ed., ed. S. N. Durlauf and L. Blume. Basingstoke, U.K.: Palgrave Macmillan.

Nash, J. F. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Neumann, J. von, and O. Morgenstern. 1944. *Theory of games and economic behavior.* Princeton, N.J.: Princeton University Press.

Nùñez Rodríguez, M. 2008. Questions stratégiques en théorie du vote (Strategic questions in voting theory). Doctoral thesis, École Polytechnique, November.

Nurmi, H. 2004. Monotonicity and its cognates in the theory of social choice. *Public Choice* 121: 25–49.

O'Brian, Patrick. 1969. *Master and commander.* Philadelphia: Lippincott. Paperback ed. New York: W.W. Norton, 1990.

Organisation Internationale de la Vigne et du Vin (OIV). 1994. Standard for international wine competitions. <http://www.oiv.int/>.

———. 2009. OIV standard for international wine and spirituous beverages of vitivinicular origin competitions. Resolution OIV/Concours 332A/2009. <http://www.oiv.int/>.

Orlov, A. 1981. The connection between mean quantities and admissible transformations. *Mathematical Notes* 30: 774–778.

Parker, R. M. Jr., with P.-A. Rovani. 2002. *Parker's wine buyer's guide.* 6th ed. New York: Simon and Schuster.

Pennisi, A. 2006. The Italian bug: A flawed procedure for bi-proportional seat allocation. In *Mathematics and democracy: Recent advances in voting systems and collective choice,* ed. F. Pukelsheim and B. Simeone, 133–150. New York: Springer.

Pennisi, A., F. Ricca, P. Serafini, and B. Simeone. 2007. Amending and enhancing electoral laws through mixed integer programming: The case of Italy. In *Proceedings of the VIII International Conference on Economic Modernization and Public Development*, 1–10. Moscow: Higher School of Economics. <http://conf.hse.ru/lingua/en/2007/>.

Peynaud, É., and J. Blouin. 1999. *Découvrir le goût du vin.* Paris: Dunod.

———. 2006. *Le goût du vin: le grand livre de la dégustation.* 4th ed. Paris: Dunod.

Pfanzagl, J. 1971. *Theory of measurement.* Vienna: Physica-Verlag.

Piéron, H. 1963. *Examens et docimologie.* Paris: PUF.

Pliny the Elder. 1855. *The natural history.* Trans. J. Bostock and H. T. Riley. Book 14, chs. 7 and 8. <http://www.perseus.tufts.edu/hopper/>.

Plott, C. R., J. T. Little, and R. P. Parks. 1975. Individual choice when objects have "ordinal" properties. *Review of Economic Studies* 42: 403–413.

Poundstone, W. 2008. *Gaming the vote: Why elections aren't fair.* New York: Hill and Wang.

Pukelsheim, F. 2006. BAZI. <http://www.math.uni-augsburg.de/stochastik/bazi/welcome.html>.

Quandt, R. E. 2006. Measurement and inference in wine tasting. *Journal of Wine Economics* 1: 7–30.

RangeVoting.org. 2007. <http://rangevoting.org/>.

Répertoire de jurisprudence. I. Égalité des salaires et ranking. 2002. No. 02-687. Grenoble: C.A., November 13.

Richter, C. F. 1935. An instrumental earthquake magnitude scale. *Bulletin of the Seismological Society of America* 25: 1–32.

Roberts, F. S. 1979. *Measurement theory, with applications to decision making, utility and the social sciences.* Vol. 7 of Encyclopedia of Mathematics and Its Applications. Reading, Mass.: Addison-Wesley.

Rudolph, F. 1977. *Curriculum: A history of the American undergraduate course of study since 1636.* San Francisco: Jossey-Bass.

———. 1991. *The American college and university: A history.* 2d ed. Athens, Ga.: University of Georgia Press.

Saari, D. G. 1989. A dictionary for voting paradoxes. *Journal of Economic Theory* 48: 443–454.

———. 1992. Millions of election rankings from a single profile. *Social Choice and Welfare* 9: 277–306.

———. 2000. Mathematical structure of voting paradoxes. I and II. *Economic Theory* 1: 51–53, 55–102.

———. 2001a. Analyzing a nail-biting election. *Social Choice and Welfare* 18: 415–430.

———. 2001b. *Chaotic elections! A mathematician looks at voting.* Providence, R.I.: American Mathematical Society.

———. 2009. Voting. <http://cema.cufe.edu.cn/admin/data/uploadfile/200907/2009071318372272.pdf>.

Saari, D. G., and J. van Newenhizen. 1988. Is approval voting an "unmitigated evil"? *Public Choice* 59: 133–147.

Satterthwaite, M. A. 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.

Scullen, S., P. Bergey, and L. Aiman-Smith. 2005. Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology* 58: 1–32.

Sen, A. 1970. *Collective choice and social welfare.* San Francisco: Holden-Day.

Sen, A., and P. Pattanaik. 1969. Necessary and sufficient conditions for rational choice under majority decision. *Journal of Economic Theory* 1: 178–202.

Sertel, M. R., and M. R. Sanver. 2004. Strong equilibrium outcomes of voting games are the generalized Condorcet winners. *Social Choice and Welfare* 22: 331–347.

Simon, H. 1954. Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly* 18: 245–253.

Smith, J. 1973. Aggregation of preferences with variable electorate. *Econometrica* 41: 1027–1041.

Stevens, S. S. 1946. On the theory of scales of measurement. *Science* 103: 677–680.

Tocqueville, A. de. 1967. Letter dated January 5, 1851. In *Correspondance d'Alexis de Tocqueville et de Gustave de Beaumont.* Vol. 2, 355. Introduced, edited, and annotated by A. Jardin. 3 vols. Paris: Gallimard.

United States Geological Survey. 1989. The severity of an earthquake. <http://earthquake.usgs.gov/learning/topics/mercalli.php>.

University Interscholastic League. 2006. <http://www.uil.utexas.edu/>.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Vieth v. Jubelirer. 2004. 541 U.S. 267.

Vinitaly. 2006. Regolamento. <http://www.vinitaly.it/concorsoenologico/home.asp>.

Weber, R. J. 1977. Comparison of public choice systems. Cowles Foundation discussion paper 498. Yale University, New Haven, Conn.

Webster, D. 1832. *The writings and speeches of Daniel Webster.* Boston: Little, Brown, 1903.

Wells v. Rockefeller. 1969. 394 U.S. 542.

Weymark, J. A. 2005. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare* 25: 527–555.

Wikipedia. 2007. Grade (education). <http://en.wikipedia.org/wiki/Grade_(education)>.

Wilderness Emergency Medical Services Institute. 2008. Mankoski pain scale. <http://wemsi.org/painscale.html>.

Wood, H. O., and F. Neumann. 1931. Modified Mercalli intensity scale of 1931. *Bulletin of the Seismological Society of America* 21: 277–283.

World Skating Federation, Plaintiff, v. International Skating Union and Ottavio Cinquanta, Defendants. 2005. 03 Civ. 9800 (JES), U.S. District Court, Southern District of New York, February 15.

Young, H. P. 1974. A note on preference aggregation. *Econometrica* 42: 1129–1131.

———. 1975. Social choice scoring functions. *SIAM Journal of Applied Mathematics* 28: 824–838.

———. 1986. Optimal ranking and choice from pairwise comparisons. In *Information pooling and group decision making,* ed. B. Grofman and G. Owen, 113–122. Greenwich, Conn.: JAI Press.

———. 1988. Condorcet's theory of voting. *American Political Science Review* 82: 1231–1244.

Young, H. P., and A. Levenglick. 1978. A consistent extension of Condorcet's election principle. *SIAM Journal of Applied Mathematics* 35: 285–300.

Zahid, M. A. 2009. Majority judgement theory and paradoxical results. Working paper. Tilburg University, Netherlands.

Zi, É. 1894. Pratique des examens littéraires en Chine. Shanghai: Imprimerie de la mission catholique.

Zitzewitz, E. 2006. Nationalism in winter sports judging and its lessons for organizational decision making. *Journal of Economics and Management Strategy* 15: 67–100.

Name Index

Subject Index

Formal definitions of concepts are found on pages set in **bold**.