The Journal of Political Philosophy: Volume 0, Number 0, 2020, pp. 1–23

Democracy without Enlightenment: A Jury Theorem for Evaluative Voting*

Michael Morreau
Philosophy, The Arctic University of Norway, Tromsø

SAY a jury is going to decide who wins a competition. First, each member evaluates all the competitors by grading them; then, for each competitor, a collective grade is derived from all the judgments of all the members; finally, the jury chooses as the winner the competitor with the highest collective grade. This is *collective grading*. The grades that are used might typically be numerical scores, or evaluative expressions of a natural language, such as "good," "fair," and "bad." They could be any signs at all, though, that come in a "top" to "bottom" order: thumbs up and down; happy, neutral, and sad emojis; or cheering, clapping, booing, and angry hissing at public events. Panels, boards, and committees throughout society evaluate all manner of things by grading them. Thus risks are prioritized, research proposals are funded, and candidates are shortlisted for jobs. Apart from acclamation in special cases, collective grading is not a usual way to pick winners in political elections.

This article takes up a question about the quality of judgments and decisions made by collective grading: under which conditions are outcomes likely to be right? An answer comes in the form of a jury theorem for *median grading*. Here, the collective grade for a thing is the median of its individually assigned grades—the one in the middle, when all of them are listed from top to bottom. Section III prepares the ground for this theorem by discussing different senses in which grades can be the *right* ones for things, or the wrong ones as the case may be, independently of which grades are assigned in the end. These notions of right and wrong are relevant to judgments of different kinds of things: risks, research proposals, job candidates, options in referendums and elections. The grading-jury theorem in Section V identifies conditions on the grading competence of individual people under which median grades, and decisions that follow them, are likely to be, independently, right.

*I thank for helpful comments and suggestions Fuad Aleskerov, Philip Ebert, David Estlund, Paul Grünke, Alan Hájek, Rida Laraki, Fredrik Nyseth, Attila Tanyi, Christopher Thompson, John Weymark, and three anonymous referees.

doi: 10.1111/jopp.12226

A second objective of this article is to suggest a solution to problems of voter ignorance in democracies. The idea developed here is to use voting methods that make more of people's limited knowledge than do traditional methods such as majority voting. Grading holds promise, because voters can express themselves more fully by grading the options on their ballots rather than simply choosing one or ranking them all. To count the one option "good" and the other "bad" is, for instance, to rank the first above the second; but there is more information in these grades than just this order, because while counting the first option instead "fair" would put them in the same order, the expressions "good" and "fair" mean different things. By tapping into the richer information carried by graded ballots, collective grading methods could, in principle, allow more of voters' knowledge to find its way into collective decisions than traditional voting methods do.

Median grading sometimes does make more of voters' knowledge than majority voting possibly can. Section VI draws from the grading-jury theorem the consequence that median grading is—in a special sense, presently explained—*forgiving* of the incompetence of voters who, perhaps as a result of their ignorance and prejudices, are not likely to make right decisions on their own. This agreeable nature of median grading is on display, in Section VI, in the example of an assembly that, going by median grades, reliably picks out the better option from a pair, even though the individual members, in their bemusement, are far more likely to vote for the wrong one. Under these circumstances, as Condorcet warned long ago, and as is explained in the next section, letting the majority decide is likely to make things worse. It would expose the assembly to a risk of making false decisions. The upshot is that, in theory anyway, and perhaps also in practice, median grading can enable unenlightened assemblies to "track the truth"—even as majority voting would run them off the rails.

I. THE RISK OF MAKING FALSE DECISIONS

Identifying good policy options requires knowledge of their advantages and disadvantages. Getting them accepted and implemented, and tracing responsibility for successes and failures, requires knowledge of how government works. People do not in general have much technical or political expertise, and critics of democracy since classical times have questioned the capacity of ordinary citizens and politicians to govern well. In the parable of the ship of state, told by the character of Socrates in Plato's Republic (Book VI), the captain has all the authority, but lacks knowledge needed to set the course. The sailors, wrangling over the helm, cannot even tell who is competent to take control. The point of the story is that only those with professional training and expertise are fit to rule the state.

The French Enlightenment brought a remarkable rebuttal. Say a jury, committee, or electorate is going to choose between two options, and that one of these is, irrespective of the outcome, the right one to choose. They could be policy

alternatives of which one really is better, in some decisive way, or candidates in an election, or what have you. Nicolas de Caritat, the marquis of Condorcet, discovered that, under certain conditions, the chance of a majority settling on this right option is greater than any given voter's chance of doing so. The larger the assembly the better and, with enough voters, the outcome of a majority vote is almost certainly right. This insight, made technically precise and demonstrated using the law of large numbers, is Condorcet's jury theorem.[1]

Condorcet's jury theorem puts paid to the notion that only experts are fit to govern. Majority voting can compound the modest knowledge of ordinary citizens, creating a greater knowledge that is of all the people. Democratic assemblies under favorable conditions reliably reach decisions that are as well informed as those of an elite few, while their inclusiveness and diversity promote liberty, equality, and political legitimacy.[2]

At the same time, Condorcet's jury theorem marks a precise point at which optimism about the wisdom of majority voting must give way to pessimism. One assumption is that individual voters are *minimally competent* to decide the question at hand. Each voter is more likely to choose the right option than the wrong one.[3] Now, this is critical. If it is the other way around, and individual voters are more likely to choose the wrong option, then, according to the jury theorem, the majority is even more likely to choose it, and the more voters there are the worse. Majority voting can compound the decision competence of citizens only if they have enough of it to begin with. Otherwise, it compounds their incompetence instead.[4]

Minimal competence might not seem much to expect. Asking a voter need be only slightly more reliable a way to discover the truth than tossing a coin! Even then, it is enough that the individual chance of choosing the right option is on average greater than ½, with some more enlightened voters making up for others who are less so.[5] In fact, people must often fail to clear even this seemingly low bar. Ignorance of relevant facts by itself can bring someone's chance of settling on

[1]For a technical statement, see Theorem 1 of Bernard Grofman, Guillermo Owen, and Scott L. Feld, "Thirteen theorems in search of the truth," *Theory and Decision*, 15 (1983), 261–78.

[2]Diversity makes a further epistemic contribution of its own. See Robert E. Goodin and Kai Spiekermann, *An Epistemic Theory of Democracy* (Oxford: Oxford University Press, 2018), esp. chs 7 and 8. It is cognitive diversity—in sources of information, perspectives, heuristics, and models of the world—that gives groups an epistemic edge. This is not the same as social diversity in people's gender, or their ethnicity or age. Still, extensions of the democratic franchise must often have increased both kinds of diversity at once.

[3]The other main assumption of Condorcet's jury theorem is that vote decisions are probabilistically independent. The chance of any given voter's settling on the right option is the same, whether or not any other voter does. The independence assumption need not be contentious in this "fixed problem" case, where there is some particular choice under consideration. See Franz Dietrich, "The premises of Condorcet's jury theorem are not simultaneously justified," *Episteme*, 5 (2008), 56–73.

[4]The mathematics of Condorcet's jury theorem does not distinguish in any substantive way right from wrong, or competence from incompetence. There are two options, people's chances of choosing them, and that is that. Now, the moment someone's chance of choosing any given one of their two options drops below ½, the chance of their choosing the other one rises above this critical mark. As far as the theorem is concerned, incompetence in choosing either option just *is* competence in choosing the other.

[5]See Theorem 5 of Grofman et al., "Thirteen theorems in search of the truth."

the right option all the way down to ½—the chance of just guessing which one is right—and biases and prejudices can drag it even lower than this. Ignorance, bias, and prejudice are common failings, and Condorcet warned of the epistemic risk they entail when majorities decide:

> A very numerous assembly cannot be composed of very enlightened men. It is even probable that those comprising this assembly will on many matters combine great ignorance with many prejudices. Thus there will be a great number of questions on which the probability of the truth of each voter will be below ½. It follows that the more numerous the assembly, the more it will be exposed to the risk of making false decisions.[6]

Plurality voting is a common way to decide when there are more options on the table than just two: an option wins if it gets more votes than any other single option does. Generalizing the jury theorem, pluralities track the truth when the individual chance of choosing the right option is for each voter greater than the chance of choosing any other single option.[7] Ignorance and prejudice can drag a voter's chance of choosing it down past ½, though, past the chances of choosing other options, and as close to nil as you like. Assemblies that decide by plurality voting among several options are at risk of making false decisions, no less than assemblies that decide by majority voting among two.

Democracy has spread throughout the world since Condorcet achieved his insights into its opportunities and risks, on the cusp of the French Revolution. Enlightenment has not. One recent survey covering 38 countries found that people everywhere are very mistaken about such factual matters as the effectiveness of vaccines, the proportion of foreigners in prisons, and trends in murder rates and deaths due to terrorism.[8] Nor are these findings unusual; on the contrary, they are consistent with the results of decades of research in the social sciences. Ignorance, bias, and prejudice have many causes, in human psychology and in the nature of mass communications and social media.[9] They put democracies at risk of making wrong decisions, no less now than in Condorcet's time.

[6]Marie-Jean-Antoine-Nicolas Caritat, marquis de Condorcet, "Essay on the application of mathematics to the theory of decision-making" [1785], *Condorcet: Selected Writings*, ed. Keith Michael Baker (Indianapolis: Bobbs-Merrill, 1976), pp. 33–70, at p. 49. Condorcet's own persecution, imprisonment, and death a few years later, in the Jacobin terror, makes this measured warning of the dangers of democracy without enlightenment all the more poignant.

[7]Christian List and Robert E. Goodin, "Epistemic democracy: generalizing the Condorcet jury theorem," *Journal of Political Philosophy*, 9 (2001), 277–306, proposition 1.

[8]Ipsos, "Perils of perception 2017," <https://www.ipsos.com/sites/default/files/ct/news/documents/2018-02/ipsos-mori-perils-of-perception-2017-charts_0.pdf>.

[9]Confronting people with facts that conflict with their ideological commitments can, perversely, strengthen misperceptions; see Brendan Nyhan and Jason Reifler, "When corrections fail: the persistence of political misperceptions," *Political Behavior*, 32 (2010), 303–30, fig. 1. Deliberation can cause beliefs to become extreme which, when the truth lies in the middle, must tend to create false beliefs; see Cass R. Sunstein, "The law of group polarization," *Journal of Political Philosophy*, 10 (2002), 175–95. Social media broadcast false news items more quickly than truths, and to more people; see Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, 359 (2018), 1146–51.

There are well-known ways to contain epistemic risk. One is to involve in decisions only people who have the relevant knowledge or, more generally, to allow those with a greater knowledge more say. This is common on panels and committees, but not in political elections, where it conflicts with the egalitarian ideal of "one person one vote."[10] Another familiar way to contain epistemic risk is to restrict the matters that people may decide to those in which they are likely to make good decisions. The role of citizens can be limited to electing representatives to govern on their behalf. Politically sensitive business, such as setting monetary policy and the interpretation of government performance statistics, is kept out of the hands of elected officials by entrusting it to civil servants, to avoid conflicts with electoral interests.

This article proposes another way to limit exposure to epistemic risk. It is to set up democratic institutions with decision methods that make more of voters' limited knowledge than do traditional methods such as majority voting.

To see what it could be for one decision method to tap into knowledge and make good use of it while another method does not, it will help to have some special vocabulary. Suppose there is a choice to be made. There are two options, or three, or more. They are policy alternatives, perhaps, or candidates in an election. Alternatively, the options are scores or grades, and the question is which of them to assign to some item that is under consideration. The nature of the options is unimportant, but they are some particular ones, and one of them is, independently, the right one to choose. An assembly of voters will make the choice, first individually and then collectively. Following Condorcet, in the cited passage, let a voter's *probability of truth* be their chance of choosing the right option on their own. Condorcet warned of voters whose probability of truth is below the *critical level*. This is the chance of choosing the right option at random: it is ½ with two options, the case Condorcet had in mind; ⅓ for a choice among three, and so on.

Consider now some method by which the assembly can make its choice. Call this method *reliable*, under given circumstances and for a given probability of truth, if a sufficiently large assembly, using this method under these circumstances, its individual members having this probability of truth, is likely to choose the right option. Call this method *forgiving* if there are *some* possible circumstances, and *some* probabilities of truth below the critical level, for which it is reliable. Forgiving choice methods put voters' knowledge to good use under some circumstances in which other methods, lacking this agreeable characteristic, do not.[11]

[10]Harry Brighouse and Marc Fleurbaey propose to replace "one person one vote" with a principle of proportionality, distributing power among people according to their stakes in decisions; see their "Democracy and proportionality," *Journal of Political Philosophy*, 18 (2010), 137–55. This entails distributing power roughly according to decision competence when, as must often be the case, people know more about decisions to the extent that their own interests are at stake.

[11]The forgiving nature of a voting method does not by itself entail that it is *voters'* knowledge that is used. One forgiving method, for instance, is to have a vote, but then ignore the result and refer the question to an omniscient and beneficent god, who knows what is right and chooses it no matter how everyone has voted. Assume that relevant choice methods share with majority voting the democratic feature that collective choices depend only on the people's votes: there is to be no change in an assembly's choice without a change in some member's vote.

Majority voting is a reliable method for choosing between two options under any circumstances, for any probability of truth above the critical ½ mark. This is Condorcet's jury theorem. But majority voting is not a forgiving method for binary choice. There are no circumstances under which majorities are likely to settle on the right option when each voter's probability of truth is below ½. This is the ground for Condorcet's warning about ignorant and prejudiced voters.[12] Majority voting has long been a hallmark of democracy. It is the method by which decisions were made in the citizens' assembly of ancient Athens, the *ekklesia*. Other voting methods are no less democratic, though, and some of them are forgiving.

Collective grading works, more precisely now, like this. There is given a scale or *language of grades*. This is a vocabulary of signs, the grades, that come in a fixed linear order (from "top" to "bottom").[13] An assembly has some items to grade and proceeds as follows. First, each member judges all the items by assigning to each one a grade in this language of grades. Then, for each item, a single collective grade is derived, using some suitable aggregation method, from all the judgments of it by the different members of the assembly.[14]

The main contribution of this article is the grading-jury theorem. It concerns *median grading*. This is that form of collective grading in which the collective grade for an item is derived by putting all the individual judgments of it in the top–bottom order of the grades, and then choosing the middlemost, or median one. The collective grade is the grade according to this median judgment.[15] Under the conditions of the grading-jury theorem, median-grading assemblies are more likely to assign right grades than are their individual members. The more members there are the better and, with enough of them, median grades almost certainly are right.

One consequence of the grading-jury theorem is that median grading is a forgiving method for choosing among a pair of options. Section VI demonstrates this with the example of an assembly whose members not only are strongly biased, but also have their biases crossed. So strong is the bias against the superior option, but in favor of the inferior one, that for each member separately the

[12]These claims about majority voting are made under the jury theorem's assumption that individual vote decisions are suitably independent. Where, on the other hand, votes are coordinated in just the right way, it can happen that majorities are most likely to settle on the right option, even though individual voters are most likely to choose the wrong one.

[13]Often, the grades are understood to express levels of approval, or of merit, but they can be interpreted with respect to any dimension along which items of one kind or another can be compared: likelihood, in the case of possible events; strength, in the case of storms; or what have you.

[14]Collective grades need not come from the same language of grades that individual voters use to provide their inputs. They don't with range voting, for instance, discussed at the end of Section IV. More generally, different voters could provide their inputs in different languages of grades. See Michael Morreau, "Supergrading: how diverse standards can improve collective performance in ranking tasks," *Theory and Decision*, 88 (2020), 541–65.

[15]The median is a suitable measure of centrality when scores and grades have an ordinal, not a cardinal significance. They express *levels* of something, not amounts. A restaurant with three Michelin stars, for instance, need not have three times the merit of a one-star restaurant (though it might). A three-liter bottle, in contrast, must have three times the capacity of a one-liter bottle.

chance of choosing the superior option is less than the critical ½. The outcome of a majority vote is under these circumstances, as Condorcet warned, most likely wrong. Members even so satisfy the conditions of the grading-jury theorem. Median grading is likely to deliver the right grades for the options, a higher grade for the superior option and a lower grade for the inferior one. Going by median grades, the assembly is likely to make the right choice.

One could read too much into the forgiving nature of median grading. By itself this means only that there are *some* fortunate circumstances—familiar and common circumstances, hopefully, but they could be arcane and only rarely encountered in real life—under which right collective judgments and decisions may be expected from unenlightened voters. It is important to realize that the extent of any real advantages from this depends on, among other things, how frequently these fortunate circumstances obtain in real panels, committees, and electorates (or, rather, on how frequently they *would* obtain, were relevant decisions framed as grading problems).

This article leaves the matter of practical implications, which is in large part empirical, to one side. Some present business touches indirectly on it, though. The unenlightened voters of Section VI, whose probability of truth is brought below ½ by their biases, are of a kind with Condorcet's ignorant and prejudiced men. Median grading presumably would limit the risk of false decision in some of the cases he had in mind. Computer modeling in Section VI suggests, furthermore, that, while having more voters is better, when the conditions of the grading-jury theorem obtain, the epistemic machinery of median grading is effective also in assemblies of a realistic size. In one simulation, 501 voters, though sorely unenlightened, are almost certain to make the right judgments and decisions. Many deliberative assemblies throughout history have been large enough for a powerful effect. The Athenian *boule* had 500 members (after 508 BCE). The US House of Representatives is somewhat smaller, and both houses of the UK parliament are larger.

Some political thinkers, echoing the misgivings of classical philosophers, propose to solve problems of voter ignorance by restricting the scope of democracy, or even abandoning democracy entirely. Thus Ilya Somin builds a case for smaller government on the fact of political ignorance,[16] and Jason Brennan argues for epistocracy, the rule of the knowledgeable.[17] The solution suggested here does not require cutting down on democracy or rolling it back. It raises the prospect of extending democracy, instead. With more effective use of their knowledge, imperfect as it might be, perhaps more people could participate in more decisions without too much exposure to the risk of making false decisions.

[16]Ilya Somin, *Democracy and Political Ignorance: Why Smaller Government Is Smarter* (Stanford: Stanford University Press, 2013).

[17]Jason Brennan, *Against Democracy* (Princeton: Princeton University Press, 2016).

II. COLLECTIVE GRADING

Collective grading is common in decision making by panels and committees. In the UK, the Arts and Humanities Research Council (AHRC) scores research proposals on a scale from *6*, their top score, down to *1*.[18] The Society for Social Choice and Welfare uses approval voting to elect members to its council. This is collective grading with just two grades, *Approved* and *Not Approved*.[19] Collective grading hasn't often been used for making political decisions, but there are a few examples from history. Candidates were elected to ancient Sparta's *gerousia*, the Council of Elders, by the loudness of voters' shouting in their support.[20] Here, the grades are shouts of varying loudness, naturally aggregated into collective expressions of approval and disapproval by the superposition of sound waves. The Venetian Republic used a complicated procedure for electing the doge, alternating rounds of random choice and approval voting.[21] More recently, Michel Balinski and Rida Laraki ran an experiment with collective grading in conjunction with the 2007 French presidential elections. Voters evaluated candidates on a scale of six grades: *Très Bien* (Excellent), *Bien* (Very Good), *Assez Bien* (Good), *Passable* (Acceptable), *Insuffisant* (Poor) and *à Rejeter* (To Reject).[22]

Collective grading is a good way to make decisions, quite apart from any epistemic advantages. For one thing, it guarantees *social ordering*: the outcome of voting about several options is always a ranking from top to bottom, perhaps with ties. The top-ranked options are those whose collective grade is highest; directly below them are any with the next-highest grade, and so on down. Social ordering ensures that it is possible to maximize, by choosing an option regarded second to none. It promotes rationality in decision making.

Majority voting, on the other hand, taking options pair by pair, does not always result in a ranking. Sometimes, depending on how everybody votes, there is for each option some other that finds majority support in a comparison of the two. Each option is second to some other and, with a finite number of them, the collective preference forms a cycle; there is no ranking from top to bottom, and no maximizing choice. This was another important discovery of Condorcet, known nowadays as the "paradox" of voting, or "Condorcet's paradox."[23]

Collective grading not only promotes collective rationality. It also embodies core democratic values. Median grading (among other grading methods) is *anonymous*. While the outcome of a vote depends on the inputs of all the people, it

[18]"Grading scale," *AHRC Peer Review Handbook* (2018), <https://ahrc.ukri.org/funding/research/researchfundingguide/peerreview/peerreviewgradingscale/>.

[19]Steven J. Brams and Peter C. Fishburn, *Approval Voting*, 2nd edn (New York: Springer, 2007).

[20]Plutarch, "Lycurgus," trans. Knightly Chetwood, *Lives of Illustrious Men*, vol. 1, ed. Arthur Hugh Clough (Philadelphia: John C. Winston Co., 1908), pp. 76–115, at p. 108.

[21]Robert Finlay, *Politics in Renaissance Venice* (New Brunswick: Rutgers University Press, 1980), pp. 141–2.

[22]Michel Balinski and Rida Laraki, *Majority Judgment: Measuring, Ranking, and Electing* (Cambridge, MA: MIT Press, 2011), p. 252.

[23]See Christian List, "Social choice theory," *Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta (Winter 2013), <https://plato.stanford.edu/archives/win2013/entries/social-choice/>.

is the same no matter which of the people supply which of the inputs. Anonymity entails a certain equality among voters: nobody's inputs count differently because of their wisdom, wealth, race, or gender.

Collective grading furthermore dissolves what to some has seemed a serious problem with the very idea of democracy. Kenneth Arrow's "impossibility" theorem tells us that there is no method for deriving a single "social" ranking of options from voters' individual rankings—none satisfying several conditions that Arrow took to "express the doctrines of citizens' sovereignty and rationality in a very general form."[24] One assumption of Arrow's theorem is social ordering. Anonymity secures, furthermore, Arrow's condition of *nondictatorship*, which rules out the existence of any one voter with whose strict preferences the social ordering must always agree, irrespective of other voters' preferences. All collective grading methods satisfy social ordering, and many satisfy anonymity as well. Indeed, it is not difficult to see that median grading satisfies all assumptions and conditions of Arrow's theorem, suitably formulated for graded ballots.[25]

The grading-jury theorem in Section V applies to grading problems in which there are right grades for things and wrong ones, independently of whether anyone actually assigns them these grades. The idea of holding grade decisions to an independent standard—the idea of an *epistemology* of grading—might seem dubious. How, you wonder, could there be an independent factual matter for a panel to find out of whether a policy option is *Very Good*, say, as opposed to *Excellent* or merely *Acceptable*? To what independent standard could we hold grades for options in referendums, or for candidates in political elections? The next section prepares the ground for the grading-jury theorem by introducing several suitably decision-independent conceptions of truth in grading, relevant in different kinds of grading problems.[26]

III. TRUTH IN GRADING

Consider first a simple kind of case. Sometimes grades have definitions that pin down their meanings in completely precise and unambiguous terms. Take, for instance, the language of likelihood grades that the Intergovernmental Panel on Climate Change (IPCC) stipulates for use in its publications. The top grade is

[24]Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd edn (New York: Wiley, 1963), p. 31.

[25]Several authors have noted that scoring and grading enable an "escape" from Arrow's theorem; see Claude Hillinger, "The case for utilitarian voting," *Homo Oeconomicus*, 22 (2005), 295–321; Balinski and Laraki, *Majority Judgment*; and Michael Morreau, "Grading in groups," *Economics and Philosophy*, 32 (2016), 323–52.

[26]Epistemic accounts of democracy point to the capacity of democratic decisions to "track the truth." Some seem to think that unchanging and absolute truths are the ones to track; Nadia Urbinati dismisses epistemic democracy as "democratic Platonism" in *Democracy Disfigured: Opinion, Truth, and the People* (Cambridge, MA: Harvard University Press, 2014). Perhaps there are "higher" truths, and democracies more than other forms of government tend to make decisions that are informed by them, and this adds to their prestige and legitimacy. Perhaps not. Be this as it may, good government certainly requires adequate responses to such mundane facts as magnitudes of risks, merits of policies, and people's wants and needs. This discussion concentrates on right and wrong judgment in these down-to-earth matters.

Virtually Certain; next comes *Very Likely*, and so on down to *Exceptionally Unlikely*. To avoid misunderstandings, the IPCC specifies precise intervals of probability for the different expressions: an event is *Virtually Certain* if its likelihood is at least 0.99; *Very Likely* covers events whose likelihood is at least 0.90, and similarly for the other five grades.[27] Here there is no problem about truth in grading. An IPCC probability grade is right for any event that it covers, and it is wrong for all other events.

In general, grades do not have precise definitions. Often, they have imprecise ones. The grading language used by AHRC panels for peer review is typical. There are six scores, from *6* at the top down to *1*. The AHRC publishes interpretational guidance, defining the scores in natural language. The score *4*, for instance, indicates "Work that demonstrates high international standards of scholarship, originality, quality and significance. Will advance the field of research."[28]

This definition ties down the meaning of the score *4* to some extent: the evaluation criteria are scholarship, originality, quality, and significance. It does not pin it down, though. For some project proposals, it is going to be indeterminate whether standards of scholarship, originality, and the rest are "high," and indeterminate whether the requirements for a *4* are satisfied. Since panels have some authority to set thresholds and decide in these borderline cases, you can wonder how their grade judgments could be, independently of their decisions, right or wrong.

In cases such as peer review, where expert judgment is expected, we may take the judgments of ideal experts as an independent standard. Imagine an elite panel. Its members are experts in the scientific field of the proposals, and in peer review. They are conscientious, and cooperative, and have all they need to do as good a job as can possibly be done. The judgments of this elite panel are a "gold standard" to which the grade decisions of ordinary panels can be held. We can count scores for project proposals as right if they are the scores that the elite panel would award.

The idea of an independent standard of truth for political decisions strikes many people as particularly farfetched. One reason for this, perhaps, is that in a referendum, or a political election, voters do not merely consider their options in light of some fixed criteria, settled in advance. In public deliberation during the run up and privately, when weighing their vote decisions, people also decide which issues are the important ones for them, and what is at stake. Independently of this deliberation and these decisions, perhaps no outcome is uniquely "right."

[27]Michael D. Mastrandrea, Katharine J. Mach, Gian-Kasper Plattner, et al., "The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups," *Climatic Change*, 108 (2011), 675–91.

[28]This definition came from the AHRC's *Research Funding Guide*, <ahrc.ukri.org/documents/guides/research-funding-guide1/>. It is quite typical. Many scores and grades are defined in vague natural language.

Balinski and Laraki recognized the authority of voters to decide what is at stake in their experiment with grading in the French presidential election of 2007. The grades they provided for the voters to use were ordinary adjectival expressions, including *Très Bien*, *Passable*, and *à Rejeter*. There were no strict definitions, like those of the IPCC. There were no authoritative guidelines specifying relevant evaluation criteria, like those of the AHRC. Balinski and Laraki allowed the different grades to carry their everyday meanings in French, simply asking voters to grade the candidates "having taken every consideration into account."[29] They left it to the voters to decide for themselves just which considerations these would be.

A further complication is that different voters may have different considerations. As James Urmson remarks of the criteria for being a good apple, or a bad one, "No one can give the precise list; some will omit a criterion I have given, add another, vary the emphasis and none of them need be wrong (though we could produce a list which would be certainly wrong)."[30] This scope for faultless difference makes the notion of a *single* standard of truth genuinely perplexing. I say some proposed law or policy option is "good." You, omitting or adding criteria, say it is "very bad"; and someone else, varying the emphasis, offers "neither here nor there." Somehow, we all disagree; and it is not merely about words; and yet, there being no factual matter of the precise list of criteria, none of us is wrong (though one could have an opinion which is certainly wrong). How under such circumstances could there be a sane, down-to-earth sense in which one option is, simply, right, and another wrong—not just for me, and not just for you or someone else, but for everyone?

Here is one decision-independent standard of truth for political decisions. It allows voters to decide what is at stake, and to disagree among themselves and with the outcome of a vote without anyone being wrong.

Consider some option in a political decision. It is a policy option or a candidate in an election; or else it is a grade for evaluating one of these. Consider also some voter. This option is *individually right* for this voter if it is the one that this voter would settle on under conditions of full and correct information.[31] Choosing it is an authentic expression of the voter's own criteria and emphasis, undistorted by any lack of information or misinformation.[32] Now generalize from individual voters to the assemblies they make up. An option is *collectively right* for an assembly if it is the one that this assembly would settle on, using its characteristic decision method, were all its members to choose the options that are right for themselves.

[29]Balinski and Laraki, *Majority Judgment*, p. 252.

[30]James O. Urmson, "On grading," *Mind*, 59 (1950), 145–69, at p. 160.

[31]Compare the notion of a correct vote decision in Richard R. Lau and David P. Redlawsk, "Voting correctly," *American Political Science Review*, 91 (1997), 585–98, at p. 586.

[32]There is a wrinkle: someone with full information might have no interest in gathering more information; yet we, without this advantage, are right to have a high opinion of, say, freedom of information laws. There is a way to iron it out: our individually right options are the ones that fully informed versions of us would recommend—not for themselves, but for us.

Individual and collective rightness are in the relevant sense decision-independent. What an individual or group would decide under ideal conditions is what it is, whether or not anyone reaches these same decisions on ordinary occasions for choice. Collective rightness is not a standard for judging *individual* vote decisions. People can dissent from the outcome of a vote while being true to themselves—while voting for what is right for them. It is, though, a standard for judging *collective* decisions. This is recognized in practice. In 2019, the Swiss Federal Supreme Court overturned the result of a 2016 referendum on taxation of married couples. In the run up, the Swiss government had announced that proposed changes to the law would affect some 80,000 couples; but later, after the vote, the number was revised upwards to almost half a million. One ground for voiding the result was that it might have been different had voters been provided with complete information. The result might, in other words, not have been, collectively, right.[33]

The ideal of full information is all very well in theory. In practice, meanwhile, citizens often vote on matters about which they know only little. They are exposed to misinformation, some of it deliberately misleading, and they are swayed by chance events that should not affect their votes but do.[34] Happily, there are non-ideal conditions under which assemblies reliably make right decisions even so.

Condorcet's jury theorem tells us that majorities reliably choose the right option from a pair, provided the individual chance of truth stays above the critical ½. Section VI shows that, going instead by median grades, under certain circumstances the collective choice reliably is right, even though the individual chance of truth drops below ½, and voters on their own are most likely to choose the wrong option: they'd be more likely to get things right just by guessing! This might sound like a sort of alchemy, the transmutation of error into truth, but there is no magic to it. Before going into technical details in Sections V and VI, let us first see intuitively how the epistemic mechanism of median grading works.

IV. BRACKETING THE TRUTH

Errors on either side of a correct value or setting can be instructive. A photographer finds the right exposure for an image by under- and over-exposing in a series of approximations. The helmsman of a ship and the pilot of an airplane identify the correct heading for an intended course by erring first on one side and then on the other. The errors indicate what is right or correct by surrounding or *bracketing* it.[35]

[33]The referendum on the popular initiative "For the couple and the family—no to the penalty of marriage" was on 28 February 2016. The Federal Supreme Court's decision invalidating it, widely reported at the time in the press, came on 10 April 2019.

[34]Such as wins and losses in sporting events; Andrew J. Healy, Neil Malhotra, and Cecilia Hyunjung Mo, "Irrelevant events affect voters' evaluations of government performance," *Proceedings of the National Academy of Sciences*, 107 (2010), 12804–9.

[35]There is in philosophy a different usage, in which to "bracket" something is to set it aside, or to hide it; in phenomenology, this term translates Husserl's "einklammern."

Many false judgments of different people in some matter can bracket the truth. This is not new. Over a century ago, Sir Francis Galton analyzed a competition at a country fair to guess the weight of an ox, once slaughtered and "dressed." He found that the median of many precise weight estimates, provided by the competitors, was much closer to the actual weight of the dressed ox than, on average, these estimates themselves. Galton already remarked on the connection to political philosophy, finding his results "more creditable to the trustworthiness of a democratic judgment than might have been expected."[36]

There is an important difference, though, between Galton's weight-guessing competition and typical decisions by panels, committees, and, especially, electorates. Among Galton's competitors there were many professionals—farmers, butchers, stock auctioneers—who were really good at sizing up livestock. Having had a look at this ox, they could estimate the dressed weight with high precision.[37] Specialists in some scientific fields, on the other hand, often struggle to make quantitative judgments that would bring their knowledge to bear on public policy.[38] For ordinary citizens, short on time, information, and relevant expertise, precise estimation of likelihoods, harms, and other matters relevant to their votes must in general be completely out of the question.

This is where grades come in. Someone might be able to say with confidence that a risk is "high," or that a policy is "very likely" to succeed, without being able to pin down the precise risk, or the precise likelihood of success. This is just because these expressions are *imprecise*. They cover ranges of precise degrees. Similarly, someone might be able to say that a political candidate is *Bien*, even though, perhaps due to indeterminacy in the criteria, or in weights or priorities, there isn't even a precise degree to which this candidate is good, or suitable for office.

Many grade judgments can bracket a correct grade, just as many precise estimates can bracket the true weight of an ox. Listing from "top" to "bottom" all the grade judgments of the different members of an assembly, we choose the median grade for it from the middle of the list. Under favorable conditions this median grade reliably is the right one, even if most of the judgments are wrong, either too low or too high. This, intuitively, is how median grading can track the truth, also under circumstances in which the individual probability of truth is low.

The grading-jury theorem tells us a certain amount about collective grading methods other than median grading. Take *Majority Judgment*, the method that Balinski and Laraki propose for judging competitions and picking winners in

[36]Francis Galton, "Vox populi," *Nature*, 75 (1907), 450–1, at p. 451.

[37]Galton reports in whole pounds estimates variously recorded by the competitors in pounds, hundredweight, or quarters.

[38]M. Granger Morgan gives the example of a toxicologist who, despite thinking that the questioning made sense, simply could not bring himself to make quantitative judgments on subjective probabilities relating to the slope of a health-damage function; M. Granger Morgan, "Use (and abuse) of expert elicitation in support of decision making for public policy," *Proceedings of the National Academy of the Sciences*, 111 (2014), 7176–84, at p. 7177.

political elections.[39] It is median grading with an innovative method for breaking ties. The grading-jury theorem applies directly to all cases in which there are no ties.

There is indirect relevance for "range" voting, where numerical scores for each candidate in an election are tallied up and the candidate with the highest sum of scores wins. Arguably, these sums often are literally meaningless. This is because the scores often have a merely ordinal significance. The operation of addition, then, has no empirical interpretation in their case and, by this measurement-theoretic argument, there is no special reason to suppose that the winner of a range vote has *more* of something—more goodness, fitness for office, or what have you—than any loser. There is a way in which the results of range voting can be informative even so. If it so happens that sums of ordinal scores rise and fall together with the medians, then the grading-jury theorem tells us of conditions in which they track, in turn, the truth.[40] Having higher sums of scores can "mean" that candidates are better by correlation, in the same non-semantic way that clouds "mean" rain, and smoke "means" fire. Just as with clouds and rain though, or smoke and fire, correlations that make range voting informative are empirical, and must be established in each case separately.

V. THE GRADING-JURY THEOREM

Consider the following simple type of evaluation task. An assembly with n members grades some given item x in some given language of grades. One grade in this language (and only one) is the right grade for x, in some suitably decision-independent sense. Our question is, under which conditions is the median of n grades for x, assigned by the individual members of the assembly, likely to be this right grade?[41] Let the collective probability of truth $\widetilde{P}_n$ be the probability that the median of n grades for x is the right grade. We would like to know of any conditions under which $\widetilde{P}_n$ is high.[42]

Let an *individual probability of truth*, p, be the probability that some given individual member of the assembly assigns the right grade to x. Assume, for now, that every member has the same individual probability of truth. Suppose also that the individual probability o of assigning either the right grade or a lower one is the same for every member. Suppose furthermore that these events are probabilistically independent.[43] Suppose similarly that the individual probability q of assigning either the right grade or a higher one is the same for every member, and that these events too are probabilistically independent.[44] We have the

[39]Michel Balinski and Rida Laraki, "A theory of measuring, electing, and ranking," *Proceedings of the National Academy of Sciences*, 104 (2007), 8720–5; Balinski and Laraki, *Majority Judgment*.

[40]With symmetrical distributions means and medians coincide, and the correlation is as close as can be.

[41]Assume n is an odd number, so that the median is defined.

[42]Where there are several items to grade, not just one, there are several tasks of this simple sort.

[43]That is, the chance of any given member's awarding a grade that is either right or too low is the same, whether or not any other given member does.

[44]That these probabilities are the same for everyone is not realistic. Section VII considers *heterogeneous* assemblies, whose members differ in their chances of truth and error.

Grading-jury theorem. If $|o - q| < p < 1$, then

(1) The larger the number n of members, the higher is $\widetilde{P}_n$, and

(2) $\widetilde{P}_n$ can be brought arbitrarily close to 1 by increasing n.[45]

There is a proof in the appendix. It establishes that the condition $|o - q| < p < 1$ is equivalent to: $o > 0.5$, $q > 0.5$, and $o < 1$ or $q < 1$. This, on either formulation, is the *competence condition* of the grading-jury theorem.

Here, intuitively, is how the competence condition makes it likely that median grades are right. One provision is that individual voters are more likely than not to assign either the right grade or else a higher one ($q > 0.5$). When a large electorate votes, we may therefore expect that less than half the voters assign too low a grade. Similarly ($o > 0.5$), we may expect that less than half come in too high. Lining up everybody's judgments in the top–bottom order of the grades, there is one in the middle, with the same number of judgments on either side. This median judgment should not be too low (in that case, more than half the voters assign too low a grade). Similarly, it should not be too high. The competence condition leads us to expect that the median grade is right.

The grading-jury theorem tells us, intuitively, this. The members of an assembly might be ignorant (tending to decrease p, the individual chance of assigning the right grade). They might be biased (tending to increase the imbalance $|o - q|$ between the chances of error on either side). Suppose, though, that the members are not too unenlightened in both ways at once ($|o - q|$ remains lower than p). Then (1) the more numerous the assembly, the more likely it is collectively to assign the right grade, and (2) the scope for improvement is boundless: a sufficiently numerous assembly is virtually certain to assign the right grade.

The provision $|o - q| < p$ entails that $p > 0$, individuals might assign the right grade. The provision $p < 1$ means that also they might not. For any individual chance of truth between these extremes, the epistemic engine of median grading *can* kick in. Whether it does depends on the balance between the chances o and q of errors up and down.[46] If p is extremely low (close to 0), then errors have to be precisely balanced ($|o - q|$ is at least as close to 0). Where p is greater, the balance can be rough, with o and q further apart.[47] With a good individual chance of getting the right grade, on the other hand ($p > 0.5$), errors up and down can be as unbalanced as you like. Then the condition $|o - q| < p$ is satisfied no matter what.[48]

[45]That is, $\lim_{n\to\infty} \widetilde{P}_n = 1$.

[46]Competence in median grading is in this sense multi-dimensional, not a simple question of the individual probability of truth (as it is with pairwise majority voting).

[47]For instance, let $p = 0.25$, while the individual chance of awarding strictly too low a grade is 0.30 and the chance of too high a grade is the remaining 0.45. Then the difference between o and q is 0.15 and the competence condition is met: $o = 0.25 + 0.30 = 0.55$, $q = 0.25 + 0.45 = 0.70$, and $|o - q| = |0.55 - 0.70| = 0.15 < 0.25 = p < 1$.

[48]By definition of o, p and q, we have $(o - p) + p + (q - p) = 1$. If $p > 0.5$, then both $0 \leq o - p < 0.5$ and $0 \leq q - p < 0.5$. So $|o - q| = |(o - p) - (q - p)| < 0.5 < p$.

Errors might often be sufficiently well balanced. Suppose individual grade decisions are made on the basis of "generated signals," the "noisy glimpses or distortions of an outcome value."[49] People award the grades they think cover the signals they receive.[50] Where conditions are favorable, there is little noise in signals, and mistakes in identifying corresponding grades are unlikely. Then the individual chance of truth is high ($p > 0.5$), ensuring that $|o - q| < p$. Imagine now that conditions gradually worsen. Signals become noisier and more biased, and mistakes in choosing corresponding grades become more common, putting downward pressure on p. Think of probability as flowing away from the right grade. Some of it flows, as conditions worsen, downwards to lower grades (tending to increase o), and some flows upwards, to grades that are too high (increasing q). Provided these notional outflows of probability are not at the same time too great and too unbalanced, satisfaction of the competence condition is preserved—even to the point that the individual chance of truth approaches 0.

Extreme grades, at the very top or the bottom of the scale, are a special case. When the right grade is extreme, the competence condition is satisfied only if the individual probability of truth is high ($p > 0.5$).[51] The median grade of an assembly is the right one only if most members assign this grade, and the epistemic mechanism of median grading has no effect beyond that of majority voting about grades. There are consequences for the choice of grade languages, and for the design of peer-review panels, selection committees, and other decision-making bodies. First, it might often be better to choose a language whose extreme grades are needed only rarely; second, the individual probability of truth p might be increased by finding easily verified criteria for extreme grades, and by encouraging people to take extra care when considering whether to award these. This applies in particular to approval voting, and to pass–fail grading in schools and universities: with just two available grades, both are extreme.

The grading-jury theorem stated here cannot apply to many real grading problems. For one thing, its assumption that everyone has the same chance of truth, and the same chances of errors up and down, is quite unrealistic—just like the corresponding homogeneity assumption of the standard Condorcet jury theorem, of which this grading-jury theorem is a corollary. Versions that are more widely applicable are available, though, as corollaries of many variants of Condorcet's theorem, obtained over the years in order to increase realism. Section VII takes a first step in this direction. It discusses a version that covers

[49]Lu Hong and Scott Page, "Interpreted and generated signals," *Journal of Economic Theory*, 144 (2009), 2174–96, at p. 2177.

[50]For instance, someone might judge that an event is *Very Likely*, in the sense of the IPCC, having estimated its likelihood as about 0.96, and having read in the relevant IPCC guidance note that *Very Likely* covers events whose likelihood is at least 0.90; see Mastrandrea et al., "The IPCC AR5 guidance note on consistent treatment of uncertainties."

[51]Suppose the right grade is at the bottom of the scale (the reasoning is similar if it is at the top). Then the chance $o - p$ of awarding strictly too low a grade is 0, and the chance $q - p$ of awarding strictly too high a grade is $1 - p$; we have $|o - q| = |(o - p) - (q - p)| = (1 - p) < p$ only if $p > 0.5$.

heterogeneous assemblies, whose members differ in their chances of truth and error.[52]

VI. MEDIAN GRADING IS FORGIVING

Sometimes, we need to know the right grades for things. For instance, in peer review, it might matter whether the merit of a project proposal is below some given threshold of acceptability, expressed as a grade, or above it. Decisions about which grades to assign are *categorical* decisions. Sometimes, on the other hand, we only need to know the right order of things. For instance, to choose the best among some candidates for a position, we only need to know which of them are better than which. Decisions about which order to put things in are *comparative* decisions. Median grading is a forgiving method for making both kinds of decisions.

A. Categorical Decisions

Median grading is a forgiving method for making grade decisions. Suppose an assembly is grading an item in a language with some finite number g of grades. The critical probability—the chance of guessing which grade is the right one—is $1/g$. It is required that there are conditions under which the median grade of a sufficiently large assembly is likely to be the right grade for an item, even though any individual member's chance of assigning the right grade is less than $1/g$. This is a direct consequence of the grading-jury theorem.

Computer simulations give a picture of how the collective chance of truth increases as assemblies get larger. Suppose the grading language has six grades, so that the critical probability is ⅙, or 0.166… . Let the individual chance p of truth be less than this, say 0.15. Let the individual chance of awarding a grade that is strictly too low be, say, 0.45, which is greater than the chance of awarding too high a grade, the remaining 0.40 of probability (there is a certain bias against the item). Then the competence condition is satisfied: with $o = 0.45 + 0.15 = 0.60$, and $q = 0.40 + 0.15 = 0.55$, we have $|o - q| = 0.05 < 0.15 = p$.

Under these circumstances, the chance that a simulated assembly of five assigns the right grade is about 0.27.[53] This is already well above the critical ⅙. With

[52]The assumption that individual grade decisions are independent is also questionable. There are pointers towards more realistic independence conditions in Franz Dietrich and Christian List, "A model of jury decisions where all jurors have the same evidence," *Synthese*, 142 (2004), 175–202; and Franz Dietrich and Kai Spiekermann, "Epistemic democracy with defensible premises," *Economics and Philosophy*, 29 (2013), 87–120. There is further discussion of the Condorcet theorem and its variants in Goodin and Spiekermann, *An Epistemic Theory of Democracy*, pt 1, and references therein.

[53]All $\widetilde{P}_n$ are obtained using Lemma 1, in the technical appendix. Probability estimates come from a Monte Carlo simulation of majority voting written by Christopher J. Thompson. The number of trials in each case is 50,000, and probability estimates are rounded to the closest whole percentage point. With $o = 0.60$ and $q = 0.55$, we have $O_5 \approx 0.68$ and $Q_5 \approx 0.59$. By Lemma 1, $\widetilde{P}_5 = O_5 + Q_5 - 1 \approx 0.27$.

25 members, the assembly's grade is more likely to be right than wrong; the chance that it is right is about 0.55. The median grade of an assembly of 501, finally, is almost certainly right; $\widetilde{P}_{501}$ is about 0.99.

Expert panels and committees are often quite small. The simulation results underline the need for members to be fairly competent. Provided they are, the probability of truth can build quite powerfully, even with considerable bias in errors. Suppose, for instance, that the individual chance of erring on one side is 0.30, but on the other side it's only 0.10. With a panel of three, the collective probability of truth is then 0.75—a full 25 percent improvement on the individual probability of truth, which is 0.60. With five members, it rises to 0.83.

B. Comparative Decisions

Median grading is a forgiving method for ranking options and choosing among them. The critical probability for binary choice is ½. An example illustrates conditions under which a sufficiently large assembly is most likely to arrive at a higher median grade for a superior option, and to choose this one, even though the individual chance of choosing it is less than ½. What enables a large assembly reliably to judge and to choose, in the example, is that the grading language discriminates between the options (the right grade for the superior option is higher than for the inferior one), while the conditions of the grading-jury theorem are satisfied for each one.

To make things concrete, let everybody score two options x and y on a scale from *6* down to *1*; the right score for the superior x is *4*, let us say, and the right score for the inferior y is *3*. The individual probabilities of truth are as follows. Where p^x is the common individual chance of correctly awarding x a *4*, and p^y is the common individual chance of correctly awarding y a *3*, $p^x = p^y = 0.25$. Let there be a good individual chance, 0.30, of scoring x too leniently, by awarding a *5*. Even so, individual voters are strongly prejudiced against x, with a 0.45 chance of awarding x the bottom score, *1*. Thus $o^x = 0.70$ and $q^x = 0.55$. Individuals are just as strongly biased in favor of y, the inferior option, with a 0.45 chance of awarding y the top score *6*, and a 0.30 chance of awarding y a *2*: $o^y = 0.55$ and $q^y = 0.70$. These individual probabilities of assigning the different grades to the options are summarized in Table 1. Notice that individual grade decisions for x and y are assumed to be independent.

By inspection of Table 1, the only way for a member to give x a higher grade than y is by awarding x either a *5* or a *4*, while awarding y either a *3* or a *2*. The chance of each of these events is 0.55 (= 0.25 + 0.30), so the individual chance of giving x a higher grade than y is 0.55×0.55, or just over 0.30.[54] (Alternatively, this value can be calculated by adding the values in the four relevant cells in the top right of the table.) Thus the individual chance of judging that x is better than

[54]With dependencies between the grade decisions concerning x and y, the chance could be higher or lower. I thank Paul Grünke for pointing out the interest in varying the independence assumption.

Table 1. The joint probability distribution for each voter's grade assignments across the options x and y

	y					
x	***6***	***5***	***4***	***3***	***2***	***1***
6	0	0	0	0	0	0
5	0.30 × 0.45	0	0	0.30 × 0.25	0.30 × 0.30	0
4	0.25 × 0.45	0	0	0.25 × 0.25	0.25 × 0.30	0
3	0	0	0	0	0	0
2	0	0	0	0	0	0
1	0.45 × 0.45	0	0	0.45 × 0.25	0.45 × 0.30	0

y is far below the critical ½. Furthermore, by inspection, the individual chance of giving *x* the same grade as *y* is 0. It follows that there's no chance of choosing *x* instead of *y* in a tie: the chance of choosing *x* is just the chance of judging that *x* is better than *y*, and likewise far below ½.

Even so, the competence condition of the grading-jury theorem is satisfied for *x* and also for *y*: we have $|o^x - q^x| = |0.70 - 0.55| = 0.15 < 0.25 = p^x < 1$, and similarly with o^y; p^y and q^y. It follows that, with a sufficiently large assembly, the median score for *x* is likely to be the right score for *x*, namely *4*, while the median score for *y* likewise is *3*, which is lower. Thus well informed about the relative merit of the options, the assembly is likely to make the right choice among them.

Computer simulations again give a picture of how the chance of making the right choice can build as the size of the assembly increases. Under the described circumstances, the chance that *x* receives a higher score from an assembly with 15 members is, at just over 0.42, still below even odds.[55] There are not enough members for the epistemic engine of median grading to overpower everybody's misleading biases. With a group of 45, though, the chance of making the right decision is already up to about 0.57. With 501 members it is above 0.97. An assembly this size is almost certain to give the better option a higher score, and to make the right choice.

VII. HETEROGENEOUS ASSEMBLIES

The grading-jury theorem of Section V and computational examples of Section VI concern homogeneous assemblies. All voters have the same chance of awarding right grades, the same chance of awarding grades on the low side, and the same

[55]The reasoning is as follows (and similar for other estimates of this example). The chance that an assembly of 15 arrives at a collective *5* for *x* is, reckoning using the simulation as in the previous example, about 0.05, and the chance that it awards the correct *4* is about 0.60, so the chance of awarding *x* one of these scores is about 0.65. The chance of collectively awarding *y* either a *3* or a *2* is the same. Since the only way of correctly judging *x* to be better than *y* is to award *x* either a *5* or a *4*, while awarding *y* either a *3* or a *2*, the chance of this is, assuming independence of these events, about $0.4225 = 0.65 \times 0.65$.

chance of awarding grades on the high side. In reality, that cannot often be the case; but the grading-jury theorem can be generalized to cover assemblies that are in this respect diverse. Let o_i be the chance that individual i awards the right grade or a lower one to x. Similarly, p_i is the chance that i awards to x the right grade, and q_i is the chance that i awards to x the right grade or higher. Let $\overline{o}$ be the average of all the o_i, and similarly for $\overline{p}$ and $\overline{q}$. Provided $|\overline{o} - \overline{q}| < \overline{p}$, we can bring $\widetilde{P}_n$ arbitrarily close to 1 by increasing n.[56]

The condition $|\overline{o} - \overline{q}| < \overline{p}$ requires a certain balance between voters who tend to award grades that are lower than the right grade, and those who tend to award higher grades than that. It is possible to achieve this balance by adding or removing individual voters. Suppose an assembly has many voters who tend to award grades that are on the high side (then $\overline{q}$ is high). Satisfaction of this condition might be secured by adding voters whose grades tend to be on the low side (raising $\overline{o}$). In this way, increasing its polarization can actually put an assembly back on the track of the truth. Competence, in such a case, is genuinely social. It is a matter of the composition of the diverse group.

People often have different understandings of scores and grades, even people who are culturally and educationally similar, such as students and members of science panels.[57] They persist in this even when given clear interpretational guidelines.[58] Now, suppose some of us are grading a political candidate. This candidate gets *Acceptable* from you, *Good* from me and *Very Good* from the other person. What sense is there in taking the median and awarding the candidate a collective *Good* if, for all we know, my standards are lower than yours, and this *Good* from me really just amounts to *Acceptable* from you? Should we then normalize to your standards and call the candidate *Acceptable*? Mine, and award a *Good*? Normalize to someone else's standards, and award some other grade? Diversity in people's understandings of grading expressions might seem to make nonsense of putting grades together in this way.

It does not. Whatever the causes of variation in grade judgments—whether ignorance, bias, prejudice, inattention, misinterpretation of grades, normal variation in understandings of natural language grade definitions, or whatever else—sufficiently large assemblies track the truth whenever they meet the conditions of the grading-jury theorem.

[56]Again, there is a standard independence assumption. The grading-jury theorem for heterogeneous assemblies is demonstrated by using, in the proof of clause (2) of the homogeneous grading-jury theorem, not the standard Condorcet jury theorem, but instead Theorem 2 of Guillermo Owen, Bernard Grofman, and Scott L. Feld, "Proving a distribution-free generalization of the Condorcet jury theorem," *Mathematical Social Sciences*, 17 (1989), 1–16.

[57]See Thomas S. Wallsten, David V. Budescu, Amnon Rapoport, et al., "Measuring the vague meanings of probability terms," *Journal of Experimental Psychology: General*, 115 (1986), 348–65; and Morgan, "Use (and abuse) of expert elicitation in support of decision making for public policy."

[58]David V. Budescu, Stephen B. Broomell, and Han-Hui Por, "Improving communication of uncertainty in the reports of the intergovernmental panel on climate change," *Psychological Science*, 20 (2009), 299–308; and David V. Budescu, Han-Hui Por, Stephen B. Broomell, and Michael Smithson, "The interpretation of IPCC probabilistic statements around the world," *Nature Climate Change*, 4 (2014), 508–12.

VIII. CONCLUSION

The grading-jury theorem states a condition on individual grading competence under which, given standard independence assumptions, it is likely that median grades arrived at by large groups of graders are independently right. One consequence is that median grading is forgiving. Even with graders so unenlightened that they are most likely to make wrong decisions on their own, median grading under certain circumstances enables sufficiently numerous assemblies reliably to make the right collective judgments and decisions.

The grading-jury theorem suggests a solution to problems of voter ignorance in democracies. It is to use voting methods that make better use of people's limited knowledge than do traditional methods such as majority voting. Multi-agent computer simulations suggest that assemblies comparable in size to familiar deliberative assemblies can benefit from the forgiving nature of median grading.

Theorems and models are one thing, and practical implications are another. It is not known to what extent median grading or other novel voting methods really could solve problems of voter ignorance. This depends on the empirical matter of how often the conditions actually obtain under which real juries, committees, and electorates would benefit from using them.

APPENDIX

This appendix contains proofs of technical claims made in the preceding sections.

A. Medians and Majorities

We consider a grading problem in which each member of an assembly of size n assigns to some given item a grade from a fixed language of grades. One of these grades, T, is designated the target. Where an outcome includes one grade judgment from each of the n graders, and Ω is the set of all outcomes, let P be a probability function on Ω. With n an odd number, $\widetilde{P}_n$ denotes the probability that the median of n grade judgments, one from each member, is T; O_n is the probability that most are either T or some grade that is strictly lower than T; and Q_n is the probability that most are either T or strictly higher. We have

Lemma 1. $\widetilde{P}_n = O_n + Q_n - 1$.

Demonstration: let $X \subseteq \Omega$ be the outcomes in which the median of the grade judgments is T. Let Φ be the outcomes in which the median is T or lower, and let Ψ be the outcomes in which the median is T or higher. Then $\Omega\backslash X = \Omega\backslash\Phi \cup \Omega\backslash\Psi$. Since $\Omega\backslash\Phi \cap \Omega\backslash\Psi = \varnothing$, $P(\Omega\backslash X) = P(\Omega\backslash\Phi) + P(\Omega\backslash\Psi)$, and $1 - P(X) = 1 - P(\Phi) + 1 - P(\Psi)$. Since, finally, $\widetilde{P}_n = P(X)$, $O_n = P(\Phi)$ and $Q_n = P(\Psi)$, substituting and rearranging, $\widetilde{P}_n = O_n + Q_n - 1$. □

B. The Grading-Jury Theorem

The jury theorem for median grading is equivalent to Condorcet's jury theorem for majority voting, here stated first. Suppose there is a choice between two options, one of which is independently right. Let p be the probability that a voter chooses the right option, and

let P_n be the probability that most of some odd number n of choices are right. Assuming votes are independent, we have (for odd m and n) the

CONDORCET JURY THEOREM. If $0.5 < p < 1$, then

(1) $P_n > P_m$ if $n > m$, and
(2) $\lim_{n\to\infty} P_n = 1$.[59]

We turn now to a proof of the grading-jury theorem. It is assumed that one of some given available grades is the right grade for some given item that is under consideration; $\widetilde{P}_n$ is defined as in LEMMA 1, with this right grade as the target grade T (O_n and Q_n, in the proof, are defined similarly). Let the individual chance o of assigning the right grade or a lower one be the same for each grader, and let these events be probabilistically independent. Similarly, let the individual chance q of assigning the right grade or higher be the same, and let these events be independent. Let p be the individual chance of assigning the right grade. For assemblies whose size is some odd number m or n we have the

GRADING-JURY THEOREM. If $|o - q| < p < 1$, then

(1) $\widetilde{P}_n > \widetilde{P}_m$ if $n > m$, and
(2) $\lim_{n\to\infty} \widetilde{P}_n = 1$.

Demonstration: we see first that the competence condition $|o - q| < p < 1$ is equivalent to the alternative condition (a) $o > 0.5$ and $q > 0.5$, and (b) either $o < 1$ or $q < 1$. Suppose to this end that $|o - q| < p < 1$. For (a), note that $o + q - p = 1$, by definition of o, q, and p, and so $p = o + q - 1$. Wlg. $o \geq q$. Since $|o - q| < p$, we have $o - q < p$, and so $o - q < o + q - 1$. Rearranging, $2q > 1$ and $q > 0.5$. Since $o \geq q$, also $o > 0.5$. For (b), suppose $o = 1$. Since $o + q - p = 1$, $q = p$. That $q < 1$ now follows from $p < 1$. Suppose, on the other hand, (a) and (b). That $p < 1$ follows immediately from (b), since both $p \leq o$ and $p \leq q$. Wlg. $o \geq q$. Since $q > 0.5$, we have $0 < 2q - 1$ and (adding $o - q$ to each side) $o - q < o + q - 1$. Since $p = o + q - 1$, finally, $|o - q| < p$. The two competence conditions are equivalent.

Assume now that $o > 0.5$ and $q > 0.5$, while either $o < 1$ or $q < 1$. Wlg. $o < 1$. By the Condorcet jury theorem,

(3) (a) $O_n > O_m$ if $n > m$,
(4) (a) $\lim_{n\to\infty} O_n = 1$,

and similarly,

(3) (b) $Q_n \geq Q_m$ if $n > m$, and
(4) (b) $\lim_{n\to\infty} Q_n = 1$.

Now (3)(a) and (3)(b) together with LEMMA 1 entail clause (1) of the grading-jury theorem, while (4)(a) and (4)(b) together with LEMMA 1 entail (2). □

Conversely, the grading-jury theorem entails the Condorcet theorem: they are equivalent. Suppose there are two options to choose between, one of them right, with the individual probability p of choosing this right option such that $0.5 < p < 1$. Put the two options (think of them now as grades) in some (strict) order. The competence condition of the grading-jury theorem is satisfied (see note 48). Assuming voters' choices among the

[59]This is Theorem 1 of Grofman et al., "Thirteen theorems in search of the truth."

options are independent, the independence condition of the grading-jury theorem is also satisfied. Since, furthermore, the median (relative to the imposed order) is with just two options the one chosen by a majority, (1) and (2) reduce to the corresponding clauses of the Condorcet jury theorem. □