Election Simulation Sheds New Light On Voting Methods

By Roy A. Minet (Rev. 11/23/19)

[Abstract: In the 250-year-long debate about voting methods, a new election simulation project which quantifies the performance of various voting methods may shed considerable light on what is important and what is not. Indications are that no method limited to an ordinal ranking of the options can achieve more than "mediocre" performance because the information gathered simply cannot support consistently making the best decisions under all circumstances. Even some "rudimentary" cardinal methods perform better by allowing voters to provide more information. The results strongly suggest that a simple, easy-to-understand and easy-to-implement cardinal voting method would greatly improve the accuracy and consistency of decisions made using the election process, and that future work might most fruitfully focus on possible ways to increase immunity to strategic voting for this and a closely related method.**]**

Background and Definitions

The mechanism of a public "election" is normally used to choose those who will wield awesome and potentially dangerous government force over others. It is obviously critical that the best possible decisions be made. In general, elections may not be the best way to make good decisions, but they are nevertheless used in order to keep decision-making power dispersed and prevent it from falling into the hands of an oligarchic group or even an individual "dictator."

So it is assumed at the outset that, **the primary design objective for an election mechanism must be for it to consistently render the best possible decisions** (with the caveat that decision-making power be kept reasonably dispersed). In addition to a clearly stated objective, an equally clear and actionable definition of "the best possible decision" is needed before the design process can be meaningfully carried out.

The heart of an election consists of just two steps. First, some kind of predefined data is obtained from each voter. Second, these data are processed in some publicly-known predefined way to yield a decision, a selection of one (or several) of multiple alternatives. A unique combination of predefined data to be collected (along with the specific way by which it is to be obtained) and a predefined way the data are to be processed is called a voting method. The result rendered by any voting method is wholly dependent upon the data that has been obtained from voters. Regardless of veracity, we have no alternative but to assume, or at least lightheartedly hope, that the voters can collectively provide data that leads to good decisions.[1]

At the time of an election (the data collection step), each voter has, by whatever means and experience, developed some form of opinion about the candidates for a given race. For a given voter, a win by one specific candidate might result in that voter being very happy and satisfied. A win by a different candidate might cause that voter to be very unhappy and very dissatisfied. In fact, that same voter's reaction to a win by some particular candidate might result in great satisfaction, great dissatisfaction or anything in between. Also, it is virtually certain that, for substantially all voters, there are some candidates that the voter doesn't know enough about or care enough about to have any opinion at all.

For this physicist/engineer/investigator/concerned citizen, the only definition that makes sense is: **the best possible decision is that result which maximizes voter satisfaction, net of dissatisfaction, when summed over all voters who voted.**

Consider, for example, the simple case of a binary decision, a referendum. Suppose that 51% of voters vote "No" and 49% vote "Yes" on some referendum. By traditional majority rule, the measure is defeated.

However, further suppose that the 51% who voted "No" don't have a very strong opinion regarding this particular issue; they don't really care, but it was very easy to mark the "No" box. On the other hand, the 49% who voted "Yes" believe it to be a super-good thing and favor it passionately. In this case, the decision to defeat the referendum would not be the best choice by the previously stated definition. The group of voters would, overall, be much happier and more satisfied if the referendum were to pass instead.

So, the first consequence of adopting the stated definition that must be squarely faced is abandonment of the "majoritarian rule." This may be a difficult hurdle for many because they have been so strongly conditioned for years to unquestioningly accept that the majority always rules.

A Perfect Voting Method

Suppose that a machine, call it a "satometer," is available which is trusted by voters and which can accurately measure each voter's satisfaction regarding the issue at hand on some absolute scale of "sats" (dissatisfaction would be negative "sats"). If the net total sats for all voters

[1] We focus only on the voting method here, but there are other aspects of elections which need to be optimized in the larger picture. For example, it would be logical to consider at this point selecting a (not too small) subset of voters who could be expected to provide better data leading to better decisions. Of course, this is not a new idea and was attempted in the past through the use of voter qualification tests. Sadly, such tests were blatantly abused in some areas where they were deliberately designed to discriminate against minorities, etc. Consequently, such tests were effectively outlawed by the Voting Rights Act of 1965. Still, it might be possible to have such a test that would "pass muster" and be unassailably unbiased. In fact, such a test has elsewhere been proposed.

voting on a referendum is positive, the result will be “yes” and if total sats is negative, the result will be “no” (a zero-sat tie would be resolved by falling back to majority rule, subtracting the number of voters who registered negative sats from those who registered positive sats). This hypothetical referendum illustrates the ideal best possible decision that we aspire to most closely approach in real-world voting. It is based upon complete, accurate, comparable and sincere cardinal information measuring the satisfaction of each voter on the issue. Such accurate and complete information enables correct decisions to be made every time.

Considering now the general case where voters are to select n of m options. We set up our voting booth satometer to grab a quick reading on each of the m options from each voter. In essence, we conduct a separate referendum on each option. This is indeed a complete and very large data set. The winners will be the n options with the highest net sat totals (or in extremely sad cases, the positive-most or least negative totals). A bitter lesson swiftly to be learned is not to nominate candidates with large negatives.

Lacking a satometer, obtaining such information-rich and *sincere* cardinal data from voters is extremely problematic. At least one method has been proposed (True Weight Voting)[2] which hopes to obtain and utilize such data. TWV would closely approach the ideal behavior described above if a way can be found to gather sincere data from voters. The work reported here was an effort to quantitatively compare the performance (in accordance with the stated definitions) of a wide variety of voting methods that utilize data which can be obtained in the voting booth without a satometer.

Imperfect Voting Methods

All real-world voting methods will be imperfect to some degree. It should not be too surprising that, under some circumstances and conditions, they all will choose a wrong winner. Of course, it will generally be impossible to tell if or when this has happened. In order to make that determination, the original complete and sincere cardinal data would be needed, and that was left in voters’ brains in the voting booth.

Quite a few methods which forsake cardinal data and gather only ordinal information have been proposed; these have been very extensively analyzed and debated. Obtaining a complete ranking of all options from all voters is patently impossible. Even if complete rankings could in some way be forced, the data would then have a lower signal-to-noise ratio since many voters would have to submit arbitrary data for candidates about which they have insufficient or no knowledge. The real world is a messy place. The best ordinal data that can be obtained from voters is very ragged, consisting of just first choices from some voters, first and second choices from others and perhaps three or four choices from a few. Even in the best possible case, requiring voters to boil their satisfactions down to ranked choices will result in a very significant loss of information.

[2] See “Voting For Better Decisions” at http://royminet.org/voting-elections/

One method allows voters to provide so little information that it's hard to categorize. It is Plurality, the most widely used method.

Rarely does any voting method give voters a way to convey dissatisfaction. That means the data cannot support distinguishing between options the voter despises and those about which s/he has no opinion (either knows nothing about the candidate or just doesn't care one way or the other). The lack of this particular information might logically be expected to be a large impediment to making good decisions under all circumstances.

It gets worse. Without a satometer to pluck accurate and sincere data from voters' brains, the door is wide open for voters to lie, and many voters definitely do lie. Prior to recording the requested information on their ballots, they may think "strategically" and decide not to register their honest opinions. Based on whatever advance information and analysis of the election they have in their heads, they believe that an insincere vote will be more effective at achieving a result they favor (or averting an outcome they despise) than would an honest input. Mostly, insincere or strategic voting is thought to degrade the decision-making of a voting method. However, there may be cases where that is not true.

Prior Work

A sustained debate has continued for no less than the past 250 years. Hundreds of voting methods and variants have been proposed and investigated. Countless hours have been spent and a large body of papers produced (you are reading one more to be thrown onto the pile). No deep dive into the details of prior work will be done here. The point to be made is that, in spite of its obvious importance, a solid consensus on the best voting method to use still eludes us. Various factions advocate for different methods. All this quite correctly suggests that there is more complexity to the voting methods issue than just the surface simplicity that meets the eye.

Although the jury is still out on the best voting method, there is near unanimity on which is the worst voting method; it's the one most widely used: Plurality. Plurality gathers the bare minimum data (actually, less than the bare minimum) – just the first choice of all the options from each voter. In many elections, that is insufficient data to render an intelligent decision no matter how the data are processed. Whenever there are more than two options, Plurality frequently engenders an almost irresistible force on many voters to vote "strategically" for "the lesser of evils."

The 2016 US presidential election provided a dramatic example. Polling data from the Pew organization indicated that Trump and Clinton had only small cadres of voters who enthusiastically supported them, 11% for Trump and 12% for Clinton. Much larger percentages disliked both of them rather intensely. Yet 46% actually voted for Trump and 48% for Clinton. Pretty clearly, more than half of all voters voted insincerely for a candidate they didn't like because they liked the other major candidate even less! It was "vote for the lesser evil" on steroids. This is not likely to be a way to make a good decision. Many experts even point to Plurality voting as a cause of increasing polarization.

Much prior work has been devoted to the analysis of specific hypothetical voting examples, especially some which are cited as so-called “voting paradoxes.” How various voting methods handle these examples is compared and studied. A lot of attention has been paid to the concept of a Condorcet winner (a candidate that would beat all others if pairwise elections were to be held against each of the other candidates), as well as the Condorcet loser, the Smith set, etc. It has been proven that no ordinal voting method is possible which satisfies all of three very reasonable “fairness axioms.” We know that, not surprisingly, all voting methods are vulnerable to some degree to manipulation by strategic voting.

However, how important such internal details are to consistently choosing the correct winner in real elections is not clear. After all, it is the big picture performance in the real world in a wide variety of election contests that is of overriding importance. Other considerations are important only if, and to the extent that, they affect a method’s ability to consistently make the best decisions.

Efforts have been made to look at overall statistical performance by simulating many elections and testing various voting methods. This should be extremely worthwhile and illuminating **IF** the simulation actually reproduces real election characteristics well. Unfortunately, there is no guaranteed concrete way to know to what degree a simulation achieves that.

No one should trust the data produced by any election simulation project unless they have (or a trusted associate has) personally examined and understood in some detail all the computer code that was used to generate the data. A fairly detailed outline of the simulation software written for this project follows and the code certainly is available upon request.

Details of this Simulation Project

The 35,000-foot overview: A Government object instantiates multiple Election objects (typically 100,000). An Election object contains some number of Candidates (2 through 7) and Voters (typically 1,000). The Government hands each Election object to each of several VotingMethod objects which then separately obtain their particular ballot information from the Voters and determine the winner in accordance with each’s unique procedure. The VotingMethod’s winner is compared to the correct winner for the Election; statistics are accumulated.

Of course, the most critical aspect of the investigation is exactly how Elections are instantiated. First, it is important to model the fact that Voters have varying levels of satisfaction for Candidates, including dissatisfaction for some and “don’t care” for others. The range of satisfactions (arising from neural patterns in various people’s brains) must be mapped onto some absolute scale. The choice of the scale is arbitrary and mostly unimportant as it must be assumed that we have the correct “magic conversion factor” which converts brain satisfactions to the chosen scale. Although it is an adjustable software parameter, a scale of -100 “sats” to +100 “sats” (with zero representing “don’t care”) was chosen as having adequate resolution.

A group of Voters with a given average satisfaction for some Candidate are assigned satisfactions having a Gaussian distribution centered on the average and having a sigma of 20 (also a parameter). For that reason, the range for average satisfactions was limited to be 2 sigma inside the maximum range (-60 to +60). Any randomly generated satisfaction which would have lain outside the range was set to the maximum value (either -100 or +100).

It doesn’t work very well to just randomly assign “satisfaction levels” (Bayes fans may prefer to call these “voter utilities”) to Voter/Candidate pairs. As the number of Voters increases, this will all average out closer and closer to zero. Real elections have more than 20 or 40 voters and it should be possible to increase the number of Voters to whatever we like and have results approach stable (non-zero whenever appropriate) values. So, we must think about the “contest structures” that real elections have and model those. Here are the steps that were followed to accomplish that:

1. In a race in a real election, there are always candidates that are well-known to most voters, candidates that are less well-known and some candidates which very few voters have even heard of. So, the first step is to assign a “notoriety” to each Candidate. A random number >=0 and <1 is generated for each Candidate. Before assigning these notorieties to Candidates, they are sorted into descending order so that the first Candidate is always the best known and the last is the least known. This does not cause any loss of generality, but does enable us to see how wins are distributed “strongest” to “weakest” Candidate.
2. Each Candidate is “supported” by a “faction” of Voters. The total number of Voters for the Election is divided into such factions. The number of Voters in each Candidate’s faction is proportional to the Candidate’s notoriety.
3. A “profile” is next generated for each Candidate’s faction of Voters which defines how that faction views each of the Candidates. Of course, all voters of a Candidate’s faction always support their own Candidate, so their average satisfaction is set to the maximum average (+60). Their averages for all other Candidates are randomly assigned on the -60 to +60 range.
4. Finally, the Voters in each faction are instantiated. A (Gaussian) satisfaction (averaging +60) is generated for every Voter in a Candidate’s faction for that Candidate. However, for Candidates other than “their” Candidate, a new random number (>=0 and <1) is generated. If this new number is greater than the Candidate’s notoriety, satisfaction is set to zero (the Voter doesn’t know enough about this Candidate to have an opinion). Whenever the new number is less than or equal to the Candidate’s notoriety, a Gaussian satisfaction is generated centered on the Candidate’s average from the faction’s profile (generated in step 3). When each satisfaction is generated, it is added to the Candidates net total satisfaction

(which will determine the correct winner of the Election after all Voters have been instantiated). Each Voter ranks all the Candidates according to its satisfactions and is able to provide the ranking and the satisfactions upon request (e.g., to VotingMethod objects or the Election object).

5. After all Voters have been instantiated, the Election determines the correct winner along with the winning (highest) net total satisfaction value. It also determines and tallies whether there was a majority winner, whether the majority winner was the correct or an incorrect winner, whether there was a Condorcet winner, whether the Condorcet winner was the correct or an incorrect winner and also tallies the wins by Candidate. It is happy to provide most of this information to VotingMethods upon request. Thus, VotingMethods are able to easily have Voters fill out their "ballots," tally the ballots, determine a winner and then see whether or not their algorithm has chosen the correct winner. Whenever an incorrect winner is chosen, how much lower the net total satisfaction was than that for the correct winner is the error.
6. When the last Election has been processed, each VotingMethod publishes its results.
   - The percentage of incorrect choices made
   - The maximum error ("sats" on the -100 to +100 scale)
   - The average error
   - The RMS (root of the mean square) error
   - The RMS error divided by the RMS error for random selection times 100
   - The number of ties that had to be resolved (also secondary ties for MRCV)

Note that all Voters always vote sincerely (more about that later). All VotingMethods choose non-zero percentages of incorrect winners. However, this is not the best figure of merit for a VotingMethod. Choosing an incorrect winner which is nearly tied with the correct winner is not a horrible blunder. On the other hand, choosing an incorrect winner which has a radically lower net total satisfaction is indeed bad. For that reason, the RMS error is deemed the best single statistical indicator of a VotingMethod's ability to consistently render correct decisions.

These VotingMethods were coded and tested:

1. Random Selection – This "method" simply selects one of the Candidates at random. Predictably, its percentage of incorrect choices is 50% when there are two Candidates, 66.67% when there are three, 75% when there are four, etc. Random Selection turns out to have an RMS error of 21.86 when there are two Candidates and this error declines somewhat and fairly linearly to 17.36 when there are seven Candidates. If we view the RMS error of other methods as a percentage of the RMS error of Random Selection, it yields a pure number. It is hoped that this number might serve as a measure of method performance which is fairly independent of variations in the way simulations are done.

2. Plurality – Only Voters' first choices are counted and the Candidate receiving the largest number of first choices is proclaimed the winner. Ties are resolved by randomly selecting one of the tied winners.
3. IRV (Instant Runoff Voting) – A candidate having a majority of first choices is declared the winner. If there is no winner, the candidate having the fewest first choices is eliminated, lower choices, if any, are promoted and the process iterates. Ties are resolved by random selection of one of the tied Candidates.
4. Pairwise Comparison – If a Condorcet winner exists, it is declared the winner. When there is no Condorcet winner, IRV is used to determine a winner.
5. MRCV (pure) – Candidates are assigned an exponentially declining score based on the level at which the Voter has ranked them. An exponent of 0.7 was determined to work well. That is, a Candidate ranked first receives a 1.0 score for each Voter who ranked that Candidate first, plus 0.7 for each Voter that ranked the Candidate second, plus 0.49 for each Voter that ranked the Candidate third. The Candidate having the lowest total score is iteratively eliminated until only one Candidate (the winner) remains. Ties are resolved by eliminating the tied Candidate with the lowest number of first choices. In the extremely rare case of a secondary tie, one of the tied Candidates is randomly selected for elimination. After each elimination and prior to re-computation of point totals, lower-ranked choices, if any, are promoted to replace eliminated Candidates.
6. ScorePos – Voters score or rate Candidates on a scale. There are many variants in scales as well as with the processing of the data. Here, a scale of positive integers from 0 to R is used and Voters rate only Candidates about which they have positive opinions. The resolution of these ratings is determined by R, the maximum integer value of the scoring scale. This is a parameter for the Score method which may be set to any number from 0 to the maximum satisfaction value allowed for Voters (which has, so far, always been 100). Setting R to the maximum (100) allows Voters to rate the Candidates with the highest resolution and is called Score100. Setting the resolution to a lower value requires that Voters project their satisfactions onto a coarser ballot scale with less resolution. This gives rise to a family of methods which can be evaluated. The Score30 method, for example, requires that Voters rate Candidates with one of the 31 values from 0 to 30. Note that Score0 is a method identical to Random Selection since all Candidates are always tied at zero in every Election and ties are resolved by random selection. Especially note that Score1 is a special case in the sense that it is identical to the Approval method. The scores for each Candidate are totaled and the Candidate with the largest total is the winner. Ties are resolved by randomly selecting one of the tied Candidates.
7. Approval – Any and all Candidates that have positive satisfactions are "approved" by each Voter. Candidates with zero or negative satisfactions are not approved. The Candidate receiving the most approvals is declared the winner. Ties are resolved by randomly selecting one of the tied Candidates. Note that this method is identical to Score1, but was

nevertheless separately coded for verification. In all cases Approval and Score1 did produce identical results.

8. TWV (True Weight Voting) – This method simply asks Voters for their opinions (satisfaction or dissatisfaction) for each Candidate, which Voters are happy to sincerely provide. As usual, the Candidate having the highest positive (or least negative) total is the winner. Ties are resolved by random selection of one of the tied Candidates. Voters must project their satisfactions onto the scale the ballot provides. Exactly as with Score voting, the scale is determined by the R parameter and it similarly gives rise to a family of methods. TWV100 has "full" resolution and near-perfect performance. However, even it can yield an "incorrect" choice in rare cases of ties. Since ties are broken by random selection, a winner different from the one chosen by the Election object is possible. The error, however, is always zero. As a concrete example, the ballot scale for TWV30 is the 61 integers from -30 to +30. As with Score0, TWV0 is identical to Random Selection.

These voting methods either were not coded and evaluated or were not reported:

1. ScoreAvg – A Score method quite different from the one described above has been proposed, discussed and evaluated (some years ago) in a different simulation project.[3] The easiest way to gain a quick appreciation for the difference is through an illustrative example. Consider a two-candidate election with 1,000 voters. Candidate A is well known and only 100 voters do not score candidate A (express "No Opinion"). The total of all the scores of the 900 voters who do score A is 36,000. Candidate B is less well known and 800 voters do not score B ("No Opinion"). The total of all the scores of the 200 voters who do score B is 12,000. By the definition stated at the beginning of this paper, A is the overwhelming winner since A's election would result in three times the total voter satisfaction (36,000) than would B's election (12,000). However, the ScoreAvg method advocated at rangevoting.org would yield the opposite result. It calculates the average score for each candidate and identifies the winner as the candidate which has the highest <u>average</u> score. A's average score is 36,000 divided by the 900 voters who scored A, which is 40. Candidate B's average is 12,000 divided by the 200 voters who scored B, which is 60. So, B wins handily with an average of 60 compared to A's 40. Thus, the 200 voters who voted for B are able to overrule the 900 who voted for A because their votes receive a weighting 4.5 times greater the voters who scored candidate A! In addition to frequently selecting the wrong winner (by the definition stated in this paper), weighting voters scores differently would certainly seem to raise "one person, one vote" issues. The ScoreAvg method was nevertheless coded and tested. Unsurprisingly, the performance was awful – better than random selection, but somewhat worse than Plurality. Thus, for the several reasons discussed above, it was not reported as a part of this study. Recently, Warren Smith has advised that his simulation assigned utilities for every voter/candidate pair (there

[3] See Warren D. Smith's website: https://rangevoting.org/

were no "No Opinion" cases). Thus, the divisors were in every case the same – the total number of voters in the simulation. That means the method that was tested and reported was the equivalent of simply summing the scores for each candidate. That is, the advocated ScoreAvg method was not actually tested in the simulation.

2. The Borda score method, although it has been discussed for centuries, was not included in this simulation project. The original Borda scoring method requires a complete ranking of all candidates and the weighting factors for ranking levels depend upon the number of candidates. There is no way a complete ranking of all candidates can be obtained from all voters in any real public election and it's hard to come up with reasonable logic to support weightings varying with the number of candidates. A large number of variants to the original Borda score method have been proposed, some of which would be much more serviceable. However, funding for the study was extremely limited ($0.00) and the person-hours available for this project also were limited.

Plurality was included because of its widespread use. Of the ordinal methods, IRV was included because it has a significant number of fans (also, it is in use a few places), Pairwise Comparison was evaluated because of the extreme attention lavished upon it over the centuries and MRCV was included because it has been argued that MRCV is the best possible ordinal method.[4] MRCV can be a scoring method or an iterative last-man-standing method. Prior testing has shown that the latter has somewhat better performance than the former, so the latter is what was included. The spectrum of ordinal methods is considered well represented by those chosen. It is hoped that nothing of great significance was missed.

It was determined that statistics stabilized when at least 250 to 400 Voters participated in each Election and for runs of at least 25,000 Elections. Therefore, 1,000 Voters was always specified for every Election and each run was set for 100,000 Elections. Each such run took approximately two minutes (single threaded) on a four-year-old desktop with a 3.4GHz Intel i7 CPU. Runs were made for 2, 3, 4, 5, 6 and 7 Candidate Elections.

Results

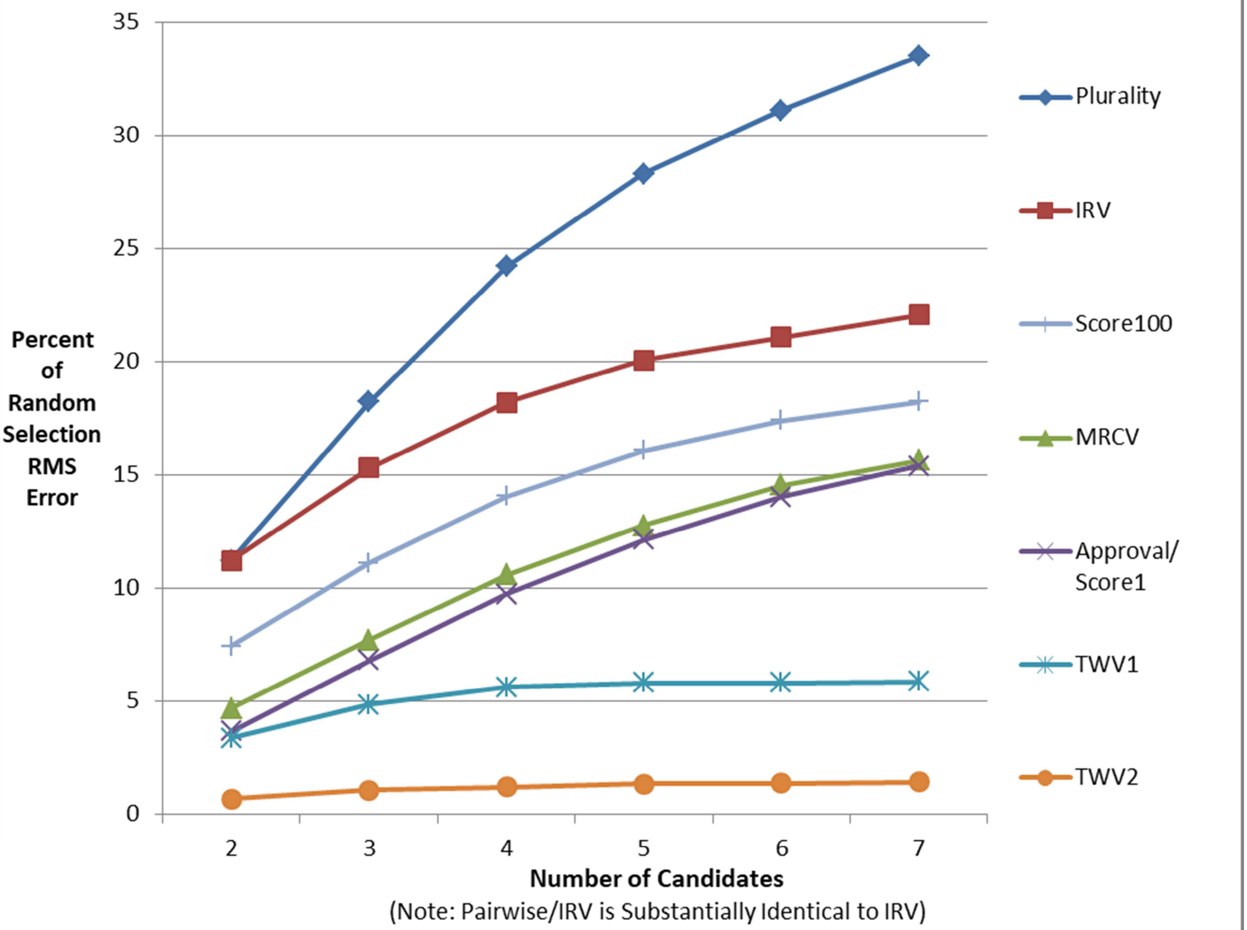
The table below summarizes the information produced by all 600,000 Election objects.

Number of Candidates	Majority Winners	Incorrect Majority Winners	Condorcet Winners	Incorrect Condorcet Winners	A Wins	B Wins	C Wins	D Wins	E Wins	F Wins	G Wins
2	99,832	7,513	99,832	7,513	86.7%	13.3%					
3	55,281	2,655	71,391	4,162	72.2%	22.7%	5.1%				
4	21,318	512	49,982	2,064	60.6%	26.9%	10.2%	2.2%			
5	6,375	91	37,315	1,206	51.4%	28.3%	14.0%	5.2%	1.1%		
6	1,588	12	29,719	780	44.7%	28.3%	16.0%	7.6%	2.8%	0.6%	
7	313	0	24,390	552	39.7%	27.2%	17.2%	9.6%	4.4%	1.5%	0.3%

[4] See "A Comprehensive, Conclusive Analysis of Ordinal Voting Methods" at http://royminet.org/voting-elections/

The data table and chart below show the "punchline," that is, the percentage of the RMS error of Random Selection that each of the tested methods achieved. Three things are not shown: 1) the plot for the reference Random Selection method which, of course, is always 100%; 2) the plot for TWV100 which is always 0%; and 3) the plot for Pairwise Comparison. Perhaps surprisingly because of the amount of attention heaped upon it for centuries, the Pairwise Comparison method was just microscopically better than IRV. Its plot is in all cases essentially on top of the IRV plot, so it was decided to eliminate that chart clutter (though the tables do show the data).

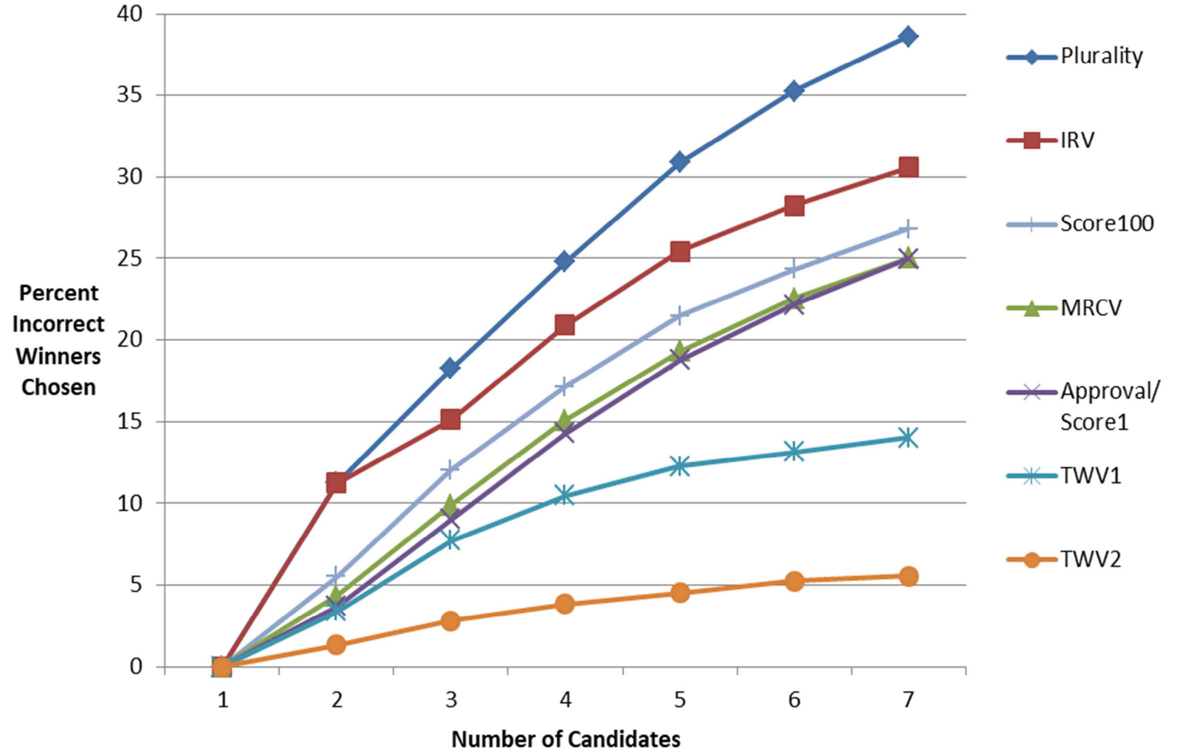
Candidates	TWV2	MRCV	Approval/Score1	TWV1	Score100	Pairwise/IRV	IRV	Plurality
2	0.698	4.692	3.673	3.373	7.428	11.215	11.228	11.224
3	1.065	7.718	6.815	4.867	11.125	15.267	15.325	18.243
4	1.223	10.602	9.737	5.626	14.063	18.171	18.210	24.215
5	1.356	12.776	12.130	5.826	16.096	20.021	20.073	28.294
6	1.380	14.543	14.022	5.829	17.410	21.052	21.083	31.095
7	1.428	15.649	15.408	5.885	18.238	22.043	22.076	33.480



(Note: Pairwise/IRV is Substantially Identical to IRV)
(Note: For any # of Candidates, TWV100 is 0.000% and Random is 100%)

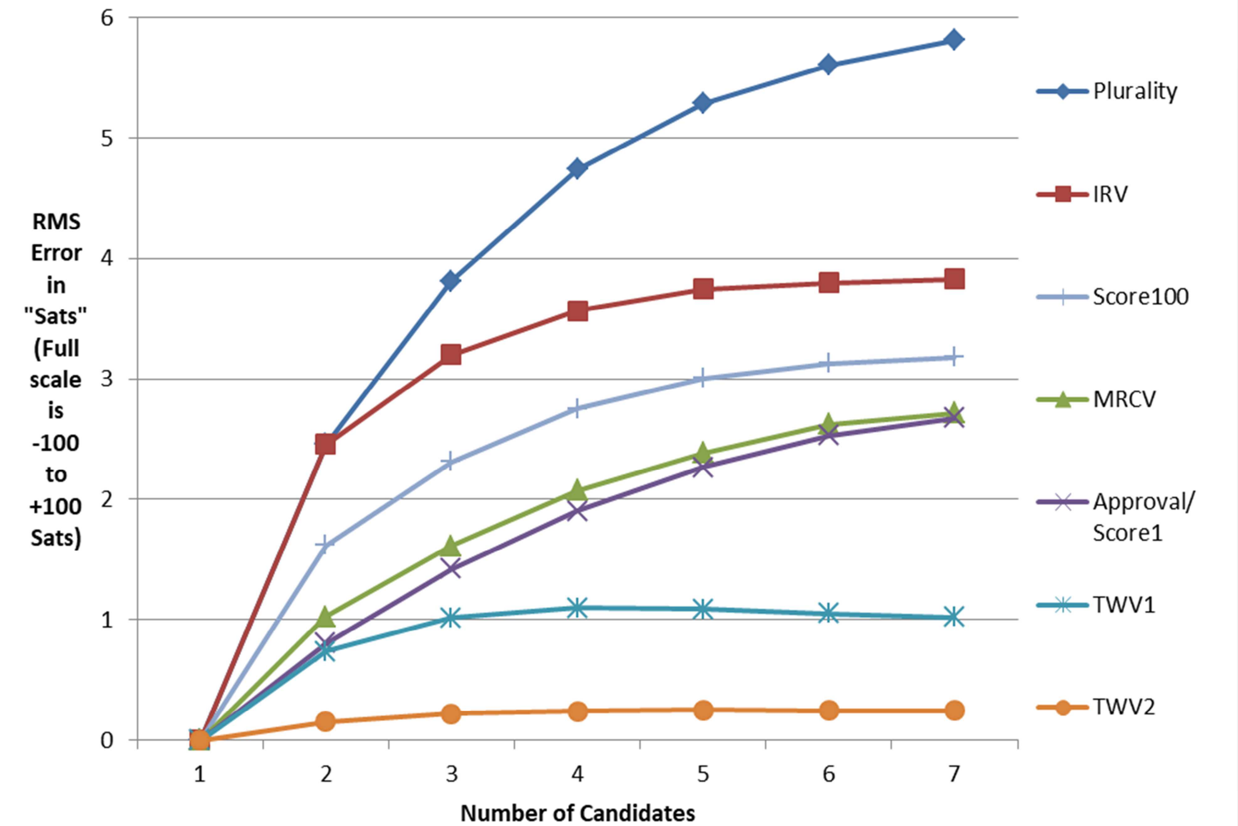
The following two sets of data tables and charts show the percentage of incorrect choices made and then the actual RMS error. The relative performance of the methods remains basically the same.

Candidates	TWV2	MRCV	Approval/Score1	TWV1	Score100	Pairwise/IRV	IRV	Plurality
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.333	4.323	3.673	3.373	5.494	11.215	11.228	11.224
3	2.804	9.893	9.033	7.722	12.019	15.052	15.105	18.243
4	3.829	15.069	14.290	10.482	17.134	20.873	20.916	24.800
5	4.519	19.294	18.768	12.238	21.483	25.405	25.453	30.879
6	5.246	22.537	22.169	13.140	24.349	28.223	28.249	35.274
7	5.554	25.013	24.980	14.001	26.798	30.537	30.566	38.609



(Note: Pairwise/IRV is substantially identical to IRV)
(Note: TWV100 is 0.000% for any number of candidates)

Candidates	TWV2	MRCV	Approval/Score1	TWV1	Score100	Pairwise/IRV	IRV	Plurality	Random
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.152	1.026	0.803	0.737	1.618	2.452	2.454	2.454	21.860
3	0.222	1.613	1.424	1.017	2.312	3.190	3.202	3.812	20.894
4	0.240	2.075	1.906	1.101	2.753	3.557	3.564	4.740	19.573
5	0.253	2.386	2.265	1.088	3.006	3.739	3.748	5.284	18.674
6	0.248	2.622	2.528	1.051	3.128	3.795	3.800	5.605	18.027
7	0.249	2.716	2.674	1.021	3.178	3.826	3.832	5.811	17.356



(Note that Pairwise/IRV is substantially the same as IRV)
(TWV100 RMS error is zero for any number of candidates)

Insincere or Strategic Voting

As previously mentioned, no provision was made in the software to simulate strategic voting. However, insincere voting likely is a serious problem (or significant effect) in real elections as was discussed with respect to the 2016 US presidential election. Clearly, the best possible decisions are supported by the data when voters are able to accurately convey the entire range of their satisfaction or dissatisfaction (or lack of either) for each candidate and they actually do so honestly. It does not seem like a foregone conclusion, though, that insincere voting always degrades decision quality under all circumstances. It is conceivable that, by voting strategically, voters may in some cases be able to at least partially compensate for the shortcomings of a bad voting method, causing it to render somewhat better decisions than it would if they had voted sincerely. In any case, it certainly makes sense to first gain an understanding of how various methods work with sincere data. The problem of minimizing voters' motivation to vote strategically can then be separately addressed.

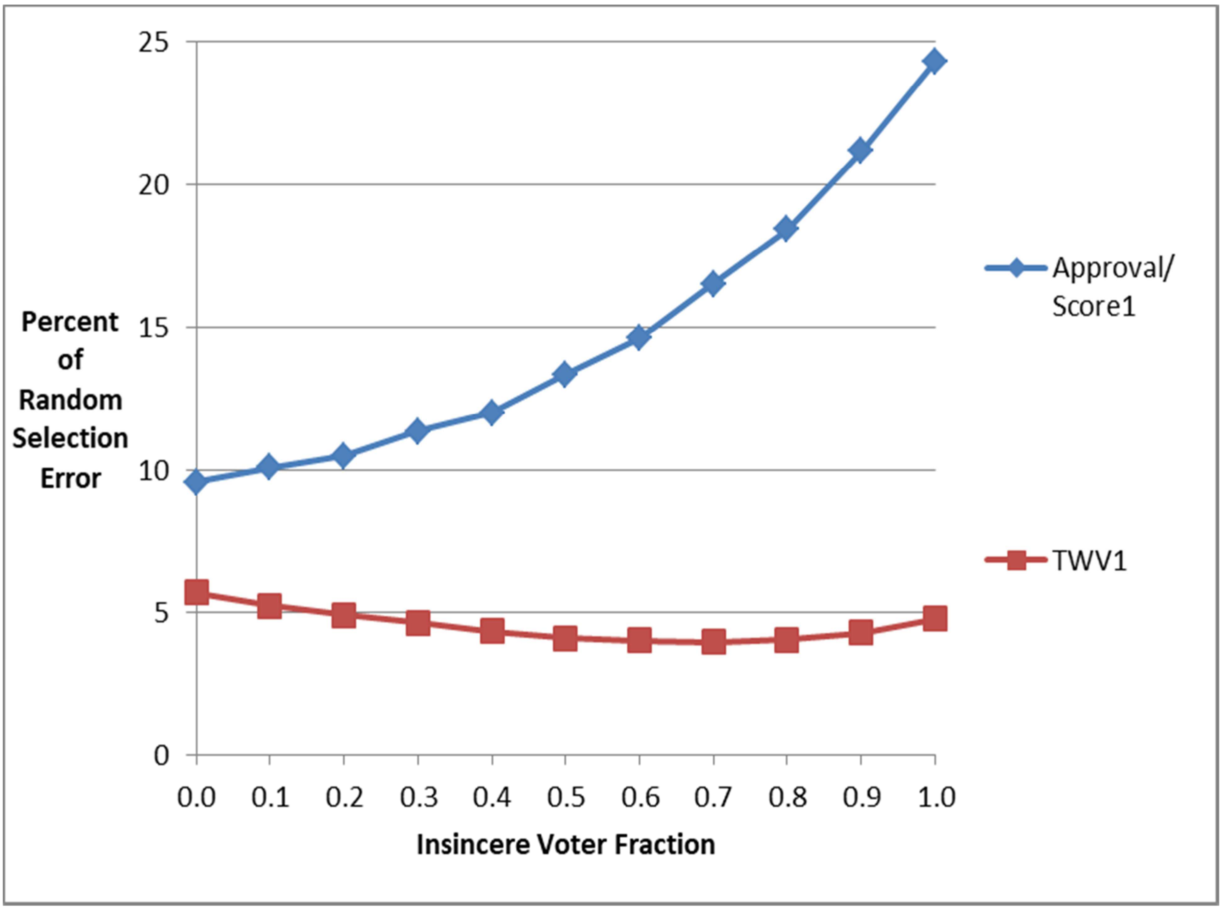
It is useful to think about insincere voting as having two main forms. In one form voters deliberately vote for a candidate they don't like very much because they think that candidate has a chance to win and thereby thwart a probable win by the candidate they most despise. For methods which allow voters to indicate dissatisfaction, the mirror of that would be voting against a candidate they like because they think that candidate has a chance to defeat the candidate they favor even more strongly. This is the previously mentioned "lesser of evils" pressure. Voters' decisions as to what strategy they may think best and what data they will actually provide is strongly dependent upon their individual perceptions of a race as conditioned by advance polling data, news reports and opinions. Representing this well in election simulation software is somewhat difficult and possibly hazardous to meaningful results. It was not attempted in this simulation project, at least not so far.

Another form of strategic voting is simpler and more straightforward. It is just the decision not to provide any "extra" information which might help a candidate other than the voter's first choice. For ordinal methods, that means not registering a second or third choice. For Approval/Score1, it's just not indicating approval for all candidates of which the voter actually approves. In fact, Approval/Score1 may be especially susceptible to this kind of insincerity since it provides no way for a voter to specify a first choice – all approvals are equal. Many voters may wish to identify their first choice. Implementing such behavior in software is straightforward and reliable.

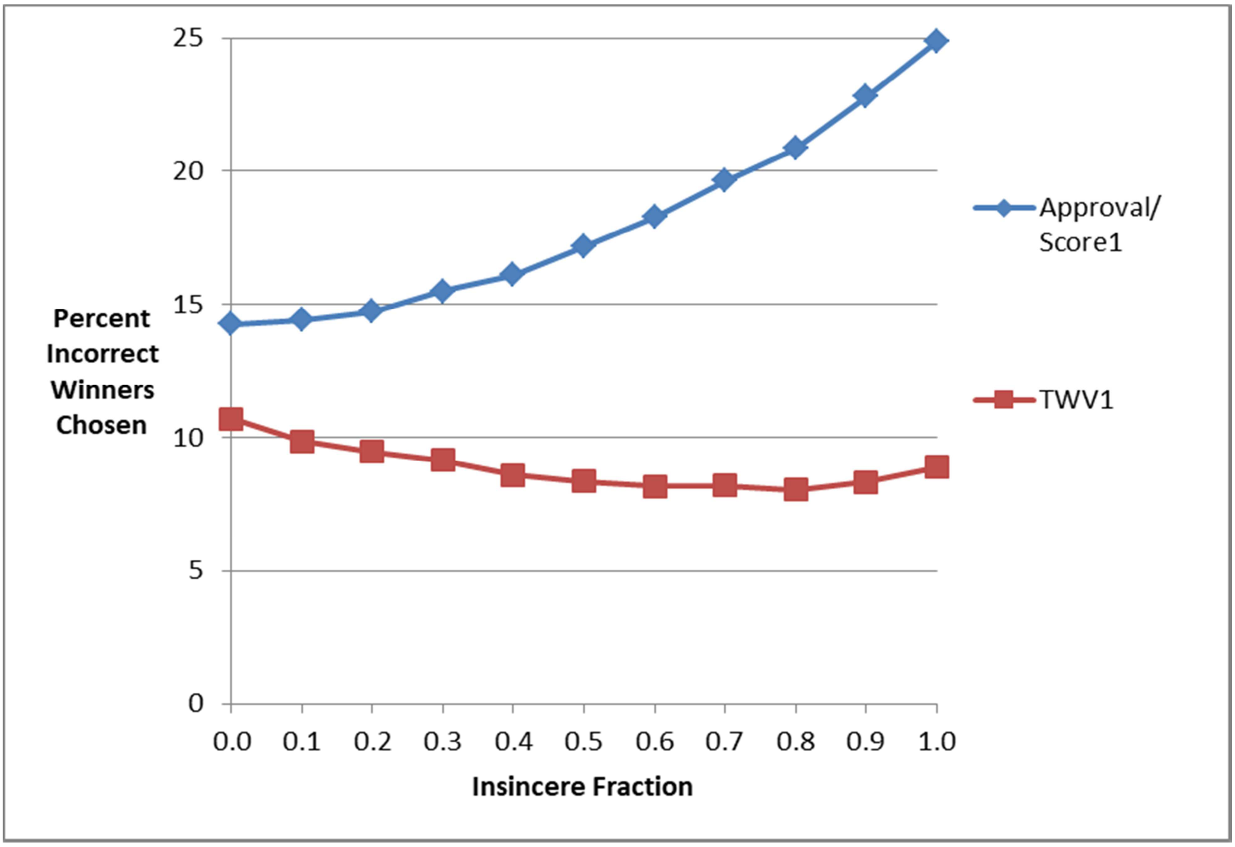
Thus, it was decided to investigate what effect this latter type of strategic voting may have on the Approval/Score1 and TWV1 methods.[5] An "Insincere Fraction" parameter was introduced which randomly divides voters into sincere and insincere categories. At 0.0, all Voters were sincere (just as was the case in all prior runs). At 0.5, a randomly selected half voted sincerely and half voted insincerely. At 1.0, all Voters voted insincerely; that is, if they had one or more positive satisfaction Candidates, they approved only the highest one (also, for TWV1, if they had one or more negative satisfaction Candidates, they disapproved only the most negative one). An additional 1.1 million 4-Candidate Elections were conducted, 100,000 for each 0.1 increment of the Insincere Fraction. The resulting data tables and charts appear below. Again, note that all these Elections had four Candidates.

[5] A method substantially the same as TWV(1) has been proposed several times under differing names. The first time in recent history may have been in 1989 by Dan Felsenthal who called it CAV for Combined Approval Voting. It does not appear to have attracted much attention. Also, a similar method was employed in the Republic of Venice for many years prior to Napoleon's rule.

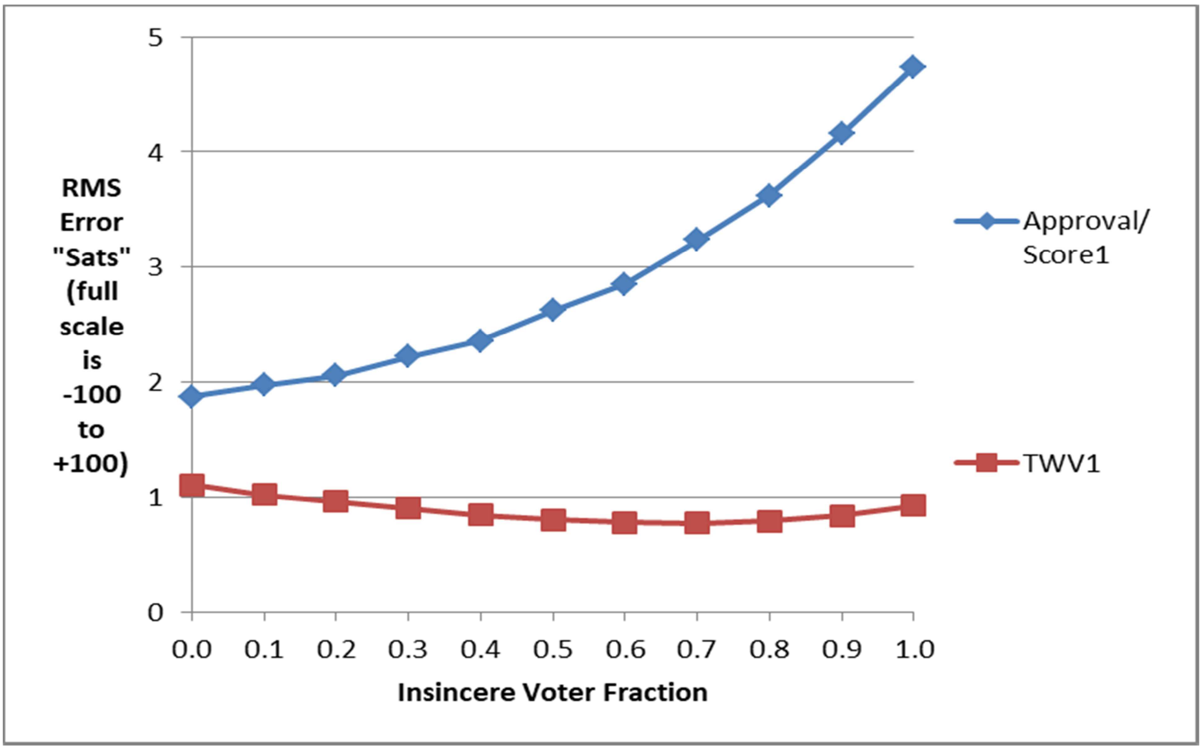
Insincere Fraction	Approval/ Score1	TWV1
0.0	9.551	5.665
0.1	10.061	5.198
0.2	10.462	4.884
0.3	11.330	4.604
0.4	11.991	4.307
0.5	13.335	4.079
0.6	14.590	3.987
0.7	16.504	3.947
0.8	18.400	4.018
0.9	21.152	4.271
1.0	24.263	4.733



Insincere Fraction	Approval/ Score1	TWV1
0.0	14.266	10.719
0.1	14.414	9.873
0.2	14.753	9.478
0.3	15.506	9.163
0.4	16.121	8.597
0.5	17.168	8.357
0.6	18.272	8.172
0.7	19.631	8.214
0.8	20.864	8.059
0.9	22.815	8.337
1.0	24.862	8.901



Insincere Fraction	Approval	TWV1
0.0	1.875	1.112
0.1	1.975	1.020
0.2	2.057	0.960
0.3	2.226	0.904
0.4	2.357	0.847
0.5	2.623	0.802
0.6	2.856	0.780
0.7	3.236	0.774
0.8	3.618	0.790
0.9	4.154	0.839
1.0	4.738	0.924



As can be clearly seen, Approval/Score1 deteriorates markedly, but TWV1 cruises along remarkably unaffected by this type of strategic voting. Its performance actually improves slightly with fractions in the 0.4 to 0.9 range. This appears to be a mild example of insincere voting compensating for a deficiency (in this case, resolution) of the method.

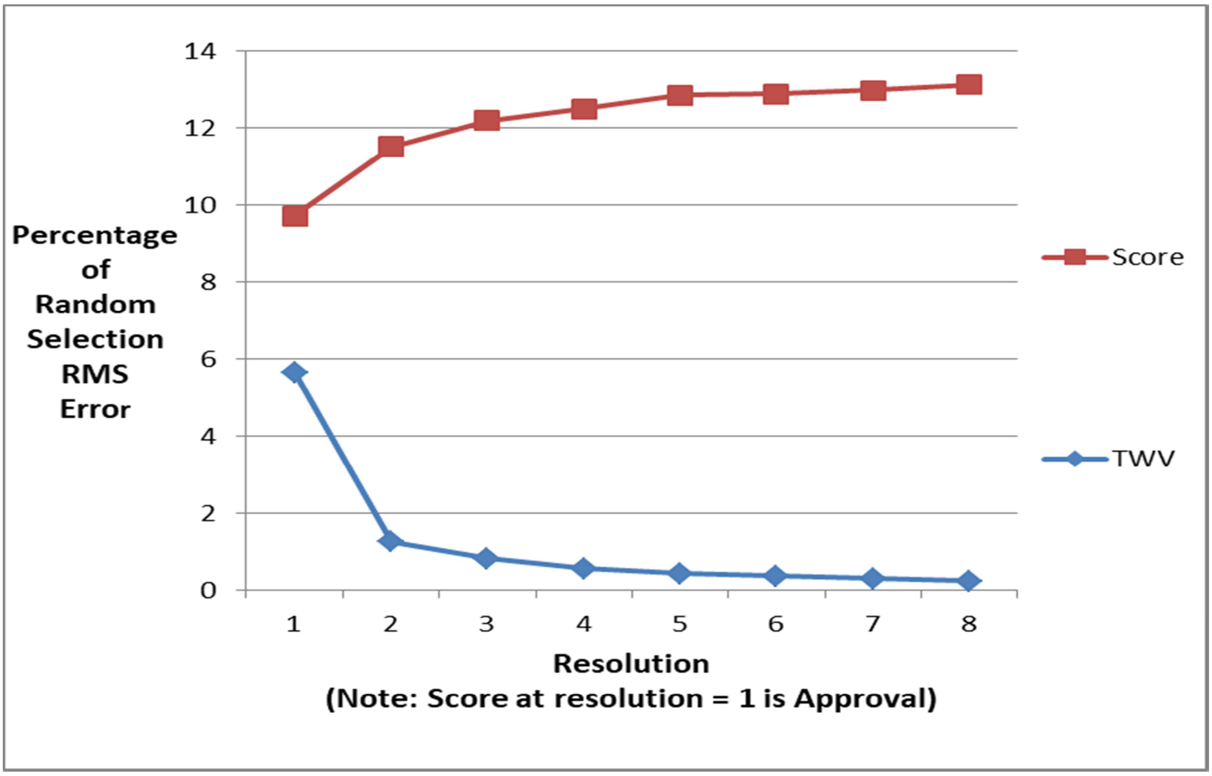
The Impact of Resolution on Cardinal Methods (TWV and Score)

Generally, one thinks better resolution is a good thing and reducing resolution would be expected to impair performance. Indeed, that is the case for TWV. Note that TWV100 has near-perfect performance. TWV2 works less well and TWV1 is worse yet (although TWV1 still significantly outperforms any non-TWV method).

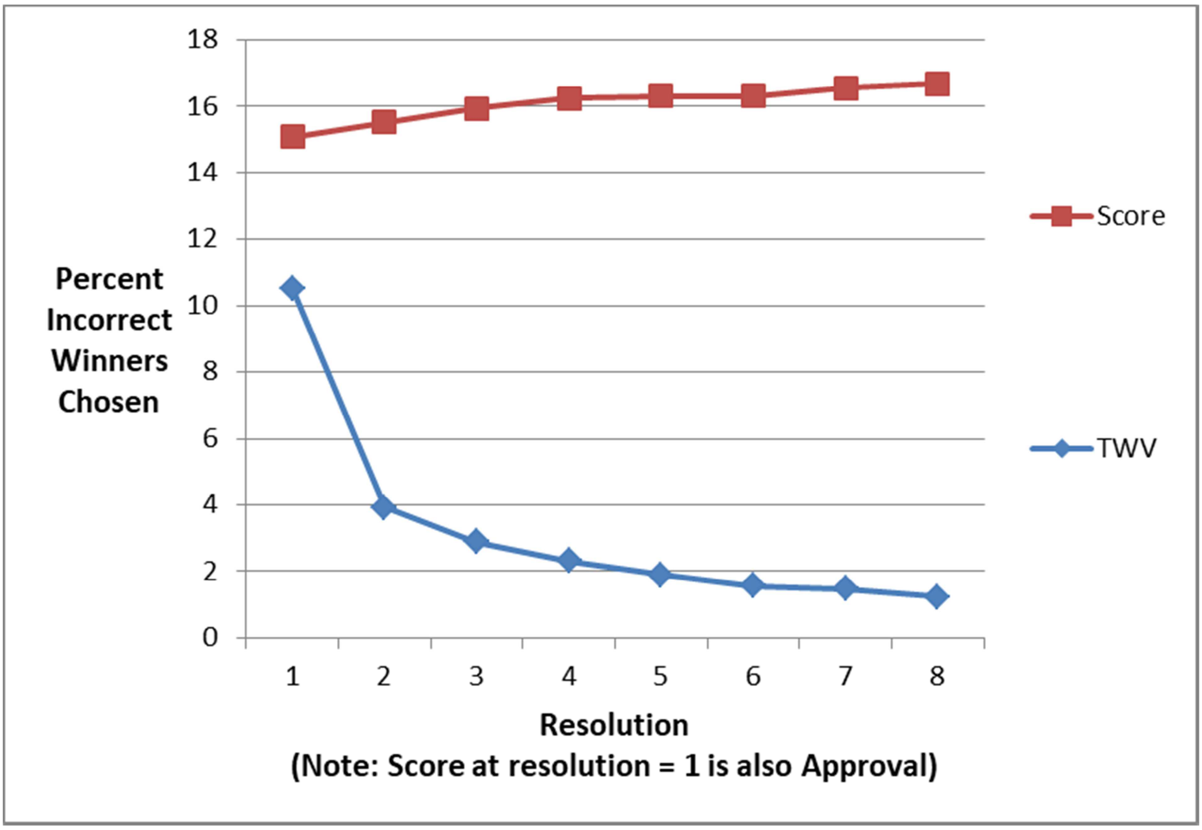
However, the usual expectations do not appear to apply to Score. The performance of high-resolution Score100 hangs out about halfway between that of ordinal methods IRV and MRCV. Low-resolution Score1, the simplest of all cardinal methods, turns out to be somewhat better. It closely matches and slightly betters the performance of MRCV, the best of the ordinal methods.

The impact of resolution was more closely examined with a series of simulations which varied resolution for TWV and Score in four-candidate elections. The three sets of tables and charts below present those results.

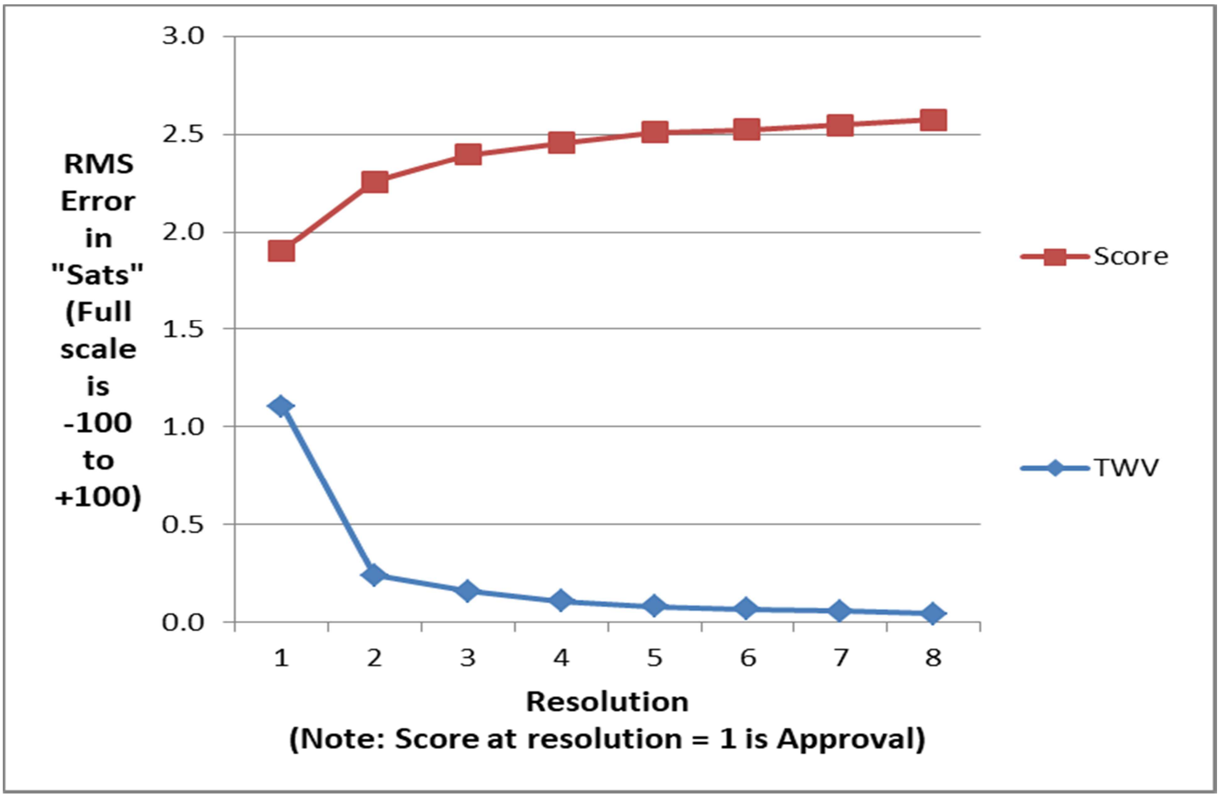
(4 Candidates)	% of Random	
Resolution	**TWV**	**Score**
1	5.649	9.716
2	1.236	11.508
3	0.809	12.187
4	0.553	12.490
5	0.423	12.841
6	0.346	12.880
7	0.290	12.973
8	0.228	13.114
16	0.085	13.473
32	0.031	13.781
64	0.010	13.957
96	0.006	13.904
100	0.000	14.063



(4 Candidates)	% Incorrect Choices	
Resolution	TWV	Score
1	10.517	15.061
2	3.925	15.515
3	2.881	15.928
4	2.312	16.239
5	1.900	16.298
6	1.563	16.310
7	1.473	16.546
8	1.234	16.685
16	0.649	16.853
32	0.347	17.007
64	0.156	17.310
96	0.108	17.152
100	0.003	17.204



(4 Candidates)	RMS Error	
Resolution	TWV	Score
1	1.107	1.904
2	0.242	2.256
3	0.159	2.394
4	0.109	2.455
5	0.083	2.508
6	0.068	2.522
7	0.057	2.547
8	0.045	2.573
16	0.017	2.642
32	0.006	2.698
64	0.002	2.743
96	0.001	2.730
100	0.000	2.757



TWV resolution can be reduced to TWV2 with little degradation, so that appears to be the "sweet spot." It is suspected (but was not tested) that, although the results shown should apply for elections with large numbers of voters, higher resolutions may become more important in elections with small numbers of voters.

Score is highly susceptible to strategic voting pressures. In addition to the normal array of strategic techniques, there is an additional extremely obvious one. One can most help candidate(s) one likes by always voting the maximum allowed satisfaction value for those you like and zero for the ones you don't like. Thus, Score>1 morphs into Score1 (Approval voting) as voters figure that out. Based upon the above data, that would be a good thing and appears to provide an example of insincere voting partially compensating for method deficiencies and enabling it to render better results. A deeper understanding likely could be achieved by detailed analysis of the characteristics of selected Elections in which Score1 chooses wisely and Score100 chooses poorly. Of course, the same strategic vulnerability applies for TWV>1 methods; however, voters motivation may be somewhat lessened by their ability to vote against candidates they don't like.

Conclusions

To the extent that real-world elections have been meaningfully simulated by this project, the following conclusions are supported.

1. It has long been known and it has again been quantitatively shown that Plurality is truly awful and is the worst of all voting methods. It should be replaced as soon as is reasonably possible.
2. IRV does improve upon Plurality somewhat in elections which have four or more candidates. It should relieve Plurality's "vote for the lesser of evils" insincerity problem to some small extent. However, much better methods are known. IRV is not recommended and should be replaced wherever it may already be in use.
3. Pairwise Comparison is only a microscopic improvement over IRV. The concept of a Condorcet winner sounds like it ought to be great, but does not appear likely to have much significance in real elections.
4. A new method was proposed in 2007 called MRCV. More recently, it was argued that MRCV is the best possible ordinal method. These simulation results support that claim. MRCV does perform significantly better than Plurality, Pairwise Comparison or IRV. However, the most rudimentary cardinal method (Score1/Approval) achieves substantially the same or slightly better performance and is certainly to be preferred for its simplicity.
5. Although Score1/Approval voting substantially duplicates the good performance of MRCV with sincere voters, with a large fraction of insincere voters and 4 or more candidates, it approaches the poor performance of IRV and quickly exceeds (sincere Voter) MRCV error as the fraction of insincere voters grows.
6. TWV1 not only achieves significantly better performance than Approval and MRCV with sincere voters, but it also demonstrates remarkable immunity to the most common and obvious form of strategic voting. Equally remarkable is that TWV1 maintains its ability to pick correct winners quite well as the number of Candidates increases, while all other methods deteriorate significantly. The remaining problem is to remedy strategic pressures to vote for the lesser of evils (see 2 under Future Work) and the strategy of preemptively voting against all but one's first-choice candidate. The importance of giving voters a way to register dissatisfaction with candidates they do not like is indicated as well. TWV1 is still

simple, understandable and very easy to implement. It appears so far to be the best method with which to replace Plurality and IRV. There does not seem to be any strong reason to delay doing so. Note that Plurality has about 5 times the error of TWV1.

7. Very high resolution is not required for good performance. With sincere voters, TWV1's three levels work surprisingly well. TWV2's five levels are noticeably better (Plurality exhibits 16 to 22 times its error). Further increases in resolution yield only small incremental improvements. However, increasing resolution to anything better than TWV1 exposes the method to the same obvious strategic voting problem as was described for Score voting. The best final choice probably depends upon a complexity of tradeoffs if or when modifications are introduced to minimize strategic voting. Possibly something like the point system described in 4 under Future Work could help. However, switching to TWV1 is such a major improvement that its adoption should not be delayed.
8. The importance of providing a way for voters to input dissatisfaction for one or more candidates is again noted. Otherwise, there is no way for a method to distinguish between candidates voters do not like and those they don't care about. It seems highly likely that this information improves a method's ability to consistently pick the option which maximizes voter satisfaction under important and difficult circumstances which probably occur quite often. This is true for elections with any number of candidates, but the data indicate it becomes essential when there are more than three or four candidates. Dissatisfaction needs to offset satisfaction to discourage nomination of highly polarizing candidates and encourage nomination of more widely supportable ones. It seems likely that allowing voters an "outlet" to indicate their dissatisfaction(s) may also somewhat reduce pressures to vote insincerely. As less polarizing, more widely supportable candidates become the norm, the pressure to vote strategically will decrease significantly. Surely, the best way of all to reduce the pressure to "vote for the lesser evil" is to have less "evil" candidates!
9. Indications are that no method limited to ordinal ranking of the options can achieve more than "mediocre" performance because the limited information gathered simply cannot support consistently making the best decisions under all circumstances. Plurality collects woefully inadequate data. IRV gathers valuable additional information in the form of second and sometimes third choice rankings. However, IRV is doomed to poor performance by two major flaws. First, the majoritarian rule is at its heart. Second, it totally ignores the valuable additional data it has collected when making the crucial decision as to which option is to be eliminated as the weakest. MRCV fixes both of these flaws, performs better and is (arguably) the best that can be done when limited to just ordinal data. However, cardinal methods pick up where ordinal methods poop out. It seems clear that the most fruitful path to realizing the truly excellent performance of TWV1 or TWV2 with sincere data is to focus on ways to reduce voters' motivation to vote strategically. It seems apparent

that radical improvements over Plurality and/or IRV election methods are achievable and should be implemented to “save the world from Plurality (and IRV) voting.”

Future Work

1. It would be good to more completely investigate TWV1 and TWV2’s resistance (or susceptibility) to various forms of strategic voting.
2. Giving voters the option to designate a candidate as a conditional top choice should be considered as a mechanism to reduce or eliminate important strategic voting pressures. Here is how that would work using TWV1. Voters may mark any number of candidates they like as +1. They also may mark any number of candidates they don’t like as -1. A third option is to mark a candidate as B, for “Backup Top Choice;” a limit of one B makes sense, but multiple Bs could be allowed. Note that a candidate marked B may or may not also be marked -1. The +1s and -1s are totaled for each candidate as usual; candidates not marked or marked only with B on a ballot count 0. Next, instead of designating the candidate with the highest total the winner, the candidate having the lowest total is eliminated. For any voter who has no +1 candidates remaining, any candidate(s) with a B designation are promoted to +1 (or to +2 if TWV2 were used). If a promoted candidate had also been marked -1, the promotion cancels the -1. The procedure iterates until only one candidate is left standing which is then named the winner. Note that marking a candidate with a B has absolutely zero effect on the outcome unless and until that candidate is promoted to a +1. There is no strategic reason a voter should hesitate to sincerely indicate conditional top choice candidate(s) as doing so cannot (initially) help or hurt any candidate. Also, there is no reason to hesitate awarding +1 only to those candidate(s) a voter sincerely favors, even though they may be “long shot(s),” since the “lesser of evils” can be marked B. If as feared, the long shots bite the dust, the voter then still is able to weigh in with a vote for the lesser evil. We can label these methods TWV1B and TWV2B to indicate the backup top choice option. However, a more popular moniker for this general family of methods might be “LHT” for “Love/Hate/Tolerate” voting. Conditional top choices would be minimally more complex for voters, but tallying the ballots becomes much more complex. Still, with the technology that is available today, the additional tallying complexity should not rule it out. Certainly, care must be taken to assure easy and complete auditability.
3. Employing the “percentage of Random Selection’s RMS error” as a voting method figure of merit, variations and tweaks to the way elections are simulated should be tested to see if the relative performance of the various voting methods changes or stays mostly the same. Invariance to such simulation differences should bolster confidence that measured method performances are fundamentally meaningful and applicable to real elections. A little of this was done (while developing the simulation software for this project) and was generally encouraging, but much more should be done.
4. The idea of giving voters some fixed number of “points” which they may then allocate to candidates in the way each voter thinks best was proposed and briefly discussed in 2007

and again in 2016.[6] Further exploration of such mechanisms, especially as a way to improve TWV2's resistance to strategic voting may be worthwhile.

[6] See "Voting for Better Decisions", Appendix B, at http://royminet.org/voting-elections/