Can Election Predictions Be Self-Defeating Prophecies?

Liam McCarty
Stanford University

This article rigorously defines a *self-defeating prophecy* and a model for a binary, democratic election. It develops a theory of rational voting behavior based on that model and proposes an experiment to explore the relationship between election predictions and outcomes. Specifically, it investigates whether predictions can be self-defeating prophecies and, if so, under what conditions. It hypothesizes that the total number of votes for each candidate will be greater than the theory predicts but by the same amount (so that the election outcomes do not change). In addition, it provides a formula for calculating theoretical voter turnout and discusses possible future research.

Introduction

In a now classic 1948 article, Robert Merton coined the term *self-fulfilling prophecy* to describe "a *false* definition of the situation evoking a new behavior which makes the originally false conception come *true*" (emphasis in the original; Merton 1948). The term has since become standard terminology in psychology and across the social sciences. Self-fulfilling prophecies have been the subject of landmark studies such as those examining the effects of teacher expectations on student performance (Brookover, Rosenthal, and Jacobson 1969). A host of other, less noted studies consider self-fulfilling prophecies in everything from how perceived popularity affects download rates in online music markets (Salganik and Watts 2007) to how expectations affect Israeli military recruits (Eden 1990).

Subsequent scholars have built on and clarified Merton's definition. This paper uses a definition proposed by Michael Biggs (Biggs 2009). Biggs defines a self-fulfilling prophecy (in its simplest form) as a dynamic process between two actors with two criteria. The first criterion is a causal sequence of the following form (A_1 and A_2 are the actors, b is a behavior of an actor, and c is a characteristic of an actor):

(1) A_1 believes that "A_2 is c."
(2) A_1 therefore does b.
(3) Because of (2), A_2 becomes c.

The second criterion is that one or both of the actors *believes* (falsely) that the causal sequence was actually the following:

(1) A_2 is c.
(2) Because of (1), A_1 believes that "A_2 is c."
(3) A_1 therefore does b.
(4) Because of (1), A_2 is still c.

The inverse of a self-fulfilling prophecy is a self-*defeating* prophecy[1]. Biggs notes that this "has received little attention, although it may have considerable importance" and does not discuss it further (Biggs 2009). As this indicates, there is a need for both a formal definition of a self-defeating prophecy and an examination of it in specific cases.

This paper defines a self-defeating prophecy and proposes an experiment to examine a particular case: whether election predictions can be self-defeating prophecies. As Biggs writes, "What distinguishes social science from natural science is the potential for reality to be altered by theory. A theory of society could, in principle, prove self-fulfilling" (Biggs 229). Election predictions are a small-scale example of just such a social theory; this study investigates whether and how they might alter reality. The experiment described here does so by considering whether and how voting behavior and election outcomes depend on what voters believe to be the probabilities of election outcomes.

In proposing the experiment, this paper defines an election model and develops a theory of rational voting behavior based on that model. It hypothesizes that more votes will be cast in the experiment than the theory predicts but by the same amount for each candidate so that election outcomes do not change.

This is a topic of considerable relevance in our politically charged time. It is not uncommon to hear arguments about how expectations may have affected the result of a contentious election. Shortly before the 2016 U.S. presidential election, for example, many if not most polls gave Hillary Clinton a very high chance of winning. When Donald Trump won, there was widespread discussion about how the predictions may have been flawed and whether they influenced the outcome.

The results of this study could shed light on election dynamics that have yet to be considered, like how the cost of voting influences voting behavior. It could thereby help psychologists better understand how people make voting decisions on the level of individuals and societies. As an addi-

[1] Merton called it a *suicidal prophecy* (Merton 1948).

tional bonus, the election model developed here provides a way to calculate theoretical voter turnout as a function of the cost of voting and the expected probabilities of outcomes. This has clear value to social scientists and anyone involved in democratic elections. Perhaps most importantly, the election model defined here and the theory developed around it can serve as a helpful foundation for future research.

Definition

Adapting Biggs' above definition of a self-fulfilling prophecy, a self-*defeating* prophecy can be defined as a dynamic process between two actors with two criteria. The first criterion is a causal sequence of the following form ($\neg c$ is a characteristic of an actor that is the opposite of c):

(1) A_1 believes that "A_2 is c."
(2) A_1 therefore does b.
(3) Because of (2), A_2 becomes $\neg c$.

The second criterion is that one or both of the actors *believes* (falsely) that the causal sequence was actually the following:

(1) A_2 is $\neg c$.
(2) In spite of (1), A_1 believes that "A_2 is c."
(3) A_1 therefore does b.
(4) Because of (1), A_2 is still $\neg c$.

In this study (as will be explained in detail below), A_1 is the group of voters, A_2 is a candidate, b is the action of voting or not voting for a candidate (which can be different for each voter), and c is "going to win the election."

Proposed Experiment

Summary

A simple model of a binary, democratic election will be used to investigate whether an election prediction can be a self-defeating prophecy and, if so, under what conditions.

Participants

The study will be conducted online to minimize cost and allow for easy data collection. 100 participants will be recruited via Mechanical Turk. Ideally, these participants would form a representative sample of the population of interest, e.g. the United States. For this preliminary experiment, however, any selection of 100 people is permissible[2]. They will participate in the study as described below, for a total of 30 minutes each. They will receive minimum wage, so the total labor cost of the study (at the California rate of $10.50 per hour) will be $525.

Procedures

An election, as modeled here, is a contest between Candidate A and Candidate B. There are N voters, each of whom can either vote for one of the candidates (which costs them v points) or choose not to vote (which costs them zero points). Each voter receives a reward of r points if Candidate A wins and $-r$ points if Candidate B wins. The value of r is different for and randomly assigned to each voter, following a normal distribution over the range [100, -100]. Each voter is presented with these three pieces of information (N, v, and r) as well as p, the stated probability[3] that Candidate A will win. In this study, p is the same for all voters (but see the discussion section for future research ideas). Each voter has 10 seconds to make their decision. They participate in 180 consecutive elections[4] (taking a total of 30 minutes), each of which immediately follows the last, and they only learn the results of the elections at the end of their participation. They are instructed to maximize their total points over all of the elections, motivated by the opportunity to win one of ten $10 Amazon gift cards. (Note that this adds $100 to the total cost of the study.)

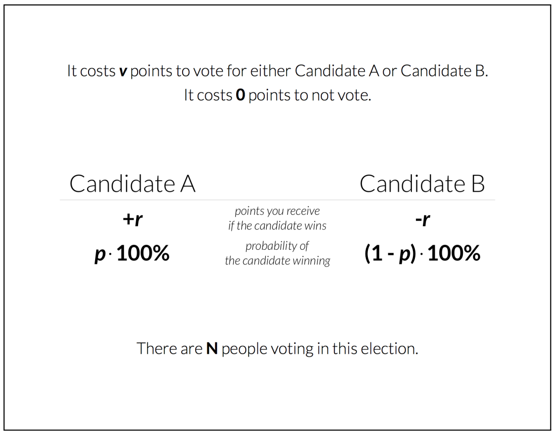
These design choices ensure several important things: (1) that the participants do not have time to think too analytically or develop a decision algorithm (which would be quite unlike a real election situation, where the benefits and costs of the outcome are largely unknown), (2) that a large number of data are collected, (3) that the time commitment is low enough to attract many participants, (4) that the results of earlier elections do not affect the outcomes of later ones, and (5) that participants are incentivized to maximize their points.

The two images below show what voters will see during elections. The first shows a general screen with the variables left unspecified. The second shows a particular example with $v = 5$, $r = 25$, $p = 0.4$, and $N = 100$. This example screen of course shows what *one* of the 100 voters in that election would see (a different voter would see the same screen but with a different value of r). In a separate part of the screen, each voter will see a ticking clock showing the remaining time and buttons to vote for A, vote for B, and not vote. Running out of time is equivalent to clicking the button for not voting.

[2] There is no obvious reason to expect demographics to influence how people behave in this very quantitative experiment. However, future research should examine this scientifically.

[3] Note that this is a *stated* probability because it's not an *actual* probability but rather a variable manipulated in the experiment. In other words, p in no way indicates the true probability of A winning, but it is presented to voters as though that is the case.

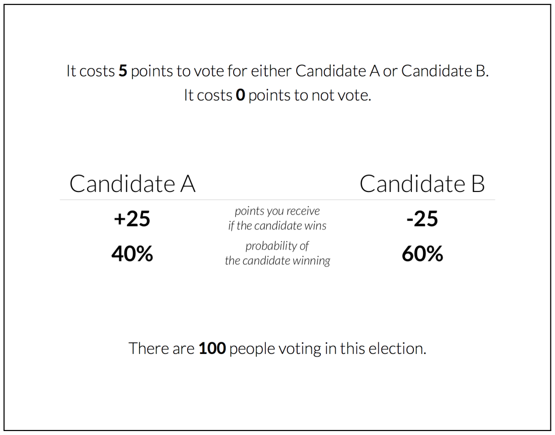
[4] Each voter participates in the same 180 elections but in randomized order.



It costs **v** points to vote for either Candidate A or Candidate B.
It costs **0** points to not vote.

Candidate A		Candidate B
+r	*points you receive if the candidate wins*	**-r**
p · 100%	*probability of the candidate winning*	**(1 - p) · 100%**

There are **N** people voting in this election.



It costs **5** points to vote for either Candidate A or Candidate B.
It costs **0** points to not vote.

Candidate A		Candidate B
+25	*points you receive if the candidate wins*	**-25**
40%	*probability of the candidate winning*	**60%**

There are **100** people voting in this election.

Measures

In the election model considered here, there are three variables: N, v, and p. In this study, N and v will be fixed: $N = 100$ since that is the number of participants (and every participant is a voter in every election, as noted above), and $v = 5$ by choice. Given these conditions[5], the proposed study examines the relationship between the stated probability of the election outcome (the independent variable) and voting behavior (the dependent variable). Voting behavior is comprised of the numbers of votes for each candidate and non-votes. The outcome of the election is derived from this voting behavior: whichever candidate has the most votes wins and, if each candidate has the same number of votes, each one wins half of the time. In this way, the proposed experiment will examine whether p affects the election outcome.

Theory

The election model defined here can be used to develop a theory of election dynamics, assuming voters behave rationally. This theory provides a helpful foundation for hypotheses about how real people might behave.

Consider a single voter, and suppose they are a rational actor who believes the probability p represents the true probability of A winning[6]. A rational actor knows that their vote changes p by some (usually small) amount ϵ by the laws of conditional probability. Their expected value of points won in the election is

$$E = \begin{cases} rp - r(1-p) & \text{[not voting]} \\ r(p+\epsilon) - r(1-p-\epsilon) - v & \text{[voting for A]} \\ r(p-\epsilon) - r(1-p+\epsilon) - v & \text{[voting for B]} \end{cases}$$

Therefore, they will choose to vote for A over not voting if

$$\begin{aligned} r(p+\epsilon) - r(1-p-\epsilon) - v &> rp - r(1-p) \\ \implies v &< 2r\epsilon \end{aligned}$$

Likewise, they will choose to vote for B over not voting if

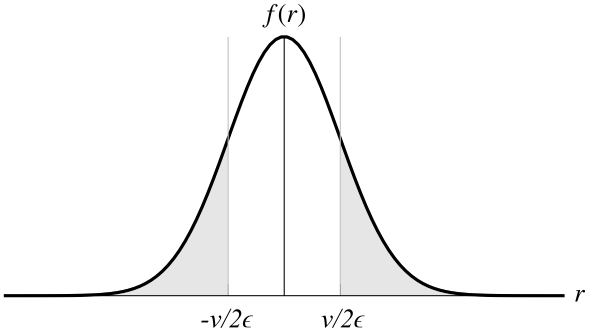
$$\begin{aligned} r(p-\epsilon) - r(1-p+\epsilon) - v &> rp - r(1-p) \\ \implies v &< -2r\epsilon \end{aligned}$$

Obviously, if they choose to vote, they will vote for A if $r > 0$, B if $r < 0$, and each half of the time if $r = 0$. Therefore, the general condition for voting over not voting is

$$v < 2|r|\epsilon.$$

This equation indicates that, if all voters are rational actors, the value of ϵ has no impact on the result of the election. This is because v is fixed and r is normally distributed: the areas under $r > v/2\epsilon$ and $-r > v/2\epsilon$ are equal, regardless of the value of ϵ. In other words, if ϵ changes the number of votes for A, it changes the number of votes for B by an exactly opposite amount. Likewise, this equation makes it clear that the election result will not change if r follows *any* Gaussian distribution, not just a normal one with $\mu = 0$ and $\sigma = 1$.

It is easiest to understand these results graphically. The figure below shows a normal distribution of r in the range [-5, 5]: $f(r)$ is the frequency of r.

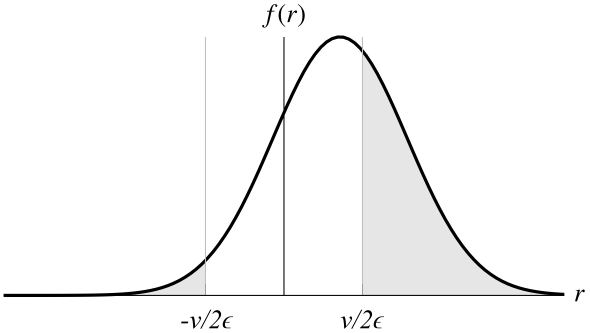


[5] These values for N and v are chosen to be reasonable but are strictly arbitrary. Future research should vary each over some appropriate range, as the discussion section notes.

[6] The experiment presents p as the actual probability of A winning, so it is rational to believe this even though it isn't true.

The shaded areas on the left and right show the number of votes for A and A, respectively. As ϵ increases, the boundaries of these areas move closer to zero but by the same amount. So, by the symmetry of the distribution, neither A nor B receives relatively more votes. Similarly, as ϵ decreases, the boundaries move further from zero but by the same amount, and neither A nor B receives relatively fewer votes. (As ϵ goes to zero, the boundaries move to $-\infty$ and ∞, and no one votes for either A or B. This makes sense: when there is a nonzero cost of voting, it is irrational to vote if that vote has no effect on the outcome of the election.) In each case, the vote differential is the same (in this case, zero), and the outcome of the election is unchanged.

Now, consider the plot below, which shows a non-normal Gaussian distribution for r (with arbitrary $\mu \neq 0$ and $\sigma \neq 1$, also in the range [-5, 5])[7].



As before, as ϵ increases, the boundaries of the areas move closer to zero by the same amount. In this case, however, A receives more additional votes than B. Similarly, as ϵ decreases, the boundaries of the areas move further from zero by the same amount, and A loses more votes than B. Regardless of the value of ϵ, however, A will still win the election, so the outcome is unchanged. Thus, provided that r follows a Gaussian distribution and all voters are rational actors, ϵ has no impact on the election outcome. Note that this is true even if ϵ depends on p and/or N.

Hypotheses and Analysis

Real people, of course, are not rational actors. It is therefore important to consider how human psychology might affect voting behavior. Loss aversion, for example, may cause the experimental results to diverge from the theoretical ones (i.e. those for rational actors). Specifically, a voter may negatively value $r < 0$ points more than they positively value $-r > 0$ points, which would increase the total number of votes. However, the increase in the number of votes for each candidate would be the same, and so the election outcome would not change. Another possibility is that framing effects will influence voting behavior. For example, a voter may prefer a definite loss of v points (the cost of voting) to a 50% chance of losing $r = 2v$ points, even though the expected value is the same. This would also increase the total number of votes but by the same amount for each candidate, therefore not changing the election outcome.

These considerations, based on well documented psychology, suggest a straightforward hypothesis: the total number of votes for each candidate will be greater (by the same amount) than the number predicted by the theory developed for rational voters. Furthermore, election outcomes will match what the theory predicts: over many elections, A and B will win the same number of elections, regardless of the value of p.

Discussion

The experiment proposed here has a few notable weaknesses. Because the election model has three variables, it is necessary to collect a large number of data to obtain meaningful overall results. By the same token, even though the more restricted investigation suggested here (which aims to vary just one variable) requires fewer data, it is correspondingly less meaningful. In addition, the election model cannot hope to capture a huge range of factors that affect real elections and are likely difficult if not impossible to model. Any results from the experiment must therefore be compared to real election dynamics only with considerable caution. Finally, the experiment may indeed show voting behavior that deviates from what the theory for rational actors predicts, but there may be no way to understand *why* that happened. The experiment is in this sense better suited to investigate *whether* and *in what way* human psychology impacts elections but not the *mechanisms* by which that occurs.

Despite these weaknesses, the experiment has many strengths. It uses an election model that is simple enough to be highly usable but complex enough to capture several election dynamics at once. It also has relatively low cost because it is conducted online and requires little time from each participant. Perhaps the experiment's greatest strength is that it paves the way for a wide range of further research.

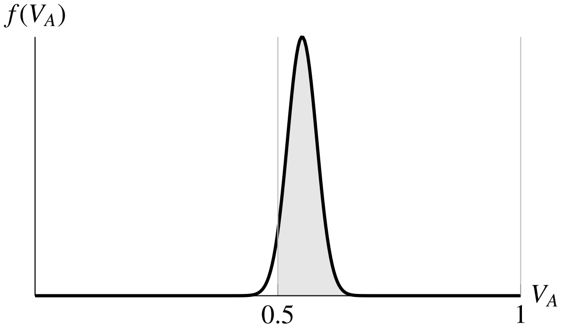
Future studies could, for example, vary N and/or v instead of or in addition to p. This might allow for very exact specification for the conditions under which predictions affect outcomes. Other studies could explore how the results change if p is not the same for all participants. This is an intriguing line of inquiry motivated by real-world election dynamics: in the 2016 U.S. presidential election, for example, it could be argued that Hillary Clinton supporters broadly believed she had a higher chance of winning than Donald Trump supporters did. Perhaps such a difference in expectations affected voting behavior and possibly even the outcome.

[7] For clarity and full generality, the vertical lines are here plotted for a different (arbitrary) value of ϵ.

Appendix

Finding the exact value of ϵ is a difficult task, but it is manageable to find a close estimate. Since the election model of this paper is binary, the distribution of votes follows a binomial distribution, which can be approximated by a Gaussian distribution. Such an approximation is quite accurate provided that the number of votes is large. In the experiment proposed here, the number is not especially large (less than or equal to N), but the election model is motivated by real-world democratic elections where $N \gg 1$ and the number of votes is an appreciable fraction of N. Therefore, for any such application, a Gaussian approximation is sufficient.

Assuming that the distribution has $\sigma = 1/V$ for number of votes V, the probability p can be found as a function of the mean μ. To see this, consider the figure below, which plots V_A, the fraction of votes for A, versus $f(V_A)$, the frequency of those fractions.



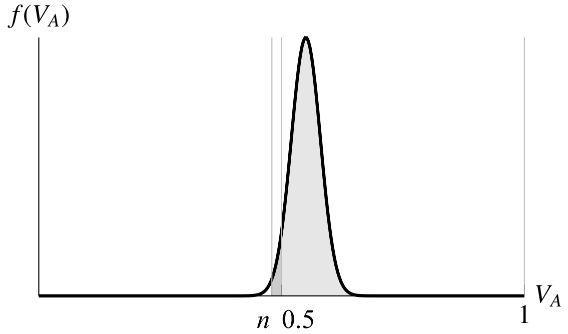
The shaded area is the value of p (the probability of A winning) because A wins whenever the fraction is greater than 0.5 (and half of the time when the fraction equals 0.5).

To find the value of ϵ, it is necessary to consider how the value of p changes when one vote is fixed. For large V, fixing one vote changes σ and μ by a negligible amount, so the distribution is approximately unchanged. However, the new value of p will no longer be the area under the curve above 0.5 but rather above

$$n = \frac{\frac{V}{2} - 1}{V - 2}.$$

This is easiest to see by example. Suppose $V = 50$. Then, A needs at least 25 votes to win, so the value of p must be the area under the curve above 25/50 = 0.5. A rational voter would know this, but they would also know to *update* their value of p based on their choice of candidate (if they decide to vote). This is a matter of conditional probability: the rational voter considers the value of p conditional on their vote. Without loss of generality, consider the case where they choose to vote for A. Then, A no longer needs at least 25/50 votes to win but rather only 24/49, consistent with the formula above. The formula above is true by inspection, and it makes intuitive sense: as V increases, n approaches 0.5.

Therefore, the *updated* value of p is the area under the curve above n. Assuming, as above, that the distribution is approximately the same (i.e. σ and μ are unchanged), this indicates that ϵ is the change in area. This is shown below as the slightly darker shaded region in the plot below.



Mathematically,

$$\begin{aligned} \epsilon &= \frac{V}{\sqrt{2\pi}} \int_n^{0.5} e^{-(V_A-\mu)^2/2V^2} \mathrm{d}V_A \\ &= K\left(\operatorname{erf}\left[\frac{\mu - n}{\sqrt{2}(1 + \log_2 V^2)}\right] - \operatorname{erf}\left[\frac{\mu - 0.5}{\sqrt{2}(1 + \log_2 V^2)}\right]\right) \end{aligned}$$

for $K = V/[2(1 + \log_2 V^2)]$. Since μ depends on p, it is clear that ϵ depends on p and V as we would intuitively expect, but it is by no means a simple relationship.

Even so, it is of interest to consider a simple approximation of ϵ. A simple dependence of ϵ on p gives a way to calculate expected voter turnout theoretically and may prove useful in other contexts. A reasonable guess is

$$\epsilon = \frac{1}{pV}.$$

This makes intuitive sense: the larger p or V is, the less each vote changes the probability of the election outcome because each vote has proportionally less impact. Voter turnout, as a function of p is then the area under $|r| > v/2\epsilon = vpV/2$ divided the total area under the curve r.

References

Biggs, M. (2009). Chapter 13: Self-Fulfilling Prophecies. Oxford Handbook of Analytical Sociology, 294-314.

Brookover, W. B., Rosenthal, R., Jacobson, L. (1969). Pygmalion in the Classroom: Teacher Expectation and Pupils Intellectual Development. American Sociological Review, 34(2), 283.

Eden, D. (1990). Pygmalion in Management: Productivity as a Self-Fulfilling Prophecy. (Lexington, Mass.: Lexington).

Merton, R. K. (1948). The Self-Fulfilling Prophecy. The Antioch Review, 8(2), 193-210.

Salganik, M. J. and Watts, D. J. (2007). An Experimental Approach to Self-Fulfilling Prophecies in Cultural Markets. Paper presented to the 2006 ASA.