Testing the Power of Arguments in Referendums
A Bradley-Terry Approach

Peter John Loewen* Daniel Rubenson† Arthur Spirling‡

July 26, 2011

Abstract

How can we determine which arguments in a referendum are most persuasive? We show that the Bradley-Terry model has several features that make it well-suited to this task, and thus preferable to other, more conventional approaches. Using a survey experiment conducted during an electoral reform referendum in Ontario, Canada in October 2007, we demonstrate how unstructured and structured Bradley-Terry models can be straightforwardly fitted and interpreted. In doing so, we gain insight into the factors which determine support for electoral reform. We identify a status quo bias and find that power varies with mention of fairness, local control over candidate selection, and the role of political parties. We conclude by discussing the limits, extensions and further applications of such models in electoral studies and political science more broadly.

*University of Toronto. peter.loewen@utoronto.ca
†Ryerson University. rubenson@ryerson.ca
‡Harvard University. aspirling@gov.harvard.edu

1 Introduction

Scholars of politics have long been centrally concerned with assessing the persuasive power of arguments (see, e.g., Aristotle (322BC/1991) for an ancient treatment and e.g. Mutz, Sniderman, and Brody (1996) for a more recent discussion). For example, does a call for greater social welfare spending which appeals to an individual's sense of equality elicit greater support than an appeal based on self-interest? Or, in a more oppositional manner, if a liberal argument for higher taxes is pitted against a conservative argument for lower taxation, which garners more support? By exposing respondents to various arguments in a controlled setting, we can learn much about the foundations of preferences, as well as rhetorical and political success in real world arenas.

Elections are perhaps the most obvious arena in which issues and arguments compete (see, e.g., Page 1978; Erikson and Wright 1997; Ansolabehere, Snyder, and Stewart 2001). However, in this case, issues and arguments are often 'bundled' together such that it becomes difficult to determine the effects of such messages independent of candidates. A more systematic discussion of positions is seen in referendums—formal plebiscites in which voters directly convey their preferences on a binary choice that may concern anything from a new policy commitment to potentially profound constitutional changes. Since so much is at stake in such national reform votes, political scientists have an obvious interest in attempting to understand *why* citizens choose the way they do. In some senses, this exercise ought to be empirically easier for referendums since, relative to elections, such votes are depersonalized: the candidate is an inanimate proposition, rather than an individual (or collection of individuals) (Pattie, Denver, Mitchell, and Bochel 1998; Tranter 2003). Voters thus receive more focused and explicit discussions of the issue at stake, with party and other (partisan) loyalties sometimes at cross-cutting angles.[1] As a result, scholars have direct access to the explicit arguments and types of reasoning used to buttress a position, rather than having to piece together the more rhetorical, nebulous and eclectic sources

[1]Consider, for example, the UK's EEC referendum vote of 1975: Prime Minister Wilson and most of his cabinet united with the majority of Margaret Thatcher's Conservatives to recommend a 'yes' vote. Meanwhile, a left-wing contingent of Labour MPs—and some right-wing Tories—campaigned for a 'no'.

of appeal that a particular party or politician uses. Using such political arguments and voter responses—observationally or experimentally—analysts ought to be able to teach us much about the outcomes of these events.

We present a method to measure the power of such arguments. By 'power', we are referring to the ability of an argument to elicit agreement with its advocated position on an issue *in the face of* an argument from the 'other side'. Such a definition implies the arguments exist in opposition to one another, and that an argument's power depends not on its ability to defeat another single, particular argument, but its ability to best a range of arguments making the case for an alternative position. Unfortunately, analysts of referendums face pronounced methodological problems when attempting to assess the persuasive 'power' of different political arguments. First, in observational studies, it is not always apparent precisely which arguments voters have received. Precisely measuring probable exposure is difficult and asking voters to self-report their exposure to arguments is problematic. Another solution is to take an experimental approach within a survey, and have subjects receive one of k total contrasting arguments, followed by a report on which argument they agreed with. For increasing values of k, however, the size of the relevant sample on which inferences are based is quickly reduced. Moreover, exposing each subject to a large number of arguments and then soliciting opinions on which is 'most convincing' risks a learning or cumulative effect where responses are a function of previous and current treatments in the series, rather than the content of the argument *per se*.

These two problems are exacerbated if we wish to pit particular arguments against one another in order to mimic the way debates take place in practice. Consider the case where we have m arguments in favour of a position, and the same number against. Randomly assigning a total of T respondents to receive one pro argument and one con argument yields just $n = \frac{T}{m^2}$ subjects per treatment. For a sample of, say, 500 individuals picking between, say, 6 pairs of arguments, $n < 15$: that is, there are less than 15 respondents per condition! With such small numbers, we

would expect low statistical power to be an issue when attempting to draw conclusions about the relative ability of arguments. And, of course, assigning multiple treatments risks the learning/fatigue effect noted above.

We suggest an innovative solution to this problem: a novel application of the Bradley-Terry model of pairwise comparisons (Bradley and Terry 1952). This model has a unique value proposition. It can give survey researchers a comparatively large amount of information in a small amount of survey space. By way of example, we use a survey experiment of just 520 respondents in which respondents are assigned one of six arguments for and six against electoral reform (creating 36 treatment groups). By conceptualizing arguments as contestants engaged in bouts with other arguments, we can use the model to estimate the probability that an argument wins. That is, we wish to compare the 'abilities of players'—in our case, the 'power of arguments'—in order to assess whether and why some arguments are better able to persuade (or 'win') when pitted against an opposing argument.

By applying the Bradley-Terry model, we can rank-order arguments according to their power to persuade. Moreover, we can also determine the sources of an argument's power, i.e. the components that make it more persuasive than another argument. We argue that the utility of this method exceeds more traditional approaches. Our principal motivation, then, is to introduce these models to election researchers. A secondary motivation is to introduce these models to political scientists more generally. As we show in the final section of our paper, these models can be applied to a wide range of substantive problems which are characterized by contests between individual actors.

Substantively, we are motivated by the question of which arguments for and against electoral reform are most powerful. To this end, we conducted a survey experiment in the midst of a province-wide referendum on whether to change the electoral system in Ontario, Canada. This

new data consists of respondents randomly assigned to receive one of six arguments in favor of the existing First Past the Post (FPTP) electoral system and one of six arguments in favor of the proposed Mixed Member Proportional (MMP) system. By knowing which arguments increase the appeal of one electoral system over another and by knowing the sources of arguments' power we can gain important insight into support or opposition to electoral reform. By understanding *why* certain arguments are more powerful, we also contribute to more general knowledge in electoral studies and political science more broadly. In doing so, we also take a route distinct from earlier studies of referendums that focus on voters rather than arguments (e.g. Pattie, Denver, Mitchell, and Bochel 1998; Tranter 2003; Pammett and LeDuc 2001; Hobolt 2009). Such an approach can thus compliment work that focusses on the determinants of individuals' decisions.

We proceed as follows. In the next section we describe our experimental design and data. In Section 3 we introduce our novel statistical approach for pairwise comparison, the Bradley-Terry model. In Sections 4 and 5 we describe our results. We conclude by suggesting other questions to which this model might be applied in political science.

2 Data and Experimental Design

Our experiment occurred within the context of a referendum on electoral reform conducted in the province of Ontario, Canada, in October 2007. The referendum was conducted concurrently with a provincial election. The Ontario MMP proposal lost the ensuing referendum, garnering just 38% of the vote and a majority in only five districts. It was not a loss directly attributable to the proposing government, as they won the concurrent election rather overwhelmingly. Thus, this loss was much to the consternation of electoral reform advocates who believed that the arguments for electoral reform were clearly superior to those in support of the current system. Indeed, in the days following the referendum, advocates claimed that had the public been better educated on the proposed system, especially its underlying values, support would have been higher, perhaps

surpassing the 60% threshold required for reform (Fenlon 2007).

To put claims that the arguments for the MMP system are more powerful upon exposure than those for the FPTP system to the test, we developed six arguments in favor of each position. These are drawn from campaign materials produced by both sides of the campaign, conversations with advocates on both sides and academic literature on electoral systems (e.g. Blais and Massicotte 2002; Powell 2000).

To assess the comparative power of arguments, we conducted an online survey experiment with 520 voting-aged Ontarians in the last week of the referendum campaign. Our experimental module was contained at the end of the survey, after three sections related to federal politics, provincial politics and environmental issues. No questions prior to the experimental module were related to electoral reform. A profile of our subjects is found in Appendix A.

In the experiment, subjects read: "As you may know, there will also be a referendum during the October 10th election. The purpose of the referendum is to determine whether Ontario should change its electoral system from the current first past the post system to a mixed member proportional system. We'd like to present you with an argument for each system and then get your view on which system you prefer." Respondents were then presented with one of six arguments in favor of the existing system and one of six arguments in favor of the proposed system. The order of the arguments was also randomized. After receiving the arguments, respondents were then asked to indicate their preference between MMP and FPTP; they were not able to give a 'Don't know' response and had to indicate a preference before proceeding to the end of the survey. Notice that each respondent is 'used' only *once* in the study: they receive one MMP argument, and one FPTP argument.

The six arguments for FPTP and MMP were:

FPTP

- A first past the post system is better because it creates strong majority governments that can implement their policies (FPTP1).
- A first past the post system is better because it makes sure that every Member of Provincial Parliament is elected from a constituency (FPTP2).
- A first past the post system is better because one ballot is less confusing than two. (FPTP3)
- A first past the post system is better because it allows local party members to choose all of a party's candidates. (FPTP4)
- A mixed member proportional system is worse because it will lead to unstable coalition governments. (FPTP5)
- A mixed member proportional system is worse because it puts too much control in the hands of parties and political elites. (FPTP6)

MMP

- A mixed member proportional system is better because it makes sure that parties that have support of some of the population still get some representation. (MMP1)
- A mixed member proportional system is better because it lets voters indicate their preference for both a local representative and a party. (MMP2)
- A mixed member proportional system is better because parties should get the same share of seats as their share of the vote. (MMP3)
- A mixed member proportional system is better because it will lead to more diversity in the legislature. (MMP4)
- A first past the post system is worse because it gives some parties a share of the seats much larger than their share of the vote. (MMP5)
- A first past the post system is worse because it shuts out small parties. (MMP6)

As the pro and con arguments were assigned randomly and independently, we have 36 approximately equal sized groups.[2]

We are not only interested in the power of arguments, but also their sources of power. In other words, what is it about the content of arguments that makes some more powerful than oth-

[2] A χ^2 test indicates that the assignments are independent of one another ($\chi^2 = 26.41$ $p = .39$).

ers? We note that several arguments share common components. For example, some arguments reference local candidates, others appeal to notions of proportionality or fairness, others reference government stability and effectiveness, others highlight the strengths or weaknesses of the ballot structure in FPTP and MMP. Many make reference to political parties. Some arguments are phrased in positive terms (i.e. a first past the post system is *better* because...), while others are phrased negatively (i.e. a mixed member proportional system is *worse* because...). And some arguments combine many of these features. Table 1 summarizes the characteristics of the arguments. We are thus not only interested in which arguments are more powerful, but also what components make one argument more persuasive than another. Accordingly, we use a structured Bradley-Terry model (in Section 5) to attempt to identify the most persuasive arguments and the components which make them persuasive.

3 Measuring the Power to Persuade: A Statistical Model

We contend that arguments vary in their 'power to persuade' citizens who receive them. The central econometric concern is to *estimate* this persuasion 'power', possibly as a function of the characteristics of the arguments themselves. We begin by supposing that in any two-way comparison in which an argument pertaining to FPTP is paired with one pertaining to MMP, one argument makes a more convincing case than the other for its respective electoral system; it is thus more likely to persuade the respondent. For any given comparison or 'contest' between arguments—or 'players'—i and j let $\pi_{i,j} \in (0,1)$ be the probability that the respondent finds i more compelling than j and thus prefers the electoral system (implicitly or explicitly) promoted by i. Write the odds that argument i 'beats' argument j as a function of their 'powers' α_i, α_j which are latent (and thus unobserved) but positive valued:

$$\frac{\pi_{i,j}}{\pi_{j,i}} = \frac{\pi_{i,j}}{1-\pi_{i,j}} = \frac{\alpha_i}{\alpha_j} \;\; \forall i, \forall j. \tag{1}$$

Table 1: Argument Characteristics

Argument	Fairness/ Proportionality	Local Members/ Control	Government Formation/ Stability/Effectiveness	Ballot Structure	FPTP	Political Parties	Negative
FPTP1			X		X		
FPTP2		X			X		
FPTP3				X	X		
FPTP4		X			X	X	
FPTP5			X		X		X
FPTP6					X	X	X
MMP1	X					X	
MMP2		X		X		X	
MMP3	X					X	
MMP4							
MMP5	X					X	X
MMP6	X					X	X

Implicit in Equation (1) is the fact that $\pi_{j,i} \equiv 1 - \pi_{i,j}$. In other words, there can be no tied contests: either i wins and j loses or j wins and i loses.[3] The task is to estimate α_i and α_j. Suppose that i and j compete against one another a total of $N_{i,j}$ times and let $n_{i,j}$ be the number of times i beats j. If all these contests are independent, then it is natural to assume $n_{i,j}$ is distributed as a binomial $(N_{i,j}, \pi_{i,j})$. With t *total* arguments competing—i.e. all the specific is and js that are actually compared in the survey—the likelihood is

$$L(\boldsymbol{\alpha}|\mathbf{n}) = \frac{\prod_i^t \alpha_i^{n_{i,j}}}{\prod_{i<j}(\alpha_i + \alpha_j)^{N_{i,j}}}. \tag{2}$$

Maximization of (2) yields estimates of the elements of $\boldsymbol{\alpha}$ subject to the identification restriction that some α_i is set equal to one.[4]

As implied by the form of Equation (1) and Equation (2),

$$\pi_{i,j} = \frac{\alpha_i}{\alpha_i + \alpha_j}. \tag{3}$$

The functional form for this 'dominance' relationship between the players i and j has particular transitivity properties. To see this, first denote the matrix of all such $\pi_{i,j}$ probabilities as $\boldsymbol{P}$. Now suppose one draws three players (a 'triplet') which we will write $\{i, j, k\}$. We say that $\boldsymbol{P}$ satisfies *strong stochastic transitivity* (SST) if when $\pi_{i,j} \geq \frac{1}{2}$ and $\pi_{j,k} \geq \frac{1}{2}$ then $\pi_{i,k} \geq \max(\pi_{i,j}, \pi_{j,i})$ (Suppes, Krantz, Luce, and Tversky 1990). Equation 3 implies that the corresponding $\boldsymbol{P}$ matrix meets SST. Specifically, if

$$\pi_{i,j} = \frac{\alpha_i}{\alpha_i + \alpha_j} \geq \frac{1}{2}$$

and

$$\pi_{j,k} = \frac{\alpha_j}{\alpha_j + \alpha_k} \geq \frac{1}{2},$$

[3]Effectively, this means not allowing 'don't know' responses.

[4]Note (a) that Bayesian estimators exist, but we do not discuss them here; (b) an alternative identification restriction is to set $\sum_i \alpha_i = 1$.

then $\alpha_i \geq \alpha_j \geq \alpha_k$ and

$$\pi_{i,k} = \frac{\alpha_i}{\alpha_i + \alpha_k} \geq \frac{\alpha_i}{\alpha_i + \alpha_j} \geq \frac{\alpha_j}{\alpha_j + \alpha_k}$$

which is stronger than SST requires. The key point is simply that 'transitivity' is implicitly assumed when modeling the data generating process this way. A consequence of this assumption is that arguments that never meet one another can be *empirically* compared post-estimation.[5]

Returning to the estimation details, note that the $\boldsymbol{\alpha}$ may be estimated even if $n_{i,j}$ is zero for some pairings, so long as there does *not* exist some subset of arguments that never meet with others in competition.[6] A logit linear form of the problem proceeds by defining $\exp(\lambda_i) \equiv \alpha_i$ and executing the maximization above. In this case, with rearrangement of Equation (1),

$$\pi_{i,j} = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}, \tag{4}$$

and, in terms of log-odds,

$$\log\left[\frac{\Pr(i \text{ beats } j)}{\Pr(j \text{ beats } i)}\right] = \lambda_i - \lambda_j. \tag{5}$$

The intuitive message from Equations (4) and (5) is clear: the larger the value of λ_i *relative* to λ_j, the more likely it is that argument i beats argument j in a pairwise contest.

Standard errors on player abilities can be obtained in the usual way: that is, via the square-root of the diagonal of the inverted Hessian matrix, although recall that one ability (of the 'reference' player) is *set* to zero and thus will be accompanied with no uncertainty estimate. The presence of the $n_{i,j}$ and $N_{i,j}$ terms in Equation (2) has a sensible consequence: players involved in more contests will have smaller variances around their estimated abilities. Furthermore, the form of Equation (2) and Equation (3) implies that if i beats a relatively powerful argument j,

[5] In principle, one could check every possible triple for transitivity violations empirically. This is difficult in the current case because a large number of the triples involve two sets of pairs that do not meet: e.g. i beats j, j beats k but i and k do not meet. This has no effect on the estimation *per se*, and only effects the verification of the modeling assumption.

[6] More technically, the design must be 'connected' though it need not be 'complete'.

ceteris paribus, player i will receive a larger boost to his estimated ability than if he defeats a weak opponent: for a fixed $\pi_{i,j}$, reducing α_j results in a smaller estimate of α_i. As a corollary to these two points, and the comment on transitivity, notice that the model does not necessarily require large numbers of subjects in any *particular* two player treatment group in order to obtain meaningful estimates of abilities in the tournament as a whole. Rather, information is leveraged from the full dataset of contests, and the relative performance of the arguments in each pairwise interaction involving them. Hence, the statistical power 'problem' that more conventional approaches such as means tests of support might encounter—and that we noted in the Introduction—is sidestepped.

The approach here may appear unusual to political scientists, but is well known to statisticians as the Bradley-Terry model[7] for pairwise comparison (Bradley and Terry 1952) and has been used by psychologists interested in subjects selecting items from choice sets for some time (see, for example, Luce 1959; Thurstone 1959).[8] The model has seen use in other fields such as biology (e.g. Stuart-Fox, Firth, Moussalli, and Whiting 2006), genetics (e.g. Sham and Curtis 1995), the investigation of journal citation patterns (e.g. Carter and Spirling 2008; Stigler 1994) and sports science (e.g. Agresti 2002). Spirling (2007) considers an application to the United States Senate, however we can find no other application of this model in political science.[9]

An important version of the Bradley-Terry model is provided by Springall (1973), in which 'player'-specific (here, argument-specific) variables $x_{i1}, \ldots, x_{ip}$ are used to predict the 'power' of the players (the $\boldsymbol{\lambda}$) directly. These independent variables enter the model via the linear predictor

[7]The model is sometimes called the Bradley-Terry-Luce model. We rest with simpler naming, as Luce's contributions to the model came some years after the model's original exposition by Bradley and Terry.

[8]In practice, Thurstone (1959) opts for a slightly different, probit form of the model. Here we will discuss the logistic regression case.

[9]Indeed, a search of all major political science journals on JSTOR resulted in only one hit; an article mentioning the Bradley-Terry model in an end note (Monroe 1995).

in the sense that

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}, \tag{6}$$

and interest focuses on the estimated coefficients $\beta_1, \ldots, \beta_p$. In the current context, these $\boldsymbol{\beta}$ inform the analyst as to the source, or 'cause', of the arguments' power. A more complicated form of Equation (6) includes a per player (here, argument) random effect such that $\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir} + U_i$, where $U_i \sim \mathcal{N}(0, \sigma^2)$. Below, we fit both structured versions, though in practice the variance of the errors is estimated as zero, and thus the forms are equivalent.

In practical applications, a general concern is that a judge—a person making a decision as to which of i and j wins—might not give independent responses from one 'row' of the data to the next. For example, we might worry that a person might is generally inclined towards a particular electoral system, before the survey begins, and that this 'bias' is propagated throughout all their decisions. In that case, one might consider a fixed effect v_r for each given respondent such that Equation (5) becomes

$$\log \left[\frac{\Pr(i \text{ beats } j)}{\Pr(j \text{ beats } i)} \right] = v_r + \lambda_i - \lambda_j.$$

This makes most sense when there are a moderate number of respondents making *multiple* decisions.[10] In our particular case however, we have as many respondents (520) as their are comparisons, since each surveyed individual is responsible for one (and only one) decision. As a result, fixed effects are not identified. Notice, in addition, that there in fact is no intercept term in the standard Bradley-Terry set up (McCullagh and Nelder 1999, 272). To see why, recall that Equation (5) has $\text{logit} p_{ij} = \lambda_i - \lambda_j$. Suppose that there were a constant, such that $\text{logit} p_{ij} = \alpha_0 + \lambda_i - \lambda_j$. Consider the complement of this quantity, $\text{logit} p_{ji}$, which simply involves switching the indices around; thus, $\text{logit} p_{ji} = \beta_0 + \lambda_j - \lambda_i$, where β_0 is a constant term, possibly taking a different value to that given by α_0. Since $\text{logit} p_{ij} = -\text{logit} p_{ji}$ by definition, we must have $\alpha_0 + \lambda_i - \lambda_j = -\beta_0 - \lambda_i + \lambda_j$, which clearly implies $\alpha_0 = -\beta_0$. Yet this equality requires

[10] Alternatively, if there are a large number of respondents, one might try a random effects route.

that $\alpha_0 = \beta_0 = 0$ and we have no intercept for our model.

In principle, different respondents may judge the arguments depending on their (differing) personal covariates. For example, older people may be more likely to respond positively to a FPTP argument since it endorses the long-standing status quo. In the current context, such effects are difficult to include formally. Generally speaking, to the extent that a particular contest-specific 'bias' (such as the sex or age of the respondent) occurs, it effects both players (here, the relevant arguments) equally. While it *is* possible to use binary 'order effects' where we have an *a priori* reason to believe that one player is (dis)advantaged in some way, in the sense of Agresti (2002), it is not obvious how to include more complicated covariates at the respondent level.

Turner and Firth (2010) devote considerable effort to designing software for the fitting of both 'unstructured' models of the form given in Equation (5), and 'structured' versions as noted in Equation (6). Our work below utilizes this implementation—`BradleyTerry2`—in conjunction with the `R` language and environment (R Development Core Team 2006).

4 The Power of Arguments: The Bradley-Terry Method

Recall that the initial goal is to report a rank-ordering of arguments according to their ability to convince voters. Two approaches may seem more obvious at this point. First, researchers could rank the arguments according to the proportion of contests in which argument i beat the alternatives it was drawn against. As an empirical example, FPTP2 is preferred over its opponent (whatever that may be) 58 times out of 99 times it is paired in the survey. Hence, it wins approximately 59% of its contests. By contrast, MMP6 wins 43 of its 84 contests, and thus its 'mean' support might be reported as 51%. There are two interrelated problems with this procedure. First, there is uncertainty in the estimates which is not reported. That is, these means should come with standard errors that reflect sampling variation. This point is impor-

tant because though the contests are plausibly independent, this was not strictly a 'round-robin' tournament in which all players met a set number of times: for example, MMP2 met FPTP2 on 19 occasions, while MMP5 met FPTP1 just 15 times. However, the arguments are of different abilities: stacking up wins against a *relatively* strong opponent should presumably count more towards the ranking than drubbing a very weak one. Yet by just comparing proportions of victories that subtlety is lost.

An alternative approach might be to estimate a logistic regression wherein the dependent variable is support for, say, the MMP option and the various FPTP options are dummied out on the right hand side. One could then flip the operation, instead looking at FPTP victories predicted by MMP argument dummies. While uncertainty would (as usual with such approaches) be estimated along with the coefficients, the problem with this strategy is that it moves away from the core quantity of interest: the specific abilities of the arguments relative to one another, such that arguments that did not meet—and did not meet the same sets of opponents in the same numbers of contests—can be sensibly compared.

4.1 Unstructured Model Results

As noted above, the Bradley-Terry model will facilitate direct estimation of abilities, allow for uncertainty statements and incorporate information on the relative abilities of contestants in terms of the estimated 'power' of those that win and lose against them. To show this we begin with the 'unstructured' model results in Table 2.

Unstructured model results consist of a power coefficient for each argument, λ_i, which can be used to compare its power with that of other arguments. The reported coefficient for FPTP2, the most powerful argument, is set to 0. All other coefficients (and their p-values), are reported in comparison to the power of this FPTP argument. The arguments are ordered from most to least powerful. By examining the coefficients we can determine which arguments are estimated as less

Argument	Power (λ)	p-value	Quasi-variance
FPTP2	0.000	—	0.04
FPTP1	-0.137	0.64	0.04
FPTP3	-0.151	0.63	0.05
MMP6	-0.197	0.50	0.05
FPTP4	-0.203	0.49	0.04
MMP3	-0.224	0.45	0.05
FPTP5	-0.278	0.34	0.04
MMP5	-0.302	0.30	0.05
MMP1	-0.343	0.23	0.05
MMP2	-0.422	0.12	0.04
MMP4	-0.593	0.04	0.04
FPTP6	-0.605	0.05	0.05

Table 2: Bradley-Terry Model of Argument Power

powerful than others. Moreover, by examining the p-values, we can determine which arguments are significantly less powerful than FPTP2 (i.e. MMP4 and FPTP6). Though we do not report it here, with such a rank ordering the Bradley-Terry model does not differ much from a mean-support type approach outlined above and the order of the arguments appears approximately the same. But, of course, this approach does offer the distinct advantage that we can easily compute the estimated probability that one argument dominates another. Recall Equation (4); if we wanted to know the probability that FPTP2 defeated MMP2, we enter the power estimates into the equation and find that:

$$\pi_{2,2} = \frac{\exp(0.00)}{\exp(0.00) + \exp(-0.422)} = 0.60. \tag{7}$$

Similarly, we could calculate the probability that FPTP2 beats the more evenly matched MMP6 at 0.55. In Figure 1 we report our results in a slightly different way. Here, the black dots are the ability point estimates, while the bars around them are confidence intervals based on their 'quasi-variances' (in the sense of Firth and De Menezes 2004).[11] The quasi-variances reported

[11]Though the term may be unfamiliar to readers, there is a straightforward analogy. In the regression context, quasi-variances are used to approximate the standard error between estimated coefficients for a categorical variable such as education, which might be coded into, say, four mutually exclusive, and jointly exhaustive bins: 'less than high school', 'high school graduate', 'some college', 'post-graduate education'. Because all standard errors (and p-values) are reported relative to some 'base' or 'reference' category, one cannot—without the variance-covariance matrix—assess the significance of the difference between two categories that do not include the reference bin. But this is precisely the situation here: we have a base category (FPTP2), yet we wish to compare *other* arguments to

Figure 1: *Power of Arguments: quasi-standard errors (via quasi-variances) for Bradley-Terry abilities. Notice that the 'comparison intervals' overlap for all arguments.*

in Table 2 allow the analyst to compare arguments even though there is a base category which has (by construction) a standard error of zero. In particular, the t-statistic associated with a comparison between arguments is (well approximated as)

$$t = \frac{\hat{\lambda}_i - \hat{\lambda}_j}{\sqrt{qv(\hat{\lambda}_i + qv(\hat{\lambda}_j)}}$$

where $qv(\cdot)$ is the relevant quasi-variance. So, for the example of $FPTP_1$ and $FPTP_2$ we have $t = \frac{-0.13-0}{\sqrt{0.04+0.04}} = -0.47$ which has an associated p-value of 0.63 (see Gayle and Lambert 2007, for discussion of the calculations). Thus the difference is not statistically significant. The quasi-variances in Table 2 allow such a comparison for any such pair. Looking at Figure 1 it is clear that every interval overlaps every other, though in fact, there is a statistically significant difference between $FPTP1$ and both $FPTP6$ and $MMP4$. Still, most arguments are not distinguishable, and it not obvious that there is much of a difference between FPTP and MMP in terms of their general appeal.

5 The Sources of an Argument's Power

In the previous section, we estimated the power of twelve arguments for and against electoral reform. Our task now is to estimate the sources of this power. In other words, what are the characteristics of an argument that make it able to overcome other arguments—in our case, for or against electoral reform? In Table 1 we summarized the attributes of each of the FPTP and MMP arguments. By including these traits as covariates in a structured model we are able to estimate the sources of an argument's power—or 'ability'—to persuade.[12]

see if they are different from one another.

[12]This is Equation (6) in Section 3.

We began with an 'all main effects' model in which all the argument characteristics reported in Table 1 were coded as binary covariate vectors. Since the model is estimated via maximum likelihood, standard procedures for stepwise finding of the 'best-fitting' model are available via the Akaike (AIC) or Bayes Information Criterion (BIC). We use the AIC and the model thus selected is as reported in Table 3. As can be seen, we finish with four predictors. We take this advantage for some components over others as evidence that at least some of our sample was responding to arguments systematically, rather than merely stating prior opinions. With the exception of *Parties*, all the variables are significant at the 10% level.

Table 3: Structured Bradley-Terry model for referendum argument data

Feature	Estimate	Std. Error	p-value
Fairness	0.53	0.24	0.03
Local	0.32	0.18	0.07
FPTP	0.28	0.15	0.06
Parties	-0.27	0.18	0.13

All else equal then, an argument in favor of the status quo appears systematically advantaged. *Local* is a characteristic of arguments for both FPTP and MMP, as an MMP system does retain local representation. However, MMP does not allow the local selection of *all* candidates, suggesting that in actual rhetoric the scope of local arguments favors FPTP. *Fairness* arguments clearly advantage MMP, as proportionality is a central component of this system and is never present in arguments for FPTP. Finally, the mention of political parties makes an argument less appealing. This is likely to the disadvantage of MMP, as the central function of this system (and thus at least some of the arguments in its favor) is to match seats to votes for political parties, rather than individual candidates. Parties are thus arguably more central to vote choice in MMP than FPTP.[13]

[13] As an aside, we note that the issue of 'negative' framing is a popular research topic in political science (see, for example, Ansolabehere and Iyengar 1997). We found that 'Negative' was not a significant predictor (on its own) for the structured model, though it had a negative sign implying that voters are (all else equal) less convinced by such phrasing.

Figure 2: *Power of arguments: structured results. The arrows show the effect of moving from one covariate profile to another. On the left side of the plot, we report the predicted probabilities for the arguments in the study based on their covariate profiles The modal arguments are FPTP1, FPTP3, and FPTP5.*

Using this model—and assuming for purposes of exposition that it is the only one we estimated—and the formula in Equation (4) we can estimate the probability that one argument dominates another according to its characteristics.[14] For example, we can estimate the probability that FPTP2 (which mentions local control, and is in the FPTP group) dominates MMP6 (which makes a fairness and parties argument):

$$\begin{aligned} \pi_{FPTP2,MMP6} &= \frac{\exp(\sum \beta_r x_{ir})}{\exp(\sum \beta_r x_{ir}) + \exp(\sum \beta_r x_{jr})} \\ &= \frac{\exp(0.318 + 0.282)}{\exp(0.318 + 0.282) + \exp(0.534 - 0.275)} = 0.58. \end{aligned} \tag{8}$$

By contrast, if we pit FPTP2 against MMP4, an argument which has no significantly powerful attributes, the probability of FPTP dominance rises to 0.65.[15]

Figure 2 presents our findings in a slightly different way. The figure compares the probability of an argument with various covariate profiles winning against the modal argument (that is, an argument which does not mention 'fairness', 'local' or 'parties', but is in the 'FPTP' group). The arrows show the effect of moving from one covariate profile to another.

For example, in a contest with the modal argument, an argument that is identical but which

[14] Of course, (a) the arguments have characteristics the coefficients for which are not reported in this 'best fitting' model and (b) the coefficients in this reduced model will differ to those in the full model. Nonetheless, for purposes of example, we use the coefficients in Table 3 in what follows.

[15] An alternative, but entirely equivalent way to conceive of the predicted powers of the arguments is via Equation (5): that is, $\lambda_i - \lambda_j = \sum \beta_r (x_{ir} - x_{jr})$

mentions fairness would do better: the probability of a win for this second argument increases from 0.5 to around 0.64. By contrast, an argument that is identical but which does not mention FPTP would do worse: the probability of a win for this second argument decreases from 0.5 to around 0.45. On the left side of the plot, we report the predicted probabilities for the arguments in the study *based on their covariate profiles*. Reading from top to bottom, we see that FPTP2 is once again the most powerful argument, in keeping with our unstructured results.

5.1 Structured or Unstructured Results?

Given structured and unstructured results, the question remains about which researchers should prefer. We believe this is a matter of what questions a researcher wishes to answer. If we wanted to understand the power of exact arguments in the referendum, we would be well-served to consider the unstructured results, as these arguments closely match those made during the campaign. However, if we were looking forward to another campaign and wished to develop new arguments, we could learn more from the structured results. Indeed, these would allow us to design optimal arguments which combined effective components and avoided less effective ones. Fortunately, despite these models giving us different types of information on power, the predictions which result from them are very similar for our case. Table 4 reports the absolute difference in predicted probabilities of FPTP dominance according to our unstructured and structured (best fitting) models. The mean absolute difference in predicted probability is just two percentage points. Of course, the situation that the typical scholar faces might be very different and depends on the degree to which the covariate profiles do a 'good' job of explaining the variation in the data.

6 Further Applications and Conclusion

We have demonstrated that Bradley-Terry models can generate meaningful results from pairwise matchings of arguments, even when the number of treatment groups is large and the sample size small. Results of similar utility cannot be uncovered, for example, by more conventional

Table 4: Difference in predicted probabilities of FPTP dominance by structured and unstructured models

	FPTP Argument⇒					
MMP Argument⇓	1	2	3	4	5	6
1	0.05	0.00	0.04	0.02	0.01	0.00
2	0.01	0.03	0.01	0.02	0.02	0.04
3	0.02	0.03	0.01	0.01	0.02	0.03
4	0.04	0.00	0.04	0.02	0.01	0.00
5	0.04	0.01	0.03	0.01	0.00	0.01
6	0.01	0.04	0.01	0.02	0.03	0.04

models that do not utilize the contest nature of the data and the fact that individual players (the arguments) are involved in multiple battles. Moreover, structured models can be used to identify the source of an argument's persuasive power. These models are thus well-suited to a situation in which a researcher wishes to test a large number of arguments on a small sample size. More generally, these models are ideal for any question that features contest data.

Contest data is perhaps more common than we think. Indeed, we can identify a number of further applications within electoral studies and political science more broadly. For example, Bradley-Terry models could be deployed in experimental studies determining the attractiveness of candidates (e.g., Sigelman 1990). Likewise, such models could be deployed to determine which factors lead voters to prefer one leader, candidate, or party over another. More generally, we could measure the comparative power of frames over a small number of respondents without the need to repeat frames (e.g., Chong and Druckman 2007). The models could also be used to identify sources of persuasion in studies using a 'jury method' (e.g., London, Meldman, and Lanckton 1970).

We can likewise identify applications in international relations and American electoral politics. For example, the large and growing literature on 'democratic effectiveness' is concerned with the modeling of 'success' in war as a function of polity characteristics (e.g. Lake 1992; Reiter and

Stam III 1998; Downes 2009). Meanwhile, the voluminous discussion of 'incumbency bias' in US politics (e.g. Fiorina 1989; Ansolabehere and Snyder 2002) asserts that, for various reasons, contests between sitting officials and challengers are not 'fair fights'. Indeed, any research design with dichotomous outcomes and direct confrontation between units or actors is a candidate for these models.

Our suggested approach is not without drawbacks. As discussed here, it requires pairwise comparison and dichotomous outcomes. Accordingly, in the form expressed above, it cannot be used if researchers are interested in measuring degrees of support for a position. On the other hand, as written (with the random effect), it is in the form of a generalized linear mixed model, and so might be extended to normal-theory measurement situations. Certainly, if researchers wish to make pairwise comparisons with dichotomous outcomes, this approach provides a powerful solution when a number of actors (be they arguments or something else) need to be tested, especially if there are a small number of observations for each contest.

Given the large number of potential applications and the likely extensions of these models in the near future, we conclude that the Bradley-Terry approach is a helpful addition to the political methodologist's toolkit and we look forward to future research that utilizes it.

A Sample, Subject Profile and Generalization

Our subjects were drawn from an internet panel containing a list of approximately 15000 specifically named individuals. The panel is managed by a commercial polling firm. Though the panel is national, our experiment was provincial. At the time of the survey the panel had 3575 registered panelists in the province of Ontario. Respondents either self-select into the panel through the surveying firm's website, or they are recruited through solicitation by phone and email. Upon registration in the panel respondents are assigned a number between 1 and 31. In each subsequent month they are invited to complete the survey in one of four weeks according to this number.

Our experiment occurred in the last week of the referendum campaign (October 2 to 9, 2007) and was limited to those living in the province of Ontario. Of 3575 Ontarians registered in the panel, 844 were invited to complete the survey during this week. 565 agreed to complete the survey. Of those 565, 520 agreed to participate in the Bradley-Terry experiment. This gives us a response rate of 61.6%.

Table A-1 presents a profile of the experimental subjects.

Assignment to treatment condition is unrelated to educational category ($\chi^2 = 87.81$ $p = .08$), income category ($\chi^2 = 138.6$) $p = .52$), age category ($\chi^2 = 145.02$, $p = .95$) and gender ($\chi^2 = 26.29$ $p = .86$).

An obvious objection to our sample is that it is not a random sample of adult-aged Ontarians. This certainly appears reasonable; compared to the general population (Canadian census, 2001), our sample is both better educated ($\chi^2 = 11.64$, $p = .00$) and wealthier ($\chi^2 = 3.79$, $p = .05$). Moreover, previous analysis with the panel suggests that respondents are more likely to identify with a political party and appear highly interested in current affairs. The panel is likely highly sophisticated politically. As a consequence, our findings may not generalize to the whole popula-

Table A-1: Sample Demographics and Political Characteristics

Variable		% or Mean
Age	19–24	1.7%
	25–34	9.6%
	35–44	16.5%
	45–54	26.5%
	55–64	31.7%
	65+	13.9%
Female		52.6%
Household Income	<$40000	12.89%
	$40000 to $60000	16.15%
	$60000 to $80000	17.12%
	>$80000	38.65%
	Refused	15.19%
Education	High School or less	8.85%
	Some College	28.46%
	Some University	62.69%
N	520	

tion of Ontario, endangering external validity (Shadish, Cook, and Campbell 2000). Objections about generalization aside, nothing in our statistical model requires that we conduct our analysis over a random sample of the population. As such, the methodological approach remains fecund for more traditional randomly sampled surveys.

References

Agresti, A. (2002): *Categorical Data Analysis*. John Wiley and Sons, New York, 2 edn.

Ansolabehere, S., and S. Iyengar (1997): *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. The Free Press, New York.

Ansolabehere, S., and J. Snyder (2002): "The Incumbency Advantage in U.S. Elections: An Analysis of State and Federal Offices, 1942-2000," *Election Law Journal*, 1(3), 315–338.

Ansolabehere, S., J. M. Snyder, and C. Stewart (2001): "Candidate Positioning in U.S. House Elections," *American Journal of Political Science*, 45(1), 136–159.

Aristotle (322BC/1991): *Rhetoric*. Oxford University Press (translation by George A. Kennedy), New York.

Blais, A., and L. Massicotte (2002): "Electoral Systems," in *Comparing Democracies 2: New Challenges in the Study of Elections and Voting*, ed. by L. LeDuc, R. G. Niemi, and P. Norris. Sage, London.

Bradley, R., and M. Terry (1952): "Rank analysis of incomplete block designs, I. The method of paired comparisons.," *Biometrika*, 39, 324–345.

Carter, D., and A. Spirling (2008): "Under the Influence? Intellectual Exchange in Political Science," *PS: Political Science and Politics*, 41(2), 375–8.

Chong, D., and J. N. Druckman (2007): "Framing Public Opinion in Competitive Democracies," *American Political Science Review*, 101(4), 637–55.

Downes, A. (2009): "How Smart and Tough are Democracies? Reassessing Theories of Democratic Victory in War," *International Security*, 33(4), 9–51.

Erikson, R. S., and G. C. Wright (1997): "Voters, Candidates, and Issues in Congressional Elections," in *Congress Reconsidered*, ed. by L. C. Dodd, and B. I. Oppenheimer. Congressional Quarterly Press, Washington DC.

Fenlon, B. (2007): "Referendum on Electoral Reform an 'Unmitigated Disaster'," *The Globe and Mail*, October 11, 2007, Online edition.

Fiorina, M. (1989): *Congress: Keystone of the Washington Establishment*. Yale University Press, New Haven, 2 edn.

Firth, D., and R. De Menezes (2004): "Quasi-variances," *Biometrika*, 91(1), 65–80.

Gayle, V., and P. Lambert (2007): "Using Quasi-variances to Communicate Sociological Results from Statistical Models," *Sociology*, 41(6).

Hobolt, S. B. (2009): *Europe in Question: Referendums on European Integration*. OUP, OUPad.

Lake, D. (1992): "Powerful pacifists: Democratic states and war," *The American Political Science Review*, pp. 24–37.

London, H., P. J. Meldman, and A. V. C. Lanckton (1970): "The Jury Method: How the Persuader Persuades," *Public Opinion Quarterly*, 34, 171–83.

Luce, R. D. (1959): *Individual Choice Behaviour: A Theoretical Analysis*. J. Wiley, New York.

McCullagh, P., and J. Nelder (1999): *Generalized Linear Models*. Chapman & Hall, New York, 2 edn.

Monroe, B. L. (1995): "Fully Proportional Representation," *American Political Science Review*, 89(4), 925–40.

Mutz, D., P. M. Sniderman, and R. Brody (eds.) (1996): *Political Persuasion and Attitude Change*. University of Michigan Press, Ann Arbor.

Page, B. (1978): *Choices and Echoes in Presidential Elections*. University of Chicago Press, Chicago.

Pammett, J. H., and L. LeDuc (2001): "Sovereignty, leadership and voting in the Quebec referendums," *Electoral Studies*, 20, 265–280.

Pattie, C., D. Denver, J. Mitchell, and H. Bochel (1998): "Partisanship, national identity and constitutional preferences: an exploration of voting in the Scottish devolution referendum of 1997," *Electoral Studies*, 18, 305–322.

Powell, G. B. (2000): *Elections as Instruments of Democracy: Majoritarian and Proportional Visions*. Yale University Press, New Haven CT.

R Development Core Team (2006): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reiter, D., and A. Stam III (1998): "Democracy and battlefield military effectiveness," *Journal of Conflict Resolution*, pp. 259–277.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2000): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.

Sham, P., and D. Curtis (1995): "An Extended Transmission/Disequilibrium Test (TDT) for Multiallele Marker Loci," *Annals of Human Genetics*, 59(3), 323–36.

Sigelman, L. (1990): "Hair Loss and Electability: The Bald Truth," *Journal of Nonverbal Behavior*, 14(4), 269–83.

Spirling, A. (2007): "Power Tool: Measuring Power in Political Science—A New Methods with Application to the United States Senate," *University of Rochester*, Working Paper.

Springall, A. (1973): "Response surface fitting using a generalization of the Bradley-Terry paired comparisons model," *Applied Statistics*, 22, 59–68.

Stigler, S. (1994): "Citation Patterns in the Journals of Statistics and Probability," *Statistical Science*, 9, 94–108.

Stuart-Fox, D. M., D. Firth, A. Moussalli, and M. J. Whiting (2006): "Multiple Signals in Chameleon Contests: Designing and Analysing Animal Contests as a Tournament," *Animal Behaviour*, 71, 1263–1271.

Suppes, P., D. Krantz, D. Luce, and A. Tversky (1990): *Foundations of Measurement: Representation, Axiomatization, and Invariance*. Academic Press, New York.

Thurstone, L. (1959): *The Measurement of Values*. University of Chicago Press, Chicago.

Tranter, B. (2003): "The Australian constitutional referendum of 1999: evaluating explanations of republic voting," *Electoral Studies*, 22, 677–701.

Turner, H., and D. Firth (2010): *Bradley-Terry models in R: The BradleyTerry2 package*R package version 0.9-4.