Advance Access publication September 24, 2013

Political Analysis (2014) 22:86–114
doi:10.1093/pan/mpt018

Scoring from Contests

Keith E. Schnakenberg
Department of Political Science, Washington University in St. Louis, 1 Brookings Drive, Campus Box 1063, St. Louis, MO 63105
e-mail: keschnak@wustl.edu (corresponding author)

Elizabeth Maggie Penn
Department of Political Science, Washington University in St. Louis
e-mail: penn@wustl.edu

Edited by Janet Box-Steffensmeier

This article presents a new model for scoring alternatives from "contest" outcomes. The model is a generalization of the method of paired comparison to accommodate comparisons between arbitrarily sized sets of alternatives in which outcomes are any division of a fixed prize. Our approach is also applicable to contests between varying quantities of alternatives. We prove that under a reasonable condition on the comparability of alternatives, there exists a unique collection of scores that produces accurate estimates of the overall performance of each alternative and satisfies a well-known axiom regarding choice probabilities. We apply the method to several problems in which varying choice sets and continuous outcomes may create problems for standard scoring methods. These problems include measuring centrality in network data and the scoring of political candidates via a "feeling thermometer." In the latter case, we also use the method to uncover and solve a potential difficulty with common methods of rescaling thermometer data to account for issues of interpersonal comparability.

1 Introduction

Many of the methods used by political scientists to make inferences from data can be understood simply as ways of assigning numbers, or *scores*, to the objects under consideration. For instance, a utility function, which may be estimated from choice data, assigns a number to each policy or alternative representing the decision-maker's preferences over the alternatives relative to some baseline. Scores are useful when we think that the objects under consideration can be meaningfully ordered along some single dimension, such as quality or utility. In such instances, data may often be structured as wins and losses, as in a voter's choice of one candidate over another, a consumer's choice of one bundle of goods from a set, or a treatment's success relative to another in a matched pair of patients. However, in spite of the structure that such comparisons possess, there remain ambiguities about how to approach the scoring problem theoretically when the number of things to be scored is more than two. These ambiguities are multiplied when more than two objects are compared simultaneously and when the outcome of each comparison is potentially continuous. In this article, we develop one particular kind of scoring technique that addresses this problem and that we believe is of broad generality and relevance to political science: scoring from contests.

A *contest* is an event in which a collection of players compete for a fixed prize. The outcome of a contest is a distribution of the prize winnings among the collection of players who competed. This outcome is assumed to be a function of the relative quality of the players. We are interested in

Authors' note: We are grateful for the helpful comments of David Darmofal, Mark Fey, Holger Kern, John Patty, and especially Stephane Wolton, along with seminar participants at Washington University, the Harris School of Public Policy, the University of South Carolina, the Stanford Graduate School of Business, the University of Maryland, and the University of Pittsburgh. This research was supported by NIH Grant (#1RC4LM010958-01), and we are particularly indebted to William Shannon, Elena Deych, Skye Buckner-Petty, and Berkley Shands at the Washington University School of Medicine for their support. Replication materials for this article are available from the *Political Analysis* Dataverse at http://hdl.handle.net/1902.1/21998.

© The Author 2013. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

scoring players according to their quality. A straightforward example is sports rankings. We would like to rank a set of teams on the basis of their quality, with "quality" conceived of as a measure of "ability to win a game." We know the particular matches each team engaged in and the outcomes of these matches, although it may be the case that certain pairs of teams never played each other, whereas other pairs played each other multiple times. The goal then is to use the data from the paired contests to score all of the teams on the basis of their capabilities.

Many different phenomena involve contest-type outcomes for objects that faced varying comparison sets. A good illustration of our method is the simple problem of ranking students on the basis of their grade point averages. Suppose, by way of example, that there are ten students and each is one of three types: either a *low*, *average*, or *high* performer. Three students are low types, three are high, and four are average. Let $S = \{l_1, l_2, l_3, a_1, a_2, a_3, a_4, h_1, h_2, h_3\}$ denote the set of students, with l_1 denoting the first *low* student, and so on. Additionally, suppose that each class is graded on a curve to ensure that the mean grade is 75. If all ten students were in a single class together, their expected grades would be 55 for the low types, 75 for the average types, and 95 for the high types. However, due to their interests, students are more likely to take classes with others of their type. A student's expected grade in a class is proportional to their performance in the class relative to the performance of the other students who also took the class.[1] Finally, suppose that five classes are offered and that the composition of the classes, and the (curved) grades of the students, are as follows:

Class	*Composition*	*(Curved) Grades*
Class 1	l_1 l_2 l_3	75 75 75
Class 2	a_1 a_2 a_3 a_4	75 75 75 75
Class 3	h_1 h_2 h_3	75 75 75
Class 4	l_1 l_2 l_3 a_1 a_2	65 65 65 90 90
Class 5	a_3 a_4 h_1 h_2 h_3	64.5 64.5 82 82 82

We would like our ranking of the students to ultimately recover their true abilities relative to each other—to be akin to a hypothetical ranking produced if all students had taken every class together. However, ranking students simply based on their average grades yields the following:

Rank	*Students*	*Average grade*	*True ability*
First	a_1 a_2	82.14	75
Second	h_1 h_2 h_3	78.45	95
Third	l_1 l_2 l_3	70.24	55
Fourth	a_3 a_4	69.83	75

In the above table not only is the ultimate ranking incorrect, but the set of average-type students—assumed to have the same true ability—has been partitioned into two groups. One of these groups is ranked *highest* overall and the other is ranked *lowest*. This, of course, happens because a_1 and a_2 chose to take a class with comparatively low performers, whereas a_3 and a_4 chose to take a class with comparatively high performers. Of course, if every class contained every student then we would have no problem; ranking students by grade point average (GPA) would be equivalent to ranking them by ability.[2]

[1] For the ultimate ranking produced by our method to be correct, we need only that a student's grade be an increasing function of their performance relative to their classmates. Class grades need not be curved. However, although different functions enable us to recover the true ranking they will not necessarily enable us to gauge the true cardinal scores representing the student abilities.

[2] GPA would be appropriate for ranking students in any situation in which for every pair of students, i,j, the set of remaining students in each of i's classes equals the set of those in each of j's classes.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Although the example we have just presented attributes the different compositions of classes to differences in student interests, it is also natural to think of this kind of problem being exacerbated in situations in which individuals, groups, or teams seek to artificially inflate their scores by selectively engaging a low-quality pool of competitors. These situations could arise in academics, sports leagues, political competition, and markets. The scoring technique that we develop in the remainder of this article is designed to deal with this difficulty. It recovers true ability from precisely this type of data, and will produce a ranking of alternatives that can differ substantially from naïve rankings that do not account for variation in comparison sets. Moreover, it can be utilized in situations in which comparison sets are strategically chosen by the contestants themselves.

2 Previous Approaches to Ranking and Scoring

The problem of ranking objects that have not faced the same objects of comparison has been analyzed extensively both statistically and analytically for the case of contests between pairs.[3] Zermelo (1929) was the first to model the pairwise matches as independent Bernoulli trials, with the probability of Team i beating Team j being a function of the teams' capabilities, which he terms "playing strengths." In particular, if v_i and v_j represent the teams' respective capabilities, then the probability that i beats j is

$$\frac{v_i}{v_i + v_j}. \tag{1}$$

Zermelo then computed estimates of the capabilities using maximum likelihood estimation. He showed that with a reasonably connected playing schedule, there exist maximum likelihood estimates of the team capabilities v that are unique up to scalar multiplication.[4] The model was later rediscovered by Ford (1957) and Bradley and Terry (1952). It is the most widely used statistical model for scoring objects on the basis of pairwise comparisons, and is known in the statistical literature as the Bradley-Terry model.[5]

Jech (1983) considers the same scoring problem axiomatically by establishing two criteria that the team scores should satisfy, and then proving the existence and uniqueness of scores satisfying these criteria. The first criterion is that if the expected outcome of a match (or division of the unit prize) between Team i and Team j is $(p_{ij}, 1 - p_{ij})$ for $p_{ij} \in (0, 1)$, then for any three teams i, j, and k, the following must hold:

$$p_{ij} \cdot p_{jk} \cdot p_{ki} = p_{ik} \cdot p_{kj} \cdot p_{ji}. \tag{2}$$

The second criterion is that the expected outcomes, or point totals received by the teams, should equal the actual point totals received by the teams, given their playing schedules. If m_{ij} is the number of matches played by i and j, and s_i is the sum of the outcomes of all matches played by Team i, this final criterion requires that

$$\sum_j m_{ij} \cdot p_{ij} = s_i. \tag{3}$$

Jech justifies equation (3) with the simple argument that no team should be expected to have won more matches than it actually did; in other words, the scores should be accurate. Although equation (2) is more difficult to interpret, Luce (1959) proved that this precise condition is equivalent to satisfaction of his *choice axiom*. The axiom can be stated as follows. Let A, B, and T be finite sets,

[3] Stob (1984) provides a clear and thorough survey of this literature.
[4] The assumption needed on the schedule, which we discuss in more detail later, is that there is no partition of the teams into two subsets S and T such that no team in S ever receives a positive score in a contest with a team in T.
[5] See Stob (1984, 278).

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

with $A \subseteq B \subseteq T$. Let $P(x, A)$ be the probability that $x \in A$ is chosen from A and $P(A, B)$ be the probability that, when presented with B, the chosen object is in A. Luce's choice axiom then requires that

$$P(x, B) = P(x, A) \cdot P(A, B). \tag{4}$$

Intuitively, the axiom is equivalent to an "independence of irrelevant alternatives" property in which the ratio of probabilities of choosing one alternative to another should be independent of the total set of alternatives available to choose from, or formally, for any $x, y,$ and S with $\{x, y\} \subseteq S$,

$$\frac{P(x, \{x, y\})}{P(y, \{x, y\})} = \frac{P(x, S)}{P(y, S)}.$$

It is a simple exercise to demonstrate that if the choice axiom is satisfied, then there exists a positive real-valued function v on T such that

$$P(x, A) = \frac{v(x)}{\sum_{y \in A} v(y)}. \tag{5}$$

Clearly in the case of pairwise contests without ties, Luce's equation (5) reduces to Zermelo's condition on playing strengths described in equation (1). Because Jech's equation (2) can be proven to be a consequence of equation (5), and because Zermelo proves that Jech's equation (3) is satisfied by his maximum likelihood estimates, it follows that Jech's and Zermelo's approaches yield the same solutions to the scoring problem, and that these solutions are the estimates yielded by the Bradley-Terry model for paired comparisons.

As Stob (p. 281) notes, although Jech's and Zermelo's approaches yield the same scores, Jech's approach to the problem is fundamentally different than Zermelo's, because Jech does not consider the problem to be one of statistical estimation. In an illustration that Stob attributes to Ford (1957), the problem of ranking teams on the basis of Jech's equations (2) and (3) is no more a statistical problem than that of ranking students on the basis of their grade point averages. As Stob puts it, "The numbers v_i are, in this view, simply measures of performance which we compute, not estimate." Although we remain neutral with respect to whether the problem should be considered one of statistical estimation or not, it should be noted that Jech's model seamlessly allows for a number of extensions that, while not disallowed by maximum likelihood estimation, require significant changes to Zermelo's model in order to accommodate. One such extension concerns the observation of ties in the pairwise matches. In Jech's model, the outcome for Team i in a pairwise contest can be any number between 0 and 1. However, because Zermelo's model assumes that the pairwise matches are Bernoulli trials, allowing for ties requires explicitly adding the likelihood of a third, "tie," outcome to the model.[6]

The second extension involves the collection of "teams" participating in a given contest. Both Jech and Zermelo consider only pairwise contests; in the statistical literature, the Bradley-Terry model has been extended to accommodate contests between triples.[7] However, Luce's choice axiom describes a very general choice scenario in which any number of objects can compete simultaneously. The likelihood of one object being chosen from a set of many is increasing in that object's capability, or quantifiable value, and decreasing in the capability of its competitors. Using a generalized version of Luce's axiom as a starting point, we show that the setting of "contests as paired matches" heretofore described can be usefully extended to encompass a far richer class of contest data; namely, data consisting of contests between subsets of alternatives, with contest outcomes that can equal any distribution of the unit prize. In allowing for contest outcomes

[6]See Davidson (1970).
[7]See Pendergrass and Bradley (1960).

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

that can distribute the prize winnings among multiple teams, we reinterpret Luce's choice probabilities in the following way: Although Luce lets $P(x, B)$ be the probability that x is chosen from set B, we interpret $P(x, B)$ to be the expected performance of object x in a single contest among the elements of B.

A benefit of our approach is its explicit connection to theories about how people make decisions. Rather than allowing our theory of choice to lurk in the background of a set of statistical assumptions, we derive our scores directly from our axioms of choice. In doing so, we develop a method that allows researchers to gain leverage on otherwise difficult inference problems. Specifically, our model is useful for situations in which the researcher (1) believes that there is a unidimensional trait (such as "quality") that usefully rationalizes contest outcomes, but where (2) choice-sets are highly variable between contests, or (3) contest outcomes may be nondiscrete or involve more than one winner.

Similar to Jech's approach, in Section 3 we prove the existence and uniqueness of scores satisfying Luce's choice axiom along with an accuracy condition requiring that actual performance equal expected performance. Our goal is to demonstrate that many problems in political science can be conceived of as contests, and many objects of interest to political scientists can be scored using our technique. To this end, we first provide a number of applications of our technique in Section 4. Our first application uses our method to rate the "likability" of different political figures among subsets of voters from the American National Election Study. Our second application uses our method to analyze various types of data possessing a "community" structure, including social network data. In our final example, we use our scoring method in a different way, to disentangle voter preferences over specific issue positions when voters are faced with candidates who take positions on many issues simultaneously. In Section 5, we present a more detailed comparison of our model to existing statistical approaches to the problem, such as conditional logit, and discuss our approach to obtaining uncertainty estimates of our scores. In Section 6, we examine several social choice theoretic implications of our scores. The first is an interpretation of the scores as a solution concept characterizing the utility function of a "representative voter." The second is the ability of our scoring technique to meaningfully capture failures of an independence of irrelevant alternatives (IIA) property, and thus provide insight into whether the alternatives we rank are complements, substitutes, independent, or "anti-complementary" of each other. Section 7 concludes.

3 The Model

Throughout we will refer to the collection of participants in a contest as *teams*. These teams are the objects that we wish to score; in the "ranking students" example described in the introduction of this article, the "teams" are the ten students. We assume that there is a set N of teams that compete in a collection of contests, with $|N| = n$. Some subset of teams participates in any given contest, and the output of a contest is some distribution of a unit prize. Our goal is then to assess the quality of each team based on its performance in the contests in which it participated.

Let b be the total number of contests, with the set of all contests denoted by

$$\mathscr{B} = \{B^1, B^2, \ldots, B^b\}.$$

A contest B is simply a collection of teams. For any $B \in \mathscr{B}$, we have $i \in B$ if and only if Team i participated in contest B. Let n_i be the number of contests in which i has participated. The set of contests in which i participated is $S_i = \{S_i^1, \ldots, S_i^{n_i}\} \subseteq \mathscr{B}$. Throughout we use subscripts to denote teams and superscripts to denote contests.

As teams are allowed more than one entry in each contest, we let $t_i(B)$ be the number of entries of team i in contest B. We assume the number of entries of i for any contest is a positive real number less than or equal to some number $\bar{t}$. We seek to score the teams, and begin by letting $P(i, B)$ be the expected performance of team i in contest B. The following axiom concerning the terms $P(i, B)$ is an extension of Luce's axiom to accommodate the fact that our contests may contain multiple entries of a single team.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Axiom 1 *(Luce, Independence).* $P(i, B) = \frac{t_i(B)v_i}{\sum_{j \in B} t_j(B)v_j}$, *with* $v_1, \ldots, v_n$ *representing numerical "values," "qualities," or "capabilities" of the teams. We let* $v = (v_1, \ldots, v_n)$, *with* $v \in \mathbb{R}^n_+$.

Although we interpret $P(i, B)$ as i's "expected performance" in contest B, this term could also be interpreted as the probability that object i is chosen from set B, as in Luce (1959). As noted earlier, satisfaction of Luce's axiom is equivalent to the assumption that the ratio of the expected winnings of one team to the expected winnings of another should not depend upon other teams in the contest. In Luce's words,

> ...the idea states that if one is comparing two alternatives according to some algebraic criterion, say preference, this comparison should be unaffected by the addition of new alternatives or the subtraction of old ones (different from the two under consideration). (Luce 1959, 9)

Axiom 1 can be interpreted similarly. Furthermore, Clark and Riis (1998) use a generalization of Luce's axiom equivalent to our Axiom 1 to "unfair" contests in which one competitor has an advantage that is unrelated to her quality; the source of the unfair advantage in our setting is simply that some teams have multiple entries in the contest.

As we have noted, Axiom 1 is equivalent to an IIA assumption. Though IIA is controversial in the discrete choice literature in econometrics, it is the most natural assumption for the scoring problem since the existence of scores that fully characterize expected performance is equivalent to IIA. Some models like the multinomial probit model specify choice probabilities using unidimensional scores but allow nonzero correlations between choices conditional on the score. In these models, knowing the scores is insufficient for predicting the performance of a team in a particular contest since one must also know all of the correlations in outcomes. Therefore, any effort to completely rationalize contest data using unidimensional scores will involve IIA or some equally strong independence assumption. Thus, though Axiom 1 is unlikely to be empirically supportable in every situation, it is a theoretically useful starting point for considering the scoring problem.

For some applications, reducing the abilities of the teams to unidimensional scores involves a great deal of information loss or conceals important aspects of the data-generating process. In the terminology of Page (2007), it may be more appropriate to think of the teams simply as possessing different "toolboxes," useful for different types of contests, rather than relying on a single measure of quality. A multidimensional approach to quality or ability calls may preclude the possibility of ranking teams at all, except perhaps on a limited subset of contests. However, unidimensional scores and rankings are theoretically important in political science and have proven to be empirically useful in many circumstances. Thus, we take unidimensionality as given and focus on the problem of recovering scores from contest outcomes.

Team i's total expected performance is then

$$E_i(v) = \sum_{B \in S_i} P(i, B). \tag{6}$$

For each contest B, there is an observed outcome distributed across the collection of teams in B. This outcome is denoted $r(B) = (r(i, B))_{i \in B}$, with the property that $r(i, B) \geq 0$ for all $i \in B$ and $\sum_i r(i, B) = 1$. Given these observed outcomes, the total actual performance for Team i is

$$A_i = \sum_{B \in S_i} r(i, B).$$

This leads to our second axiom.

Axiom 2 *For all teams i, it is the case that* $E_i(v) = A_i$, *or that total expected performance equals total actual performance.*

We seek to show that there exists a unique collection of scores satisfying Axioms 1 and 2. For the following results, and following Jech's terminology and proof structure, we assume that all teams are *pairwise comparable*, a condition guaranteeing that the observed contests provide a sufficient basis for comparing teams. This concept is defined below, and is assumed to hold for all pairs of

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

teams. Satisfaction of this requirement essentially implies that there is no partition of the collection of teams N into two sets, N_1 and N_2, in which no team in N_1 ever received a positive outcome from a contest in which a team in N_2 participated.

Definitions: "Comparability" and "pairwise comparability" of teams.

1. Team i is *comparable* to Team j if and only if there exist $i_1, i_2, \ldots, i_\ell \in N$ and $\bar{B}^1, \bar{B}^2, \ldots, \bar{B}^{\ell-1} \in \mathscr{B}$ such that $i_1 = i$, $i_\ell = j$, and for $k = 1, \ldots, \ell - 1$ it is the case that $i_k, i_{k+1} \in \bar{B}^k$ and $r_{i_k}(\bar{B}^k) > 0$.
2. Team i and Team j are *pairwise comparable* if i is comparable to j and j is comparable to i.

3.1 *Existence and Uniqueness of Scores*

To prove our existence result, we define a function T with domain $\mathscr{X} = \{x \in \mathbb{R}^n : \sum_i x_i = 0\}$ such that for each $x \in \mathscr{X}$,

$$T(x) = (x_i + A_i - E(e^{x_i}))_{i \in N}.$$

It is simple to show that for any $z = T(x)$ we have $\sum_i z_i = 0$,[8] and so T maps $\mathscr{X}$ onto itself. T is also clearly continuous. To prove that there exists a solution to our problem it suffices to show that there exists a fixed point of T, or some $x \in \mathscr{X}$ such that $T(x) = x$. At such a point, it must be the case that $E_i(e^{x_i}) = A_i$ for all i, and it follows that the scoring vector $v_i = e^{x_i}$ solves the system of equations $A_i = E_i(v)$ for all i.

Our proof is an application of Brouwer's fixed point theorem. The challenge is finding a compact subset C of $\mathscr{X}$ such that T maps C onto itself. To establish compactness of C, we proceed in two steps. First, in Lemma 1, we show that the pairwise comparability condition implies that, if there is a "gap" in scores between two teams above a particular bound, then $E_i(x) \geq A_i$ for all of the teams above the gap. Second, in Theorem 1, we use this fact to construct a closed and bounded set C such that T maps C onto itself. This allows us to conclude that T has a fixed point in C.

For the lemma, let d be a positive number, and let $x \in \mathscr{X}$. A *gap* in x of length d is an open interval of length d such that each x_i lies outside the interval. The *top of the gap* is the set of all i such that x_i lies above the interval; the *bottom of the gap* is defined similarly. If i is in the top and j is in the bottom, then $x_i - x_j \geq d$. The numbers $\bar{x}$ and $\underline{x}$ are the maximum score for any bottom team and the minimum score for any top team, respectively. We let $\varepsilon \in (0, 1)$ denote the smallest result of a contest that is not zero. Additionally, we let $\mathscr{T}_G$ and $\mathscr{B}_G$, respectively, denote the top and bottom of a (to be defined) gap, G. $B_{\mathscr{T}} = \cup_{i \in \mathscr{T}_G} S_i$ is the set of contests in which a top team participated. We let $n_{\mathscr{T}} = |B_{\mathscr{T}}|$ be the number of such contests. The number $t^{\max} = \max_{i,B} t_i(B)$ is the maximum multiplicity of any team in a single contest. And finally, we define a number q equal to the maximum number of bottom teams that participated in a contest in which a top team participated, or $q = \max_{B \in B_{\mathscr{T}}} |i : i \in B \cap \mathscr{B}_G|$.

Lemma 1 *For $x \in \mathscr{X}$, suppose that G is a gap in x of length $d \geq \log(\frac{t^{\max} qb}{\varepsilon})$ in which $x_i \geq 0$ for all $i \in \mathscr{T}_G$. Then* $\sum_{i \in \mathscr{T}_G} (A_i - E_i(x)) \leq 0$.

Proof: We start with the observation that our pairwise comparability condition implies that

$$\sum_{i \in \mathscr{T}_G} A_i \leq n_{\mathscr{T}} - \varepsilon$$

[8] This is because $\sum_i A_i = \sum_i E_i(v)$, with both sums equaling the total number of contests played.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

since some bottom team received at least ε in a match involving a top team. We also have

$$\begin{aligned}
\sum_{i\in\mathcal{T}_G} E_i(x) &= \sum_{i\in\mathcal{T}_G}\sum_{k=1}^{n_i}\frac{t_i(S_i^k)e^{x_i}}{\sum_{j\in S_i^k} t_j(S_i^k)e^{x_j}} \\
&= \sum_{B\in B_{\mathcal{T}}}\frac{\sum_{j\in\mathcal{T}_G\cap B} t_j(B)e^{x_j}}{\sum_{l\in B} t_l(B)e^{x_l}} \\
&= \sum_{B\in B_{\mathcal{T}}}\frac{1}{1+\frac{\sum_{l\in\mathcal{B}_G\cap B} t_l(B)e^{x_l}}{\sum_{j\in\mathcal{T}_G\cap B} t_j(B)e^{x_j}}} \\
&\geq \sum_{B\in B_{\mathcal{T}}}\frac{1}{1+\frac{\sum_{l\in\mathcal{B}_G\cap B} t_l(B)e^{\bar{x}}}{\sum_{j\in\mathcal{T}_G\cap B} t_j(B)e^{\underline{x}}}} \\
&\geq \sum_{B\in B_{\mathcal{T}}}\frac{1}{1+\frac{\sum_{l\in\mathcal{B}_G\cap B} t_l(B)}{\sum_{j\in\mathcal{T}_G\cap B} t_j(B)}e^{(\bar{x}-\underline{x})}} \\
&\geq \sum_{B\in B_{\mathcal{T}}}\frac{1}{1+\sum_{l\in\mathcal{B}_G\cap B} t_l(B)e^{-d}} \\
&\geq \frac{n_{\mathcal{T}}}{1+qt^{\max}e^{-d}},
\end{aligned}$$

where the penultimate inequality follows from the facts that $\bar{x}-\underline{x}<-d$ and $\sum_{j\in\mathcal{T}_G\cap B} t_j(B)\geq 1$ for each B.

Using the inequality $d\geq\log(\frac{t^{\max}qb}{\varepsilon})$, we get

$$e^{-d}\leq\frac{\varepsilon}{t^{\max}bq}\leq\frac{\varepsilon}{t^{\max}q(b-\varepsilon)}\leq\frac{\varepsilon}{t^{\max}q(n_{\mathcal{T}}-\varepsilon)}.$$

Thus,

$$1+t^{\max}qe^{-d}\leq 1+\frac{\varepsilon}{n_T-\varepsilon}.$$

It now follows that

$$\frac{1}{1+\frac{\varepsilon}{n_{\mathcal{T}}-\varepsilon}}=\frac{n_{\mathcal{T}}-\varepsilon}{\varepsilon}\geq n_{\mathcal{T}}-\varepsilon$$

since $\varepsilon\leq 1$ and $n_{\mathcal{T}}\geq 1$. Thus, $\sum_{i\in\mathcal{T}_G} E_i(x)\geq\sum_{i\in\mathcal{T}_G} A_i$ when $x_i\geq 0$ for all $i\in\mathcal{T}_G$. □

The next result shows that there exists a compact, convex set C such that T maps C onto itself. It immediately follows, by Brouwer's fixed point theorem, that there is a solution to our scoring problem.

Theorem 1. *There is a nonempty compact, convex subset C of $\mathcal{X}$ such that for each $x\in\mathcal{X}$, $T: C\to C$.*

Proof: Our proof directly follows that of Jech (1983). Let d be a positive number with $d\geq\log(\frac{bq}{\varepsilon})$ and $d\geq b$. Consider the following, strictly positive, $n-1$ numbers:

$$\begin{aligned}
a_{n-1} &= 2(n-1)!*d \\
a_{n-2} &= 2(n-2)!*d+a_{n-1} \\
&\cdots \\
a_3 &= 12d+a_4 \\
a_2 &= 4d+a_3 \\
a_1 &= 2d+a_2.
\end{aligned}$$

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Let C be the set of all $x \in \mathscr{X}$ such that

$$\begin{array}{lcl} \max\{x_i : i \leq n\} & \leq & a_1 \\ \max\{x_{i_1} + x_{i_2} : i_1 \neq i_2\} & \leq & 2a_2 \\ & \cdots & \\ \max\{x_{i_1} + x_{i_2} + \ldots + x_{i_{n-1}} : i_1, \ldots, i_{n-1}\text{distinct}\} & \leq & (n-1)a_{n-1}. \end{array}$$

C is nonempty, closed, bounded, and convex (being the intersection of a hyperplane and a finite number of half spaces). It therefore suffices to show that $T(x) \in C$ whenever $x \in C$. Let $z_1 \geq z_2 \geq \ldots \geq z_n$ be the ordering of the elements of x in decreasing order, and let $z_1^* \geq z_2^* \geq \ldots \geq z_n^*$ be the ordering of the elements of $T(x)$ in decreasing order. Because $T_i(x) = x_i + A_i - E_i(x) \leq x_i + A_i$, it follows that

$$z_1^* + z_2^* + \ldots + z_i^* \leq z_1 + z_2 + \ldots + z_i + b.$$

By the assumption of $z_i \in C$, we also have $z_i \leq a_i$ for $i = 1, \ldots, n-1$, and $z_n \leq 0$.

Beginning with z_1^*, we know that $z_1 \leq a_1$. If $z_1 \geq a_2 + d$, then since $z_2 \leq a_2$, we have that z_1 is positive (trivially) and in the top of a gap of length $\geq d$. By Lemma 1 we have $z_1^* \leq z_1 \leq a_1$. If $z_1 \leq a_2 + d$, then

$$z_1^* \leq z_1 + b \leq z_1 + d \leq (a_2 + d) + d = a_1.$$

Moving to z_2^*, if $z_2 \geq a_3 + d$, then by the same argument, $\{z_1, z_2\}$ are both positive and are the top of a gap of length $\geq d$. By Lemma 1, we have

$$z_1^* + z_2^* \leq z_1 + z_2 \leq 2a_2.$$

If $z_2 \leq a_3 + d$, then

$$z_1^* + z_2^* \leq z_1 + z_2 + b \leq a_1 + (a_3 + d) + d = (a_2 + 2d) + (a_2 - 3d) + d = 2a_2.$$

Similarly, if $z_3 \geq a_4 + d$, then $\{z_1, z_2, z_3\}$ are positive and the top of a gap of length $\geq d$. Thus, $z_1^* + z_2^* + z_3^* \leq z_1 + z_2 + z_3 \leq 3a_3$. If $z_3 \leq a_4 + d$, then

$$\begin{aligned} z_1^* + z_2^* + z_3^* \leq z_1 + z_2 + z_3 + m &\leq a_1 + a_2 + (a_4 + d) + d \\ &= (a_3 + 6d) + (a_3 + 4d) + (a_3 - 11d) + d \\ &= 3a_3. \end{aligned}$$

We continue in this way to deal successively with $z_4^*, \ldots, z_{n-1}^*$ in order to verify that $T(x) \in C$ whenever $x \in C$. If follows that there is a nonempty, compact, convex subset C of $\mathscr{X}$ such that T maps C onto itself. □

Corollary 1 *(Existence). There exists a vector $v \in \mathscr{R}_+^n$ of scores satisfying the system of equations $E_i(v) = A_i$ for all i.*

Proof: By Theorem 1 and Brouwer's fixed point theorem, there is an $x \in \mathscr{X}$ satisfying $T(x) = x$. It follows that the collection of scores $v_i = e^{x_i}$ solves $E_i(v) = A_i$ for all i. □

We now prove that the solution to our ranking problem is unique up to scalar multiplication.

Theorem 2 *(Uniqueness). Suppose that all teams are pairwise comparable. If $v, w \in \mathbb{R}_+^n$ satisfy $E_i(v) = A_i$ and $E_i(w) = A_i$ for all $i \in N$, then $w = cv$ for some scalar c.*

Proof: If all teams are pairwise comparable, then any solution to $E_i(v) = A_i$ gives $v_i > 0$ for all i, because all A_i are strictly positive. Let v and w satisfy $E(w) = E(v) = A$. Suppose that there is no scalar c such that $w = cv$. Then we must have $w_i/v_i > w_j/v_j$ for some $i, j \in N$. For any $i, j \in N$, let $\delta_{ij} = \frac{w_i v_j}{w_j v_i}$. By pairwise comparability, there is a contest $B \in \mathscr{B}$ such that $\delta_{ij} > 1$ for some $i, j \in B$. Let $D = \{i \in N$: there exists $j \in N$ and $B \in \mathscr{B}$ such that $i, j \in B$ and $\delta_{ij} > 1\}$.

Let i^o be an element of D that maximizes $w_i v_\ell$ for some ℓ. Thus, if $\delta_{i^o j} < 1$ for some j, then i^o and j were not in a contest by the fact that i^o maximizes $w_i v_\ell$ for all elements of D. Thus, for all j such that

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

$i^o, j \in B$ for some $B \in \mathscr{B}$, we have $w_{i^o} v_j \geq v_{i^o} w_j$, which implies that $w_{i^o}/v_{i^o} \geq w_j/v_j$, and the inequality is strict for at least one j. Letting $(c_1, \ldots, c_i, \ldots, c_n)$ be the set of numbers solving $w_i = c_i v_i$ for all i, we have

$$\begin{aligned}
E_{i^o}(w) &= \sum_{k=1}^{n_{i^o}} \frac{t_{i^o}(S_{i^o}^k) w_{i^o}}{\sum_{j \in S_{i^o}^k} t_j(S_{i^o}^k) w_j} \\
&= \sum_{k=1}^{n_{i^o}} \frac{t_{i^o}(S_{i^o}^k) c_{i^o} v_{i^o}}{\sum_{j \in S_{i^o}^k} t_j(S_{i^o}^k) c_j v_j} \\
&= \sum_{k=1}^{n_{i^o}} \frac{t_{i^o}(S_{i^o}^k) v_{i^o}}{\sum_{j \in S_{i^o}^k} t_j(S_{i^o}^k) \frac{c_{i^o}}{c_j} v_j} \\
&> \sum_{k=1}^{n_{i^o}} \frac{t_{i^o}(S_{i^o}^k) v_{i^o}}{\sum_{j \in S_{i^o}^k} t_j(S_{i^o}^k) v_j} \\
&= E_{i^o}(v).
\end{aligned}$$

The inequality holds because $c_{i^o}/c_j \leq 1$ for all j in a contest with i^o and is strictly less than 1 for some such j. Thus, $E_{i^o}(w) > E_{i^o}(v)$, which contradicts the hypothesis that $E_i(w) = E_i(v) = A_i$ for all $i \in N$. □

3.2 *Sampling-Based Justification for Scoring*

In the model and results that we have presented so far, there is no role for sampling. That is, our framework proceeded as if the analyst had the entire population of contests and simply desired a useful summary measure of performance. More often, political scientists analyze samples of contests that are presumed to be representative of the population. Do our scores provide reasonable estimates of population parameters? Theorem 3 shows that our scores, when estimated on a random sample of contests, provide consistent estimates of team performance.

Since many political science applications rely on sample data, many researchers would like to interpret the scores as an estimate of some population characteristic. Furthermore, consistency of our scores (along with continuity of the sampling distribution, which is also established by Theorem 3) justifies the use of the bootstrap estimator to obtain confidence intervals. This is useful for researchers who wish to quantify the estimation uncertainty when using our method, and is utilized in some of the examples in the following section.

Let $k = 1, \ldots, b^*$ be a random sample. For each k, we observe a set of contest participants B_k, the number of entries $t_i(B_k)$ for team i in contest k, and an outcome $r(i, B_k)$ for each i in B_k. The list of triples $B^* = \left(B_k, (t_i(B_k))_{i \in B_k}, (r(i, B_k))_{i \in B_k}\right)_{k=1, \ldots, b^*}$ represents all of the observed data. Assume that B^* satisfies pairwise similarity.

Assume that Axiom 1 holds, so there is a vector of numbers v such that

$$\mathbb{E}[r(i, B_k)] = E_i(v) = \frac{t_i(B_k) v_i}{\sum_{j \in B_k} t_j(B_k) v_j}$$

for all k and $i \in B_k$, and assume without loss of generality that $\sum_{i \in N} v_i = 1$.[9]

Let $v^*(B^*)$ be a vector of numbers such that $\sum_{i \in N} v_i^*(B^*) = 1$ and

$$E_i(v^*(B^*)) = A_i = \sum_{B_k \in B^*} r(i, B_k).$$

[9] Recall that $\mathbb{E}$ is the expectation operator, whereas $E_i(v)$ gives the expected *contest performance* for i as a function of v.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

for all i. We wish to show that $v^*(B^*)$ is an unbiased and consistent estimator of v, meaning that the estimated scores converge to the true scores as the sample size gets large. Since it is possible to have a different number of observations for each team, we must define a relevant notion of sample size, which we do below.

Define a set containing only the unique contests in B^*. Thus, define $\tilde{B}$ such that $B \in \tilde{B}$ if and only if, for some $k \in \{1, \ldots, b^*\}$, $B = B_k$ and $t_i(B) = t_i(B_k)$ for all $i \in B_k$. For a particular element $\tilde{B}_\ell$ of $\tilde{B}$, let $\tilde{n}(\tilde{B}_\ell, B^*) = |\{k \in \{1, \ldots, b^*\} : B_k = \tilde{B}_\ell \text{ and } t_i(B_k) = t_i(\tilde{B}_\ell) \forall i \in B_k\}|$ be the number of times that contest appeared in the data, so $\tilde{n}(\tilde{B}_\ell, B^*)/b^*$ is the proportion of times the contest $\tilde{B}_\ell$ was observed. Thus, $\tilde{B}$ and $\tilde{n}$ along with the contest outcomes contain the same information as B^* so that $(B_k, (t_i(B_k))_{i \in B_k})$ pairs are not repeated. Let $\underline{n} = \min_{\tilde{B}_\ell \in \tilde{B}} \tilde{n}(\tilde{B}_\ell, B^*)$ be the minimum number of times a contest appears in the sample.

We say that $v^*(B^*)$ is an unbiased estimator if $\mathbb{E}(v^*(B^*)) = v$ and a consistent estimator of v if $\text{plim}_{\underline{n} \to \infty} v^*(B^*) = v$. Thus, our notion of consistency is that the estimated scores converge to the true scores as the sample of each type of contest in B^* gets large. Theorem 3 shows that this is true of our scoring method.

Theorem 3 *Assume that pairwise similarity holds for B^*. Then $v^*(B^*)$ is an unbiased and consistent estimator of v. Furthermore, for each $i \in N$, $\lim_{\underline{n} \to \infty} \Pr\{v_i^*(B^*) \leq \tau\}$ is a continuous function of τ.*

Proof: By Axiom 1, we have $\mathbb{E}(A_i) = E_i(v) = \sum_{B_k \in B^*} \frac{t_i(B_k)v_i}{\sum_{j \in B_k} t_j(B_k)v_j}$ for all $i \in N$. By Axiom 2, we know that $E_i(v^*(B^*)) = A_i$ and therefore $\mathbb{E}(E_i(v^*(B^*))) = E_i(v)$ for all $i \in N$. Therefore $\mathbb{E}(v^*(B^*)) = v$, proving unbiasedness.

By the law of large numbers, we have

$$\text{plim}_{\underline{n} \to \infty} \frac{A_i}{b^*} = \sum_{\tilde{B}_\ell \in \tilde{B}} \frac{t_i(\tilde{B}_\ell)v_i}{\sum_{j \in \tilde{B}_\ell} t_j(\tilde{B}_\ell)v_j} \frac{\tilde{n}(\tilde{B}_\ell, B^*)}{b^*}$$

for all i.

Since we know that $E_i(v^*(B^*)) = A_i$ for all i, we have

$$\frac{E_i(v^*(B^*))}{b^*} = \sum_{B_k \in B^*} \frac{t_i(B_k)v_i^*(B^*)}{\sum_{j \in B_k} t_j(B_k)v_j^*(B^*)b^*} = \sum_{\tilde{B}_\ell \in \tilde{B}} \frac{t_i(\tilde{B}_\ell)v_i^*(B^*)}{\sum_{j \in \tilde{B}_\ell} t_j(\tilde{B}_\ell)v_j^*(B^*)} \frac{\tilde{n}(\tilde{B}_\ell, B^*)}{b^*} = \frac{A_i}{b^*}$$

for all i. Thus,

$$\text{plim}_{\underline{n} \to \infty} \sum_{\tilde{B}_\ell \in \tilde{B}} \frac{t_i(\tilde{B}_\ell)v_i^*(B^*)}{\sum_{j \in \tilde{B}_\ell} t_j(\tilde{B}_\ell)v_j^*(B^*)} \frac{\tilde{n}(\tilde{B}_\ell, B^*)}{b^*} = \sum_{\tilde{B}_\ell \in \tilde{B}} \frac{t_i(\tilde{B}_\ell)v_i}{\sum_{j \in \tilde{B}_\ell} t_j(\tilde{B}_\ell)v_j} \frac{\tilde{n}(\tilde{B}_\ell, B^*)}{b^*}$$

for all i, which implies that $\text{plim}_{\underline{n} \to \infty} v^*(B^*) = v$, which implies that $v^*(B^*)$ is a consistent estimate of v.

To establish continuity, note that $\Pr\{A_i/b^* \leq \tau^0\}$ is a continuous function of τ^0 for all $\tau^0 \in (0, 1)$ as $\underline{n} \to \infty$. Thus, $\Pr\{E_i(v^*(B^*))/b^* \leq \tau^1\}$ is also continuous for all $\tau^1 \in (0, 1)$. Since $E_i(v)$ is a continuous function of v and $v^*(B^*)$ is a set of numbers satisfying $E_i(v^*(B^*)) = A_i$, the solution for $v^*(B^*)$ is continuous in A_i and $\lim_{\underline{n} \to \infty} \Pr\{v_i^*(B^*) \leq \tau\}$ is continuous in τ. □

4 Applications

In this section, we apply our method to several different questions. Our first application uses our method to rate the "likability" of different political figures among subsets of voters from the American National Election Study. Our second application uses our method to analyze various types of data possessing a "community" structure, including social network data. In our final

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

example we use our scoring method in a different way, to disentangle voter preferences over specific issue positions when voters are faced with candidates who take positions on many issues simultaneously. Detailed instructions on how to generate our scores from contest data are presented in Appendix A.

4.1 *Feeling Thermometers*

Feeling thermometers, which have been included in the American National Election Study (ANES) since the 1960s, were developed to measure citizens' affect toward candidates and elected officials. Respondents are asked to rate various prominent political figures on a scale from zero to one hundred, where high "temperatures" indicate positive affect toward the political figure. In the ANES, respondents are asked to rate several national political figures, plus their local House and Senate candidates.

Deriving an aggregate measure of candidate likability from these data carries several challenges. First, the numerical ratings are not comparable across individuals. Different people may interpret the scale differently or have varying baseline levels of positive affect toward political figures in general. Furthermore, since these baseline perceptions may correlate with partisan attitudes, failure to account for such differences may bias estimates of aggregate candidate likability. Second, since each respondent is asked to rate the candidates in her own district or state, the respondents' ratings of political figures are made in relation to very different choice sets. As respondents are likely to evaluate political figures in relation to one another, the analyst may need to account for the choice sets of different respondents in order to ascertain the relative likability of the political figures.

To illustrate how the lack of interpersonal comparability and varying choice sets might complicate inferences from feeling thermometer data, consider the following example. There are two states, each consisting of two districts with a hundred voters each. There are two types of voters, called Hot and Cold. Hot voters always report an average feeling thermometer score of 50 and Cold voters always report an average of 30. Each voter is asked to rate the incumbent President (P), as well as one Senator (S) and their Representative (R). The politicians are characterized by a unidimensional measure of likability, and voters allocate thermometer ratings proportionally to this measure.[10] Consider the distribution of voters and candidates and resulting thermometer scores depicted in Table 1.

We consider two conventional approaches to summarizing these data. First, we consider the approach of simply taking the mean thermometer scores. This approach does not correct for lack of interpersonal comparability or for variation in choice sets. Some political behavior researchers, noting problems created by the lack of interpersonal comparability in feeling thermometers, have suggested subtracting out the respondent's mean temperature prior to analysis (Winter and Berinsky 1999). Thus, we also consider the results of this approach, which adjusts for lack of interpersonal comparability but not necessarily for variation in choice sets.

The "Means" and "Normalized" columns of Table 2 present the results of these two approaches. Both approaches exhibit considerable bias as estimates of the politicians' overall likability. The order of Senator A and Senator B, for instance, are reversed under both approaches. Though the second approach corrects for the composition of voter types in each district, it does not adjust for the fact that Senator A was compared with two very high-quality representatives, whereas Senator B was compared with two very low-quality representatives. In fact, the effect of the varying choice sets on the relative ratings of the senators is exaggerated under the mean-subtraction approach, since the dependence of thermometer ratings is incorporated into the treatment of the data.

Our model provides a potential solution to the problems of the approaches discussed below. We treat the set of thermometer ratings from a respondent as a contest between the politicians being rated by that respondent. As with the mean-subtraction approach, our contest approach normalizes each individual's thermometer ratings to have the same sum in order to account for differing

[10]We ignore problems that may arise because of the discrete and bounded nature of feeling thermometers, since we expect this can cause problems for any method for analyzing feeling thermometers.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table 1 A hypothetical distribution of feeling thermometer data across two states with two districts each

State A					*State B*				
District 1					*District 1*				
Voter type	*N*	*P (40)*	*S (70)*	*R (80)*	*Voter type*	*N*	*P (40)*	*S (30)*	*R (10)*
Hot	10	31.58	55.26	63.16	Hot	90	75	56.25	18.75
Cold	90	18.95	33.16	37.89	Cold	10	45	33.75	11.25
District 2					*District 2*				
Voter type	*N*	*P (40)*	*S (70)*	*R (90)*	*Voter type*	*N*	*P (40)*	*S (30)*	*R (5)*
Hot	0	–	–	–	Hot	50	80	60	10
Cold	100	18	31.5	40.5	Cold	50	48	36	6

Note. The "true" quality scores are listed to the right of each politician in parentheses. The numbers in the cells underneath each politician are the average thermometer scores for that politician for each type of voter. The numbers in the *N* columns are the numbers of each type of voter in each district.

Table 2 Summaries of hypothetical thermometer data in Table 1 using conventional approaches

	True	*Means*	*Normalized*	*Scoring*
Rep A2	90	40.5	10.5	90.004
Rep A1	80	40.4	8.4	79.99
Senator A	70	33.4	2.4	70.002
President	40	43.6	6.1	40.003
Senator B	30	51	7	30.002
Rep B1	10	18	−30	10.001
Rep B2	5	8	−20	5.000

Note. The column gives the actual politician likability scores that generated the data. The "means" column shows the results of simply taking the mean thermometer score of each candidate, and the "normalized means" column shows the results of taking the mean thermometer score after subtracting the respondent mean. The "scoring" column gives the result from our scoring method.

baseline temperatures. However, our approach explicitly models the fact that a respondent's rating of a particular politician may depend on the likability of the other politicians she is asked to evaluate. The final column of Table 2 illustrates that our scoring model returns the correct quantities for this example.[11]

As discussed in relation to the introductory course grade example, the problems with traditional approached may be exacerbated when teams can strategically select into contests. Our method is designed to recover true scores despite this problem. To see this, note that as long as the similarity condition is satisfied, any team can be connected to any other team through a "path" of contests. For example, suppose that House Candidate A tries to manipulate these rankings by entering a district with inferior House and Senate politicians. Since House Candidate A is connected to superior candidates through common national-level candidates, the national-level candidates will perform better in House Candidate A's (inferior) district than in districts with high-quality House and Senate politicians. Thus, Candidate A will be ranked lower than superior competitors.

[11]The results from the scoring method are rescaled to have the same sum as the true scores.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Similarity provides the connecting information we need to ensure that Candidate A cannot receive an artificially high score by strategically choosing its contests.

We now apply our method to the candidate feeling thermometers in the 2008 ANES. Respondents were asked to rate George W. Bush, Barack Obama, Joe Biden, John McCain, and Sarah Palin, in addition to the major party candidates for any Senate or House races in their states or districts. Thus, the structure of the data is similar to the data in Table 1, with some candidates rated by all respondents, some by all respondents in the same state, and others only by certain districts within one state. In all, there were 2322 respondents who gave thermometer ratings of a combined 244 political figures.

To apply our scoring model, we treat a set of thermometer ratings from an individual respondent as a contest between the candidates that respondent was asked to evaluate. Thus, a contest outcome for a candidate is that candidate's share of the total thermometer rating given by the respondent. This treatment is similar to the mean-subtraction approach advocated by previous researchers. Given a set of estimated likability scores, a candidate's expected performance in a contest is that candidate's share of the total likability score of the candidates in the contest. To minimize variation in candidate likability due to ideology, we estimated the scores separately using voters at each level of the seven-point partisan self-placement scale. The result is an estimate of overall likability for each candidate within each of the seven groups of voters.[12]

Figure 1 displays the model estimates for all Senate candidates by state and voter group. In each panel, the score for the Democratic candidate is marked by the letter "d" and the Republican candidates' scores are marked by the letter "r." Though there are some exceptions for very popular Senate candidates, such as Mark Warner in Virginia, the general pattern is toward higher ratings for Republicans relative to Democrats among more Republican groups of voters. Figure 2 displays scores for all House candidates for each voter group, plotted against candidate scores in the "strong Democrat" group. The figures for the House candidates indicate the same pattern as the national political figures regarding discrimination between groups of voters. The scores for the more Democratic voter groups have a high positive correlation with the scores for the strong Democrats, but the correlation is weaker for independent voters and becomes negative for more Republican groups.

In Fig. 3, we display 95% confidence intervals[13] for candidate likability ratings from feeling thermometers. In this figure, the results are displayed only for voters from the state of Texas, so the ratings are for national figures and Texas politicians. The ratings in Fig. 3 are for Democratic voters, which include those that identified as strong Democrats, weak Democrats, or Democrat-leaning independents.

Our method for generating these "likability" scores models voters as homogeneous in their preferences over public figures. Our model is similar in this respect to the qualitative choice models of McFadden (1974), which are justified using a model of a single representative consumer. An alternative justification for the model, also discussed by McFadden, is to suppose that voters have fixed decision rules, but that the selection probabilities of voter decision rules lead to an aggregate distribution over outcomes that satisfies Axiom 1. This is known as the "random utility" model in discrete choice problems. In either case, it is important to emphasize that the scores may omit important variation in preferences at the individual level. Thus, these scores are primarily useful at the aggregate level and would work poorly as a model for explaining individual voters' preferences or as an input to a voter-level regression model.

Another concern with this example is the possibility of multidimensional evaluations. There is considerable disagreement among political psychologists about whether evaluations of political objects can be adequately explained by a unidimensional trait. Though a thorough review of this

[12]Some candidates do not have scores for all types of voters for sampling reasons. For instance, some House of Representatives candidates had no "Strong Republicans" in their district in the ANES sample, so those candidates are not given a score among Strong Republicans.

[13]The confidence intervals were computed through bootstrapping as follows. First, we take m random samples of size n with replacement from n contests. Next, we estimate the model for each of the m bootstrap samples. Finally, we use quantiles from the set of m estimated parameters to construct uncertainty estimates. For more information on bootstrapping methods, see Efron (1982).

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press



Fig. 1 Results of the scoring model for all senate candidates rated by respondents in the 2008 ANES, by state and ideological type of voter. A letter "d" on the plot denotes the Democratic candidate from that election, and a letter "r" denotes the Republican candidate.

debate is beyond the scope of this article,[14] it is important to note that multidimensional evaluations would raise additional concerns about the assumptions of our model. Fortunately, in these data, the one-dimensional approach appears to organize the data well, since ratings of candidates among liberal and conservative groups are inversely related.

4.2 *Measuring Influence in Social Networks with a Community Structure*

A common undertaking in social network analysis is to identify the influential vertices, or people, within a particular network of individuals. Our scoring method provides one technique for assessing the influence of each vertex in a network by simply assuming that observed levels of influence relate to an underlying latent "quality" of the vertices. Although common methods for measuring influence in networks assume that each vertex has the potential to influence every other vertex, many networks reflect temporal, spatial, or other practical constraints that make this assumption implausible. One advantage of our scoring method is that it is appropriate for measuring influence

[14]See Jost, Federico, and Napier (N.d.) for a detailed review of this literature.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

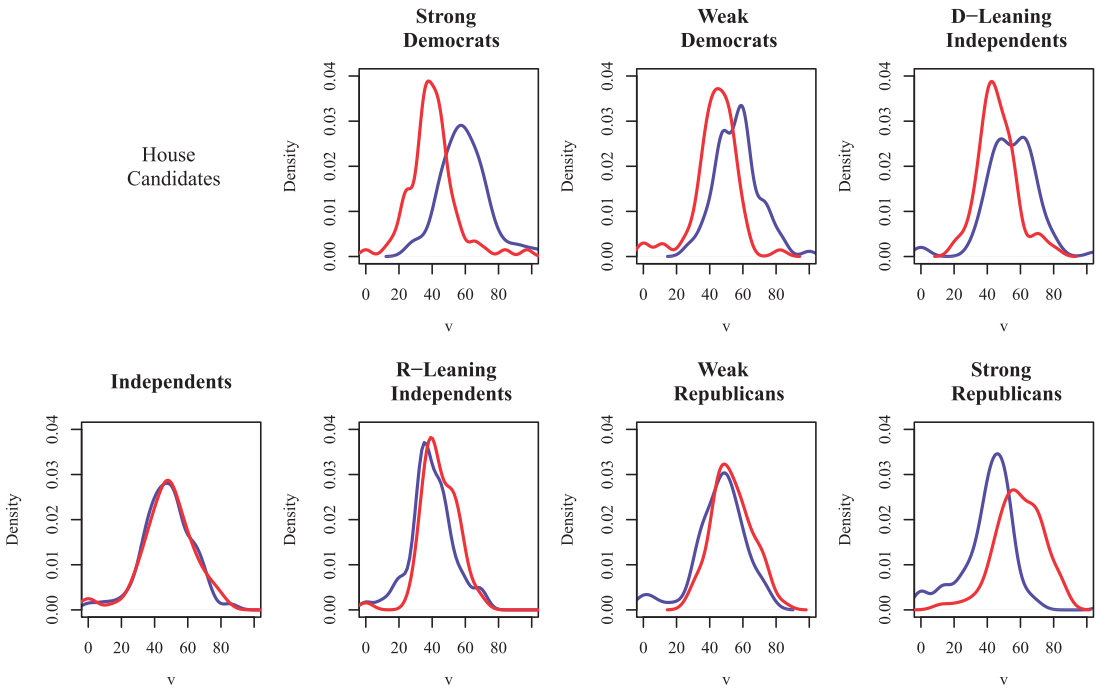


Fig. 2 Results of the scoring model for all House candidates rated by respondents in the 2008 ANES. The analysis was conducted separately for all seven ideological types of voter. The red (light) line represents the density of scores for Republican legislators, and the blue (dark) line represents the density of scores for Democratic legislators.

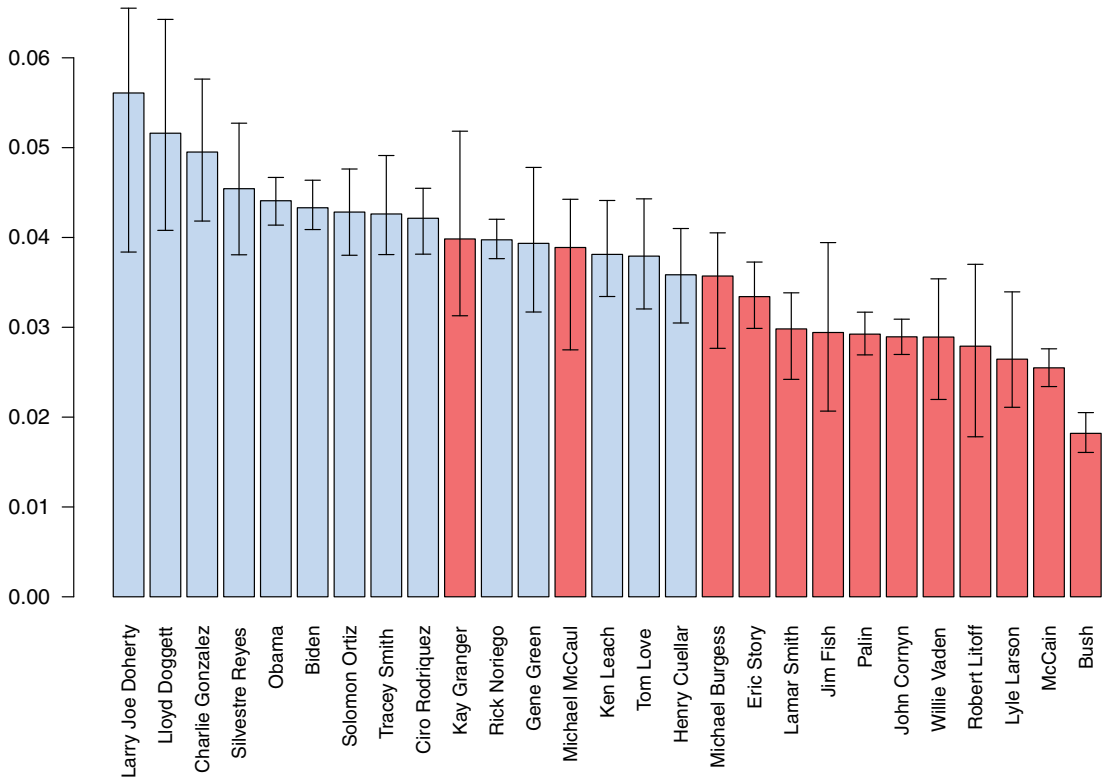


Fig. 3 Likability ratings for politicians among Democratic voters in Texas, with 95% confidence intervals obtained via bootstrapping. Republican politicians are displayed in red (dark), Democratic politicians in blue (light).

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

in networks where (1) some vertices cannot form an edge with certain vertices for reasons that are unrelated to their underlying quality and (2) each vertex may be influenced by a different number of other vertices, so that different edges reveal different amounts of information about the latent quality of the influencing vertices.

One example of a network in which different vertices have different abilities to influence other vertices, irrespective of quality, is the network consisting of Supreme Court majority decisions and the cases citing them.[15] Later decisions cannot be cited by earlier decisions, regardless of the quality of those later decisions. Here we define a vertex's "community" to be the collection of vertices that have the potential to influence it. Defined this way, we can conceive of a particular Supreme Court decision as having a community equal to the collection of decisions that predate it.

In the following example, we consider a different kind of network with a community structure. In this case, we consider a social network in which different individuals had different opportunities to interact with, and thus influence, others. Figure 4 shows fifteen vertices of an (undrawn) social network. In this example, each vertex is a "person" and is assumed to have a community equal to the set of vertices that it is circumscribed with. Loosely speaking, the members of a person's community can be thought of as the people that the person had the potential to meet. In the figure, Vertex 1 has community $\{1, 2, 3, 4, 5, 6, 7\}$, whereas Vertex 6 has community $\{1, 2, 3, 4, 5, 6, 7, 13\}$. Perhaps Vertex 6 lives near Vertices 1 through 7, but attended summer camp with Vertex 13.

We assume that a vertex can only be influenced by elements of their community. We would like to uncover the latent "potential to influence" of each vertex taking into account the fact that certain vertices are more capable of influencing some vertices than others, for reasons that are unrelated to their underlying quality. For example, Vertex 1 has the potential to influence six other vertices, whereas Vertex 14 can only potentially influence two. The colors of the vertices represent their latent influence, with darker nodes being more influential than lighter nodes. In particular, we assume that the darkest nodes are eight times more influential than the lightest nodes and twice as influential as medium-colored nodes: their true scores are represented by $v = (8, 8, 8, 8, 8, 4, 4, 4, 4, 4, 4, 4, 1, 1, 1)$, with v_i being the score of vertex i.

As each vertex can be influenced only by members of its community, we conceive of each vertex as representing a contest between the elements of its community. To generate our network data for this example, we assume that each vertex i has a hundred social interactions over some period of time. Letting C_i be i's community, the probability that i interacts with $j \in C_i$ (alternatively, the probability that j *influences* i) is $\frac{v_j}{\sum_{k \in C_i} v_k}$. The outcome of these randomly drawn interactions yields a weighted and directed network structure. It is straightforward to interpret these weights as contest outcomes and to solve for the true v_i using our contest approach. We illustrate with one particular contest that is pictured in Fig. 5, which depicts the relative influence of vertices in Vertex 1's community on Vertex 1. Our data for this example consist of fifteen such "influence" contests, one per vertex.

Figure 6 plots the v_i generated by one run of the model in which network data were randomly generated via the process described above. On the left, we plot our scores against the true v_i that generated the weighted network. On the right, we plot the outdegree centrality measure of each node against the true scores, having standardized the outdegree centrality scores to sum to one. The outdegree centrality of a vertex is simply the sum of the (in this case, weighted) links directed from that vertex. In the context of this example, it is the number of times a vertex has influenced other vertices. This measure of centrality performs poorly in this example, in part because it advantages teams that straddle communities. For example, although vertices 6 and 7 have the same latent quality, 6 has the ability to influence 13, whereas 7 does not. It also performs poorly because it cannot account for the uneven distribution of quality across the communities.

Because our model imputes unobserved relationships between objects, it is well suited to analyzing networks in which certain links are impossible to observe. As mentioned earlier, these types of networks could arise in situations in which vertices are indexed by time and a later vertex is

[15] This particular network was studied in Fowler et al. (2007). We have analyzed it using our scoring method in Patty, Penn, and Schnakenberg (2013).

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

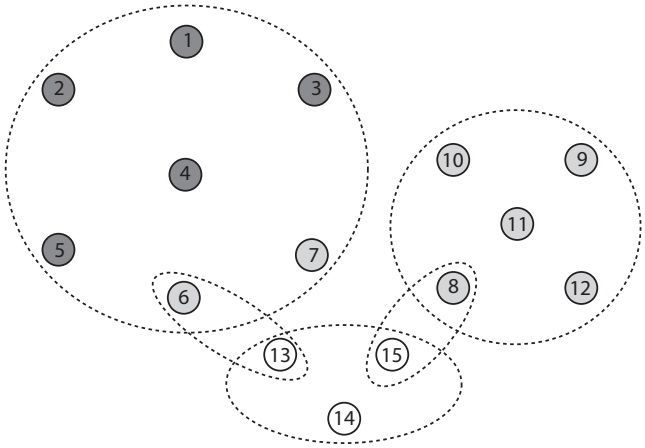


Fig. 4 An example of communities in a social network.

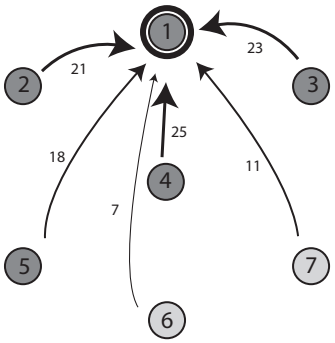


Fig. 5 The vertices that influenced Vertex 1.

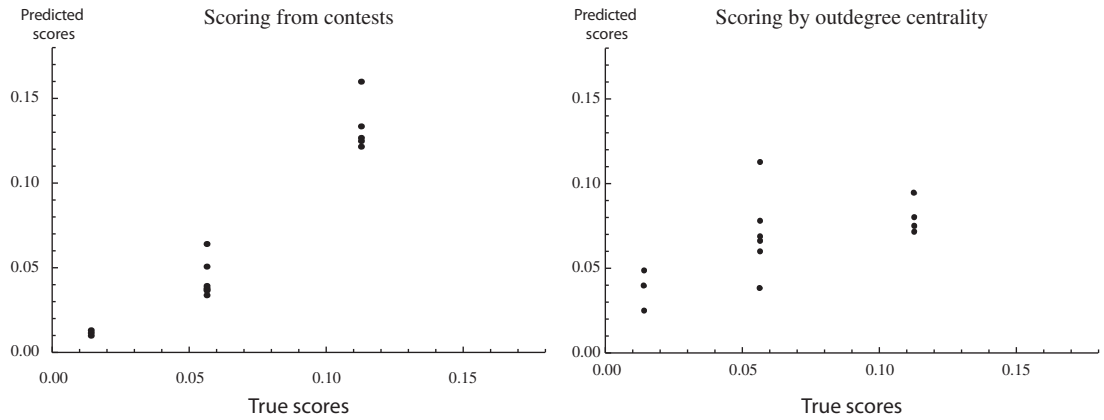


Fig. 6 Contest scores versus outdegree centrality scores.

incapable of influencing a vertex that preceded it. In generating estimates of the v_i using observed network and community data, we can impute "influence relationships" between vertices that did not have the potential to interact. For example, although Vertex 1 never had the potential to interact directly with Vertices 9 and 15, the v_i that we generate enable us to estimate which of these vertices would be more likely to influence Vertex 1 in the event that an interaction *did* occur. This leads to the following interpretation of our scores: if there were a hypothetical vertex with a community equal to the set of *all possible* vertices, then our scores represent the expected influence of each vertex on that hypothetical vertex. Thus, the method provides a different measure of node centrality that generalizes the concept of outdegree centrality.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

4.3 *Assessing the Electoral Value of "Issue Bundles"*

In this last example, we use our scoring technique to disentangle the scores of objects that are bundled together. Specifically, we construct an example to rate the relative electoral value of candidate issue stances, which are always observed in bundles of more than one stance. A single voter choosing between two or more candidates in an election can be viewed as representing a contest between the bundles of issue stances taken by each of the candidates.[16] Using choice between bundles in order to score the objects comprising each bundle presents a somewhat different challenge to our scoring approach than settings in which contest outcomes are directly and unambiguously responsive to the quality of each team in a contest. When a contest consists of a choice between two competing bundles, each of which contains a collection of objects that may potentially have very different values, we confront the problem of how to score objects in a winning bundle when our only information is whether or not the bundle, as a whole, was chosen. We may be faced with an additional problem if the bundles overlap; if, for example, a voter faces a choice between two candidates that share many of the same issue positions.

To illustrate this scenario, consider a thought experiment in which a strategist for the Democratic Party is assessing support for various campaign platforms among Democratic voters. He or she has access to the state's electoral returns from the most recent Democratic primary. In each of his state's three districts, two candidates vied for the nomination, and each candidate could potentially choose a stance on two issues. The strategist might then seek to use the election returns to assess statewide voter preferences over the issues. As described earlier, issue positions are the objects to be scored and votes for candidates represent contest outcomes. To make the example more concrete, electoral outcomes are pictured in Table 3. We assume that the voting behavior of a Democratic voter depends only on the set of issue stances that the candidates have.

In this example, four types of contests occurred, each represented by a different district. Each contest is between the collection of issues represented by the candidates in a particular district. In each of these contests, voters faced only a limited collection of issue stances: in Districts 1 and 3 all candidates were "pro" Issue 1, whereas in Districts 2 and 4 all candidates were "con" Issue 1. Thus, the information provided from these contests is necessarily coarse because voters could not choose the slate of issues that they voted on. Each vote that was cast can be thought of as representing the outcome of a contest between the bundles of stances taken by the candidates being voted upon. Although voting occurs over bundles of issues, the strategist seeks to score each *specific* issue on the basis of voter sentiment.

Let p_1, c_1, p_2, and c_2 denote the underlying values for the above issues we seek to score (p_1 represents "Pro 1" and so on). Our model assumes that the probability a voter casts a ballot for Mr. A over Mr. B, for example, is

$$\frac{p_1 + c_2}{p_1 + c_2 + p_1 + p_2}.$$

Similarly, the probability a voter casts a ballot for Ms. Q over Mr. R is[17]

$$\frac{p_1}{p_1 + p_1 + p_2}.$$

[16] This application is similar in principle to conjoint analysis in psychology, in which respondents are faced with hypothetical choices between alternatives with different attributes, with the goal of rating the importance of each attribute. Recent political science applications include Malhotra and Margalit (2010) and Bechtel and Scheve (2012). Our method addresses problems that, although less common in the experimental settings in these studies, may arise in similar analyses of observational data such as election returns.

[17] Note that we assume that the value of "no issue stance" on an issue is zero, and thus, that it is always better to have some stance on an issue than no stance. The example could just as easily be run by letting the value of a candidate's issue bundle be the weighted average of the values of their issue stances, in which case the probability a voter casts a ballot for Ms. Q over Mr. R is

$$\frac{2 * p_1}{2 * p_1 + p_1 + p_2}.$$

This would change the specific calculations that follow, but poses no challenge to the method.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table 3 A hypothetical Democratic primary as contests between issues

District 1				*District 2*			
Mr. A	Issue 1 Issue 2	Pro Con	42%	Ms. X	Issue 1 Issue 2	Con Con	35.7%
Mr. B	Issue 1 Issue 2	Pro Pro	58%	Ms. Y	Issue 1 Issue 2	Con Pro	64.3%
District 3				*District 4*			
Ms. Q	Issue 1 Issue 2	Pro –	34.8%	Ms. S	Issue 1 Issue 2	Con –	18%
Mr. R	Issue 1 Issue 2	Pro Pro	65.2%	Mr. T	Issue 1 Issue 2	Con Pro	82%

Table 3 was generated using the values $p_1 = 0.4, c_1 = 0.1, p_2 = 0.35$, and $c_2 = 0.15$, assuming that outcomes are expected vote shares.

There are many ways that the information in Table 3 could be analyzed. One could, for example, simply look at the number of votes that were cast in favor of a candidate representing each particular issue stance and then rank the issues according to those numbers. Supposing one thousand voters in each district, this would yield the "Simple tally" scores represented in Table 4. However, this score is unsatisfying because no voter faced a meaningful choice over Issue 1; in each district, the candidates had identical stances on this issue. Alternatively, we could account for the number of times a team appeared in any election and weigh the simple tallies by this amount. For example, "Pro 2" appears four times on a platform, whereas "Con 2" only appears twice, so we might divide the "Pro 2" tally by four and the "Con 2" by two in order to make those issues comparable. This yields the "Weighted tally" column, which, again, is unsatisfying because this weighting still does not account for the fact that voters faced no choice over Issue 1. As a basis of comparison to both these scores, we include a "True scores" column, which shows the vote total that each issue would be expected to receive if voters in a district were allowed to vote on issues individually, instead of in bundles. In this case, we would predict that each issue would receive $v_i * 1000$ votes.

Our method provides a straightforward way to recover true scores using data such as is provided in Table 3. We first score the observed bundles themselves: in this case, the six objects to be scored are $b_A = \{p_1, c_2\}, b_B = b_R = \{p_1, p_2\}, b_X = \{c_1, c_2\}, b_Y = b_T = \{c_1, p_2\}, b_Q = \{p_1\}$, and $b_S = \{c_1\}$, with b_i being Candidate i's stance. Although our pairwise comparability condition does not hold for the observed bundles, we can partition the bundles into two sets that each satisfy the pairwise comparability condition. Doing this yields the following scores:

Set 1		*Set 2*	
Bundle	*Score*	*Bundle*	*Score*
$\{p_1, p_2\}$	0.394	$\{c_1, p_2\}$	6.122
$\{p_1, c_2\}$	0.285	$\{c_1, c_2\}$	3.399
$\{p_1\}$	0.210	$\{c_1\}$	1.344

Recall that a solution (the "score" vector in the above table) is only unique up to multiplication by a positive constant, and because the two sets of bundles above do not satisfy pairwise comparability as a whole, the scores are not comparable across sets. To recover scores for the specific issue stances $\{p_1, p_2, c_1, c_2\}$, we perform the following steps. First, Luce's axiom implies that the sum of values of a bundle's components equals the value of the bundle. Thus, Sets 1 and 2 above describe two distinct systems of equations that can be used to solve for the bundle components, either via a least squares approximation if there are more bundles than components or explicitly if the number

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table 4 Scores generated by tallying versus "true" scores

Simple tally		*Weighted tally*		*True scores*	
Pro 2	2695	Pro 2	673	Pro 1	400
Pro 1	2000	Pro 1	500	Pro 2	350
Con 1	2000	Con 1	500	Con 2	150
Con 2	777	Con 2	388	Con 1	100

of bundles in each system equals the number of components and the system is identified, as in this example. Solving $p_1 + p_2 = 0.394$, $p_1 + c_2 = 0.285$, and $p_1 = 0.210$ yields solutions of

$$\{p_1 = 0.210, p_2 = 0.184, c_2 = 0.075\}$$

for Set 1, and a similar calculation yields

$$\{p_2 = 4.778, c_1 = 1.344, c_2 = 2.055\}$$

for Set 2. Next, as noted earlier, these solutions are not directly comparable. We now use p_2 and c_2 as "connectors" to rescale the solutions of Set 1, in order to put all scores on the same scale. Using least squares, we search for a good approximate solution to the two equations $4.778 = \lambda * 0.184$ and $2.055 = \lambda * 0.075$. We get $\lambda \approx 26.2$. Thus, the solutions to Set 2 approximately equal 26.2 times the solutions to Set 1. Rescaling the solutions from Set 1 by λ and appending with c_1 from Set 2, we get

$$\{p_1 \approx 5.502, p_2 \approx 4.821, c_1 \approx 1.344, c_2 \approx 1.965\}.$$

Finally, we normalize to get the approximate solutions:

$$\{p_1 \approx 0.404, p_2 \approx 0.354, c_1 \approx 0.098, c_2 \approx 0.144\},$$

which differ from our true scores only due to rounding error.

The approach outlined in the preceding paragraphs, of scoring the observed bundles of issues and then solving for the scores of the objects comprising the bundles using a least squares approximation, requires two conditions to hold on the observed data. First, for the least squares approximation to yield a unique vector of scores, it must be the case that the system of equations characterized by setting the linear combination of each set of scored bundles' objects equal to the bundles' totals must not be underdetermined. Second, we require that the bundles that we explicitly solve for must satisfy the "pairwise comparability" condition outlined in Section 3, although pairwise comparability may only hold on subsets of bundles. However, for all subsets of bundles that we solve for, when pairwise comparability is violated there must be "connecting information" across pairwise comparable sets that enables us to rescale in order to place the scores on a common scale. Thus, our first condition requires sufficient variation in the bundles in order to solve for their constituent parts, whereas the second condition requires either connecting information across the bundles in order to compare bundles across contests or a sufficiently rich collection of observed contests in order to satisfy pairwise comparability.

5 Relationship to Maximum Likelihood Approaches to Scoring

As we discussed in Section 1, the justification for our model is axiomatic rather than statistical. A traditional statistical approach to the scoring problem would be to assume a probability distribution for the outcome-generating process and set about the task of finding a set of scores to maximize the likelihood of the observed data. Though we depart from this traditional approach, our method is similar to popular maximum likelihood methods on key dimensions. Below, we show that the scores from our model are equivalent to McFadden's conditional logit model in some typical discrete-choice situations and that it applies in some cases where no current statistical method is available. We also discuss some limitations of our model relative to existing statistical models.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

5.1 *Relationship to Conditional Logit*

The conditional logit model (McFadden 1974) can be viewed as a scoring model for contests with more than two teams and unitary outcomes. McFadden (1974) assumes that individual behavior satisfies the Luce axiom. If v_i denotes the utility value of a team i, it follows that the probability of selecting i from a set of teams B is

$$P(i, B) = \frac{e^{v_i}}{\sum_{j \in B} e^{v_j}}.$$

If we eliminate the possibility that a team has multiple "entries" in a contest as in our examples, this is identical to our definition of $P(i, B)$ in Axiom 1. Thus, in order to show that a conditional logit model can generate scores identical to ours for contests with more than two teams and unitary outcomes, it is only necessary to demonstrate that, at a maximum likelihood estimate of v, the resulting probabilities satisfy the property that the expected total number of times each team is chosen is equal to the actual total number of times the team is chosen. We demonstrate this fact below for a conditional logit model with no team-level covariates.

The joint likelihood of a set of contest outcomes under the conditional logit model is

$$\mathcal{L}(\mathbf{x}|R) = \prod_{B \in \mathcal{B}} \prod_{i \in B} \left(\frac{e^{x_i}}{\sum_{j \in B} e^{x_j}} \right)^{r_B^i},$$

where in this case $r_B^i = 1$ for one and only one element i of B and is equal to zero for all other elements. The log-likelihood function is

$$\ell(\mathbf{x}|R) = \sum_{B \in \mathcal{B}} \sum_{i \in B} \left[r_B^i x_i - r_B^i \ln \left(\sum_{j \in B} e^{x_j} \right) \right].$$

The gradient of the log-likelihood function with respect to $\mathbf{v}$ is

$$\begin{aligned} \nabla \ell_i &= \sum_{B \in B^i} r_B^i - \sum_{B \in B^i} \sum_{h \in B} \frac{r_B^h e^{x_i}}{\sum_{j \in B} e^{x_j}} \\ &= \sum_{B \in B^i} r_B^i - \sum_{B \in B^i} \frac{e^{x_i}}{\sum_{j \in B} e^{x_j}}, \end{aligned}$$

where B^i denotes the set of contests in which i competed. After substituting $v_i = e^{x_i}$, the last line is equivalent to

$$A_i - E_i(v),$$

so the first-order condition for maximizing the log-likelihood with respect to $\mathbf{x}$ implies our Axiom 2. Hence, a conditional logit model (with no covariates) can be considered a special case of our model in which only unitary outcomes are allowed. In addition to allowing for continuous outcomes, our model also allows for choices from sets containing varying quantities of the teams under consideration. Thus, our model allows us to consider situations in which a consumer can pick from a menu of choices (e.g., internet ads appearing on his screen), but some objects appear on the menu with far greater frequency than others.

5.2 *Special Features of Our Method Compared with Statistical Approaches*

Though are model reduces to maximum likelihood approaches in some cases, it also applies in some situations where no current statistical methods are available. Much of the related work in statistics and econometrics, following the work of McFadden, was developed in the context of discrete choice.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Application of choice models to situations where choices are continuous—for instance, allocation of time to various activities, or allocation of a budget to different programs—are more limited.

The analysis of compositional data is also similar to our scoring problem. Compositional data are descriptions of the parts of a whole, such as the components of a sample of soil, the vote shares of different parties in electoral competition, or the allocation of a consumer's budget to spending on different goods. Compositional data are subject to a natural unit-sum constraint, which makes the independence assumptions of most multivariate statistical models problematic (Aitchison 1986). Therefore, the quantitative study of compositional data uses probability distributions on the unit simplex. Compositional data methods have been particularly important in the study of vote shares from multiparty elections (Katz and King 1999).

Compositional data methods as methods of scoring from contests accommodate more than two players per contest and continuous contest outcomes. However, they are not designed for variable, or asymmetric, schedules. Thus, our model may provide a useful generalization of compositional data methods to situations in which choice sets vary from contest to contest, such as in partially contested multiparty elections where voters in different districts face different choice sets of political parties.[18]

Furthermore, a persistent problem with compositional data methods arises when one team receives a zero in some contest. Since compositional methods are based on probability distributions on the unit simplex, contests in which one participant received a score of zero are not well defined, and researchers must impute nonzero outcomes or employ two-stage models (Aitchison and Egozcue 2005). Since our model solves for conditions on the entire set of contests, zero outcomes for a particular contest are unproblematic for our method.

Table 5 summarizes the types of data that can be accommodated by several existing models that could be used for scoring teams in certain contests. A checked box under "More than two players" indicates that the model applies when contests may contain more than two players. "Continuous outcomes" indicates that the method applies when the outcomes of contests are any division of a fixed prize between the teams. "Variable schedules" indicates whether or not the model applies when teams play different schedules. "Continuous entries" indicates whether or not the model can be applied to scenarios in which multiple entries of one player are allowed in a single contest. Each model produces results that are consistent with our model in the situations corresponding to the checked boxes.

5.3 *Limitations Relative to Statistical Methods*

Though our method applies in many situations where existing statistical methods are unhelpful or difficult to apply, it is also limited relative to those methods in two respects. First, additional assumptions would be required in order for our model to incorporate covariates. For many applications, the researcher simply desires a quantitative summary of the observed contests, and the effects covariates are not directly interesting. However, this limitation is particularly important for cases in which the IIA assumption is only plausible conditional on a set of covariates.[19] Thus, extending the model to include covariates would be an interesting extension for future work.

Second, many models in the discrete choice literature allow IIA to be relaxed in a variety of ways. For instance, several papers on vote choice estimate mixed logit (Glasgow 2001) or (computationally difficult) multinomial probit models in order to avoid IIA assumptions (Alvarez and Nagler 1995, 1998, 2000) or incorporate varying choice sets into multinomial logit models as a way of relaxing IIA (Yamamoto 2011). Nested logit models have also been presented as an alternative discrete choice model that relaxes the IIA assumption (Greene 2008, 847–50).[20] As we have argued above, the IIA assumption is most consistent with the goal of scoring from contests, and

[18]See Yamamoto (2011) for an example of a partially contested multiparty election, with an application of a multinomial model for varying choice sets to individual choices.

[19]In these cases, a researcher should only condition on covariates that affect outcomes but cannot effect quality.

[20]Nested logit captures correlations between alternatives by dividing the set of alternatives into "nests" or levels, but maintains IIA within each level. Section 6.2 presents an application of our model that is similar to nested logit in this way.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table 5 Features of various scoring models

	More than two players	*Continuous outcomes*	*Asymmetric schedules*
Bradley-Terry			✓
Jech		✓	✓
Discrete choice	✓		✓
Compositional models	✓	✓	
Our model	✓	✓	✓

researchers should not expect a single set of scores to fully rationalize contest outcomes if IIA is implausible. However, we note that these alternative methods are available to describe choices in this instance.

6 Social Choice Theoretic Issues

In the previous section, we discussed the fact that a statistical approach to generating scores from contest data might assume a probability distribution over the outcome-generating process and then solve for the set of scores maximizing the likelihood of the observed data. Our axiomatic approach makes as strong of an assumption about the data-generating process by assuming that choices over sets of alternatives are consistent with Luce's choice axiom. The scores we generate are therefore an attractive solution for applications in which the Luce axiom is a reasonable assumption on the data-generating process. But under what conditions might we expect this axiom to be reasonable?

Luce's motivating idea stemmed from the view, common in psychology, that individual choice behavior is a probabilistic phenomenon. The choice axiom was conceived as a probabilistic response to "algebraic," or deterministic, theories of choice more commonly employed in economic modeling. In this vein, the scores produced by satisfaction of the axiom provide, in Luce's words, "... a formal counterpart to the intuitive notion of utility (or value) in economics, of incentive value in motivation, of subjective sensation in psychophysics, and of response strength in learning theory" (Luce 1959, 3). In developing his theory of choice as a probabilistic extension of more classic theories of choice, Luce devotes attention to two important axioms in nonprobabilistic choice theory: IIA and transitivity. As we noted earlier, satisfaction of the choice axiom implies satisfaction of a probabilistic form of IIA in which the ratio of the choice probabilities for any two alternatives is independent of the total set of alternatives under consideration. Similarly, satisfaction of the axiom also implies that when choice behavior is deterministic, then pairwise choices are transitive.

A critical assumption underlying Luce's axiom is that choice outcomes are driven by a numerical score assigned to each alternative; the better an alternative, the more likely it is to be chosen. In the case of individual choice behavior, this score might be interpreted as a cardinal measure of the individual's preferences over the choice set.[21] In this respect, the critical assumption underlying our contest data is that team performance is driven by a single "quality" score given to each of the competing teams. Our goal is then to uncover these scores using data on team performance. If contest outcomes are driven by an underlying process that cannot be measured in this way—if choice behavior violates IIA in that the ratios of the choice probabilities are dependent on the choice sets, for example, or if our contest outcomes represent the aggregation of multiple quality scores (such as the preferences of a collection of heterogeneous voters)—then we are faced with a problem of interpreting what it is that our scores mean.

[21] In this case, the utility function represented by the scores is unique up to scalar multiplication but not addition.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

6.1 *Cyclic Group Choice*

The most common example in political science of a collection of contests whose outcomes cannot be derived from numerical scores assigned to the competing teams is the majority preference cycle, in which a majority of individuals prefer alternative x to y, y to z, and z to x. In this case, if contest outcomes are generated by votes between pairs, then it is well known that there is no transitive method of ranking the alternatives in question that is consistent with the contest outcomes (put differently, there is no "group utility function"). At the same time, the scores we derive are not meaningless in this type of example. As discussed earlier, when contests occur between pairs, then these scores represent the most likely estimates of the alternatives' quality given the contest data. Although we know that there is no group utility function that generated the observed outcomes, *if there were* a group utility function generating probabilistic choice outcomes, then it would most likely be represented by our estimates.

As an example, consider the voter information in Table 6. There are a hundred total voters with a distribution of preferences over the alternatives such that 65% of the population prefers A to B, 80% of the population prefers B to C, and 55% of the population prefers C to A. Clearly, there is no way of assigning quality scores to the alternatives such that A is better than B, B is better than C, and C is better than A.

However, if we must score the alternatives because a collective decision over the value of the alternatives in question needs to be reached, then the best we might hope for is a set of scores that are accurate in the following sense: If we treat each individual's vote over a pair of alternatives as a contest between that pair, then the total observed number of wins for each alternative equals the number of wins that would be predicted if every individual votes probabilistically according to the scores we generate. In this regard, our scores are the closest we can get to describing the behavior of a hypothetical "representative voter" of the group.

Finally, in social choice theory, the problem of scoring alternatives from choice data is commonly considered for situations involving complete paired comparisons. Scholars have applied axiomatic (Rubinstein 1980; Laslier 1997) and evolutionary (Laslier and Laslier 2013) approaches to this problem. Our model provides a natural extension of axiomatic scoring methods to more general types of contests.

6.2 *Violations of IIA*

A different weakness in the examples we have presented is that we might expect contest outcomes to be driven by interactions between the competing teams. For example, in our curved grade example, two students may decide to form a study group. In this case, we might expect these students to perform better in classes they take together—in other words, they are complements. Similarly, in our network example it might be the case that a husband and wife are unpleasant to be around individually but delightful as a pair. Or perhaps more realistically, they are delightful individually but awful as a pair (in this case they are "anti-complementary"). It may also be the case that teams are substitutes for each other, such as a pair of individuals who are both pianists. Having both present at a party is no more wonderful than having just one. In this instance, bundling the pianists may result in a score for the bundle that is no greater than the sum of its parts. An example of this is Debreu's classic criticism of Luce's axiom,[22] and the motivating example behind Tversky's alternative theory of choice, *elimination by aspects*:

> Suppose you are offered a choice among the following three records: a suite by Debussy, denoted D, and two different recordings of the same Beethoven symphony, denoted B_1 and B_2. Assume that the two Beethoven recordings are of equal quality, and that you are undecided between adding a Debussy or a Beethoven to your record collection. Hence,
>
> $$P(B_1, \{B_1, B_2\}) = P(D, \{B_1, D\}) = P(D, \{B_2, D\}) = \frac{1}{2}.$$

[22] As an anonymous reviewer pointed out, this example is sometimes presented in econometrics as the red bus–blue bus problem.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table 6 Scoring "group utility" as a representative voter

No. of voters	*Preferences*	*Scores*
45	$A \succ B \succ C$	$v_A = 0.379$
35	$B \succ C \succ A$	$v_B = 0.390$
20	$C \succ A \succ B$	$v_C = 0.231$

It follows readily from the choice axiom that

$$P(D, \{B_1, B_2, D\}) = \frac{1}{3}.$$

This conclusion, however, is unacceptable on intuitive grounds because the basic conflict between Debussy and Beethoven is not likely to be affected by the addition of another Beethoven recording. Instead, it is suggested that in choosing among the three records, B_1 and B_2 are treated as one alternative to be compared with D. Consequently, one would expect that $P(D, \{B_1, B_2, D\})$ will be close to one-half, while

$$P(B_1, \{B_1, B_2, D\}) = P(B_2, \{B_1, B_2, D\})$$

will be close to one-fourth, contrary to [IIA].[23]

In the cases described above, and with sufficiently rich data, our model can be modified to capture these types of interaction effects between alternatives. We accomplish this by directly ranking the bundle of alternatives as a single alternative, and then comparing the score of the bundle to the scores of its constituent parts.

To use Debreu's example, suppose that we hypothesize that B_1 and B_2 are substitutes for each other, so that the score of bundle $\{B_1, B_2\}$ equals both the scores of B_1 and B_2 individually. In a contest including bundle $\{B_1, B_2\}$, we treat any selection from the bundle as a win for the bundle. Suppose we observe consumer behavior in settings in which all three records are available, and in settings in which only two of the three are available, and that behavior looks as follows (with the bottom line of Table 7 showing the number of times we observe any particular choice from a set).

Table 7 Debreu's example of an IIA failure in choosing record albums

Choice set	$\{D, B_1, B_2\}$			$\{D, B_1\}$		$\{D, B_2\}$		$\{B_1, B_2\}$	
Observed choices	D	B_1	B_2	D	B_1	D	B_2	B_1	B_2
	100	52	48	31	28	20	21	40	41

If we simply score the three alternatives from these contests, we get the scores

$$v_D = 0.45,\ v_{B_1} = 0.28, \text{and } v_{B_2} = 0.27.$$

However, if we treat $\{B_1, B_2\}$[24] as a bundle and score the (now four) alternatives from these contests we get the scores

$$v_D = 0.255,\ v_{B_1} = 0.24,\ v_{B_2} = 0.25, \text{and } v_{\{B_1, B_2\}} = 0.255,$$

which suggest a failure of IIA, and, moreover, confirm our hypothesis that B_1 and B_2 are substitutes, as the value $\{B_1, B_2\}$ approximately equals the value of both B_1 and B_2.[25] We can similarly see what happens when we bundle D and B_1, treating them as a single alternative. In this case, we get the scores

$$v_D = 0.16,\ v_{B_1} = 0.16,\ v_{B_2} = 0.16, \text{and } v_{\{D, B_1\}} = 0.52,$$

[23] In Debreu (1960), and quoted in Tversky (1972).

[24] Note that in this case there is a type of contest, $\{B_1, B_2\}$, with only one alternative, and the information from this contest is not utilized.

[25] If a researcher desires a more formal evaluation of the hypothesis of an IIA violation, confidence intervals for these quantities are easily available through bootstrapping.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

suggesting that D and B_1 are complementary in that the score of their bundle is higher than the sum of their individual scores. This is reasonable, because even though the records have the same value and would thus be equally likely to be selected in the absence of any complementarities, in a choice from all three alternatives there is a three out of four chance that the choice will be either D or B_1.

It is useful to note the similarity between this application and the nested logit model. In the nested logit model, we would first model the choice between bundles (i.e., Debussy versus Beethoven) and then the choice between items within the bundles. As with nested logit, this approach relies on the researcher to have some knowledge about the relevant bundles.

7 Conclusions and Extensions

In this article, we have provided a general approach to scoring the alternatives in a set of contests. We assumed that contest outcomes are driven by a unidimensional measure of "quality" and showed that, if the data provide a sufficient basis for comparing alternatives, there is a unique set of scores providing an accurate prediction of the alternatives' overall performance. Our approach generalizes the method of paired comparison to settings in which contests may be between any collection of alternatives and outcomes can be any division of a fixed prize. Our intention was to introduce the reader to the concept of a contest and to demonstrate that a wide variety of problems can be reconceptualized as generating contest-type data. Many of these problems would be difficult (if not impossible) to analyze with standard methods, such as linear or multinomial regression. In particular, by using Luce's IIA axiom to impute unobserved relationships between objects we can, for example, analyze choice behavior in settings where choice sets vary across observations, choices are continuous, and objects may appear a fractional number of times within the choice set.

Our method will prove useful when the analyst believes that the contest outcomes are explained by a unidimensional trait, when the relative expected performance of two alternatives does not depend on which other alternatives are involved in the contest (IIA), and when the data are rich enough that every pair of alternatives can be connected by some chain of contests. In case the first two (closely related) conditions fail, we have demonstrated how to expand the set of alternatives to include "bundles" of related alternatives to test the IIA assumption and possibly redefine the problem such that IIA is plausible. When it is not possible to bundle alternatives in a way that satisfies IIA, the method does not provide reliable estimates of the quality of the alternatives. Though there exist many discrete choice models such as those discussed in Section 5 that relax the IIA assumption, we are unaware of any such models that apply to the general contest outcomes accommodated by our model. Similarly, if the data do not provide a rich enough basis for comparing alternatives to yield a solution to our model, we are not aware of other models that satisfactorily solve the scoring problem.

Several examples were presented that detailed potential uses of our approach. These examples included scoring students in a law school; calculating "feeling" scores for candidates based on self-reported thermometer data from heterogeneous districts; computing a network centrality score that accounts for the fact that different individuals have different communities of acquaintances; disaggregating the effects of issue positions on vote choice when choices are between bundles of issues; and exploring potential violations of IIA. In each of these instances, our model provides one approach to reducing complex data into a single set of simple-to-compute, easy-to-interpret numbers. We believe that these examples capture only a small set of the types of problems that can be analyzed with a contest approach.

Appendix A: How to calculate our scores

We now provide brief instructions on how to score from contests. The first step is to gather the data into a format that captures the alternatives (teams), the contests (collections of alternatives being compared), and the outcomes (division of unit prize among alternatives). A simple example is shown in Table A1. Each row represents a contest (total fifteen) and each column under "Contest Participants" and "Contest Outcomes" represents a team (total ten). Thus, the table depicts

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

Table A1 An example of a data set capturing contest outcomes

	Contest participants										*Contest outcome*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	0	1	1	0	0	1	0	1	0	0	0	0.2	0	0	0	0.3	0	0.5	0	0
2	1	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	1	1	0	1	1	0	0	0	0	0	0.1	0.1	0	0.7	0.1	0	0
4	0	0	0	1	0	1	0	0	0	0	0	0	0	0.8	0	0.2	0	0	0	0
5	0	0	1	0	0	1	0	0	1	0	0	0	0.7	0	0	0.2	0	0	0.1	0
6	1	0	1	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0.5
7	0	0	0	1	0	0	1	0	0	0	0	0	0	0.3	0	0	0.7	0	0	0
8	0	0	1	0	0	0	1	0	1	0	0	0	0.9	0	0	0	0.1	0	0	0
9	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0.2	0	0.2	0.6	0	0
10	1	0	0	0	0	0	1	1	1	0	0.1	0	0	0	0	0	0.1	0.7	0.1	0
11	0	0	1	0	1	0	0	0	0	0	0	0	0.3	0	0.7	0	0	0	0	0
12	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
13	0	0	0	1	1	0	0	0	0	0	0	0	0	0.2	0.8	0	0	0	0	0
14	1	0	0	0	1	0	0	1	1	0	0.4	0	0	0	0.4	0	0	0.1	0.1	0
15	0	0	0	1	0	1	0	1	1	0	0	0	0	0.6	0	0.2	0	0.1	0.1	0

fifteen contests among ten alternatives. Columns two through eleven indicate whether or not an alternative was in a given contest. For example, Contest 1 was between alternatives $\{2, 3, 6, 8\}$; each of these alternatives is assigned a "1," whereas the other alternatives are assigned a "0." Columns twelve through twenty-one indicate the outcomes of each contest. In Contest 1, scores were assigned as follows: Alternative 2 scored 0.2, Alternative 3 scored 0, Alternative 6 scored 0.3, and Alternative 8 scored 0.5. Note that for each contest these numbers sum to one.

Team i's *actual performance*, A_i, is the total sum of the prizes won by Team i in all of the contests it participated in. For Alternative 1 in Table A1, for example, A_1 is the total of the entries in Column 12, or $A_1 = 2$.

If each team has a "quality score" of v_i, then Team i's *expected performance*, $E_i(v)$, is the performance we would expect from Team i given the contests Team i participated in. In Table A1, Alternative 1 participated in Contests 2, 6, 10, and 14, which, respectively, consisted of teams: $\{1, 3, 8, 9\}, \{1, 3, 10\}, \{1, 7, 8, 9\}, \{1, 8, 9\}$. Therefore,

$$E_1(v) = \frac{v_1}{v_1 + v_3 + v_8 + v_9} + \frac{v_1}{v_1 + v_3 + v_{10}} + \frac{v_1}{v_1 + v_7 + v_8 + v_9} + \frac{v_1}{v_1 + v_8 + v_9}.$$

In this same way, we calculate A_i and $E_i(v)$ for all the alternatives. Setting $E_i(v) = A_i$ for all alternatives i yields a nonlinear system of n equations (one per team) in n unknowns $(v_1, \ldots, v_n)$. Corollary 1 proves that there is a solution to this system of equations, and Theorem 2 proves that the solution is unique up to multiplication by a positive constant.

We can solve the system of equations in Mathematica, for example, using the FindRoot function or in `R` using the `BB` package (Varadhan and Gilbert 2009). When inputting the data it is essential to avoid rounding numbers in the system before solving (e.g., when appropriate, fractional data should be entered as fractions and not decimals). We also note that for solutions to be correct, our *pairwise comparability* condition must be satisfied. In this case, we checked it by hand. In settings where it is uncertain whether the condition is satisfied and the condition is not readily checked by hand, the condition can be checked computationally (see attached `R` code).

As noted, our scores are only unique up to multiplication by a positive constant. Given the data above, a quick run of Mathematica yields the particular solution (rounded to):

$$\tilde{v} = (0.038, 0.008, 0.013, 0.010, 0.018, 0.004, 0.013, 0.015, 0.002, 0.051).$$

If we wish to interpret our v_is as choice probabilities, we can standardize $\tilde{v}$ to sum to one:

$$v = (0.219, 0.046, 0.075, 0.058, 0.107, 0.022, 0.080, 0.087, 0.011, 0.294).$$

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press

References

Aitchison, John. 1986. *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Aitchison, J., and J. J. Egozcue. 2005. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology* 37(7):829–50.

Alvarez, R. Michael, and Jonathan Nagler. 1995. Economics, issues and the Perot candidacy: Voter choice in the 1992 presidential election. *American Journal of Political Science* 39(3):714–44.

———. 1998. Economics, entitlements, and social issues: Voter choice in the 1996 presidential election. *American Journal of Political Science* 42(4):1349–63.

———. 2000. A new approach for modelling strategic voting in multiparty elections. *British Journal of Political Science* 30:57–75.

Bechtel, Michael, and Kenneth Scheve. 2012. *Public support for global climate cooperation*. Working paper.

Bradley, Ralph Allan, and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3/4):324–45.

Clark, Derek J., and Christian Riis. 1998. Contest success functions: An extension. *Economic Theory* 11(1):201–4.

Davidson, Roger R. 1970. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* 65(329):317–28.

Debreu, Gerard. 1960. Review of RD Luce, individual choice behavior: A theoretical analysis. *American Economic Review* 50(1):186–88.

Efron, B. 1982. *The jackknife, the bootstrap and other resampling plans*. Vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

Ford, L. R., Jr 1957. Solution of a ranking problem from binary comparisons. *American Mathematical Monthly* 64(8):28–33.

Fowler, J. H., T. R. Johnson, J. F. Spriggs II, S. Jeon, and J. Paul. 2007. Network analysis and the law: Measuring the legal importance of Supreme Court precedents. *Political Analysis* 15(3):324–46.

Glasgow, Garrett. 2001. Mixed logit models for multiparty elections. *Political Analysis* 9(2):116–36.

Greene, William H. 2008. *Econometric analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.

Jech, Thomas. 1983. The ranking of incomplete tournaments: A mathematician's guide to popular sports. *American Mathematical Monthly* 90(4):246–64, 265–66.

Jost, John T., Christopher M. Federico, and Jaime L. Napier N.d. Forthcoming.

Katz, Jonathan, and Gary King. 1999. A statistical model for multiparty electoral data. *American Political Science Review* 93:15–32.

Laslier, Benoit, and Jean-François Laslier. 2013. Reinforcement learning from comparisons: Three alternatives is enough, two is not. Working paper.

Laslier, Jean-François. 1997. *Tournament solutions and majority voting*. Studies in Economic Theory Series. Berlin: Springer.

Luce, R. Duncan. 1959. *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Malhotra, Neil, and Yotam Margalit. 2010. Short-term communication effects or longstanding dispositions? The publics response to the financial crisis of 2008. *Journal of Politics* 72:852–67.

McFadden, Daniel. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka, 105–42. New York: Academic Press.

Page, S. E. 2007. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Patty, John W., Elizabeth M. Penn, and Keith E. Schnakenberg. 2013. Measuring the latent quality of precedent: Scoring vertices in a network. In *Advances in political economy: Institutions, modelling and empirical analysis*, eds. Daniel Kselman, Gonzalo Caballero, and Norman Schofield. Berlin: Springer.

Pendergrass, Robert N., and Ralph A. Bradley. 1960. Ranking in triple comparisons. In *Contributions to probability and statistics*, eds. O. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, 331–51. Stanford, CA: Stanford University Press.

Rubinstein, Ariel. 1980. Ranking the participants in a tournament. *SIAM Journal on Applied Mathematics* 38(1):108–11.

Schnakenberg, Keith. 2013. Replication data for: Replication materials for "Scoring from Contests". http://hdl.handle.net/1902.1/21998 IQSS Dataverse Network. V1.

Stob, Michael. 1984. A Supplement to "A Mathematician's Guide to Popular Sports." *American Mathematical Monthly* 91(5):277–82.

Tversky, Amos. 1972. Elimination by aspects: A theory of choice. *Psychological Review* 79(4):281.

Varadhan, Ravi, and Paul Gilbert. 2009. BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* 32(4):1–26.

Winter, Nicholas, and Adam Berinsky. 1999. What's your temperature? Thermometer ratings and political analysis. Working paper.

Yamamoto, Teppei. 2011. A multinomial response model for varying choice sets, with application to partially contested multiparty elections. Working paper.

Zermelo, Ernst. 1929. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 29:346–60.

https://doi.org/10.1093/pan/mpt018 Published online by Cambridge University Press