Predicting Candidate Preference Shares in Preferential-List PR Systems: A Logical Model of Intra-Party Competition

Leonardo Carella[†]

Preliminary draft dated March 12, 2023

Abstract

This paper derives and tests quantitative predictions for three indicators of intra-party competition under preferential-list proportional representation (PLPR) rules: the share of preference votes of the first-ranked candidate (v_1), the effective number of candidates (N_c), and the share of preference votes of the last eligible candidate (v_s). First, it presents a model where these variables are functions of input quantities measurable at election time: the number of candidates c, the number of seats won s and the maximum number of preference votes p. Then, it is shown that these equations can be recast in terms of purely institutional quantities to predict average expected values of v_1, N_c and v_s for any seat-winning list in a district. Given a district magnitude M, a maximum number of preference votes allowed p and a parameter r which captures the permissiveness of over-nomination rules, v_1 is predicted to be $(prM^{\frac{11}{8}})^{-\frac{1}{4}}$, N_c is predicted to be $(prM^{\frac{11}{8}})^{\frac{3}{8}}$ and v_s is predicted to be $(prM^{\frac{5}{2}})^{-\frac{1}{4}}$. These relationships are tested on a diverse sample of data from 31 PLPR elections in nine countries. The model's predictions come remarkably close to describe the empirical relationship between the observations in the sample, and are substantially less biased than those of existing models of intra-party competition. It is also shown that the district-level model can predict *average* values of intra-party competition in a district about as well as the seat-product model predicts analogous inter-party quantities. Implications for further research on the intra-party dimension of electoral systems and institutional design are discussed in the conclusion.

[†]Nuffield College and University of Oxford. email: `leonardo.carella@nuffield.ox.ac.uk`

1 Introduction

The seat-product model (SPM) represents a major achievement of the 'Duvergerian agenda', the scholarly quest to identify regularities in the relationship between institutional features of electoral systems and political outcomes (Taagepera and Grofman, 1985; Taagepera and Shugart, 1993; Taagepera, 2007; Taagepera and Sikk, 2010; Sikk and Taagepera, 2014; Li and Shugart, 2016; Shugart and Taagepera, 2017). Through their career's work, Rein Taagepera and Matthew Søberg Shugart – occasionally in collaboration with other authors – have shown that key features of party systems (and the democratic political process more broadly) can be derived deductively from a small set of quantities that characterise an electoral system: primarily, district magnitude (M) and assembly size (S). These relationships between institutional input variables and predicted average outcomes 'in expectation' are normally in the form of $Y = X^k$, an exponential functional form more common to the laws of natural sciences than the linear expressions used in much of social science modelling (Taagepera, 2008, pp. 52-70). For instance, the formula for the expected effective number of parties N_S in an assembly of size S elected from districts of mean magnitude M is $N_S = (MS)^{\frac{1}{6}}$ and for the expected fractional share of the first party σ_1 it is $\sigma_1 = (MS)^{-\frac{1}{8}}$ (Taagepera, 2007; Shugart and Taagepera, 2017).[1] The accuracy of these equations in predicting empirical distributions of party system quantities across repeated elections in large samples is not only substantively interesting for the comparative study of electoral institutions, insofar as they offer a guide as to the expected effects of electoral system reform, but also a testimony to the potential of the theory-building method behind their derivation: logical modelling.

This paper follows in this line of theoretical reasoning, investigating whether features of intra-party competition follow predictable patterns similar to those identified by the seat-product model for inter-party competition. Specifically, it proposes a quantitatively predictive logical model of intra-party competition in preferential-list proportional representation (PLPR) systems, the most common category of preferential voting system. PLPR is defined as an electoral system with the following characteristics: (1) voters can or must cast a personal vote for a candidate within a list of co-partisans, (2) the number of candidates elected in each list is determined by the *pooled* number votes cast at list level, and (3) the attribution of seats to candidates within a list is determined, at least in part, by personal votes (Karvonen, 2004; Shugart, 2005; Passarelli, 2020).

Three quantities of interest that characterise the intra-party distribution of preference votes for a (seat-

[1] I use σ_1 to notate the fractional share of the largest party, whereas this is generally indicated as s_1, to avoid confusion with s (the 'raw' number of seats won by a list, which is central to the model developed in the rest of the paper) and S (the assembly size, which is a key component of the seat-product model).

winning) list competing under preferential-list rules are examined:

— *The share of preference votes obtained by the first candidate in a list (v_1)*: this is simply the number of votes of the first candidate over the total number of preferences cast for the party in that district. The lower this value, the more competitive the list.

— *The effective number of candidates (N_c)*: this is the intra-party analogue of the inter-party notion of 'effective number of parties' (Laakso and Taagepera, 1979), a measure of competition where the count of parties is weighted by their fractional share.[2] In a list with c candidates, N_c is $\frac{1}{\sum_1^c({v_1}^2+{v_2}^2+\ldots{v_c}^2)}$, i.e. the inverse of the sum of squares of candidate shares. The higher this value, the more competitive the list.

— *The share of preference votes obtained by the last eligible candidate (v_s)*: for a list winning s seats in a district, it is the share of preferences gained by the s^{th} candidate.[3] While a higher value of v_s is not *prima facie* an indicator of lower (or higher) competitiveness, the quantity may be of substantive and practical interests for two reasons: in pure open-list PR (OLPR), where only preference votes matter to candidates' election, it identifies the minimum share a candidate must get to win a seat in expectation; in flexible-list PR (FLPR), where threshold constraints apply, it indicates how 'low' the legal threshold should be for the list to function as fully open.

The paper proceeds as follows. In section 2, I illustrate the derivation of the basic building block of the existing models of intra-party competition quantities (the Shugart-Bergman-Watt model): the formula for the expected value of first-ranked candidate's preference share (v_1). I then proceed to propose a refinement of this simple model that extends its scope conditions to systems where voters may express more than one preference vote and incorporates an intuitive, but previously overlooked, assumption about the relationship between intra-party competitiveness of a list and actors' expectations of the list's inter-party performance. In section 3, I argue that, from v_1, it is possible to compute two additional quantities of interest: the effective number of candidates (N_c) and the preference shares for the last eligible candidate in a list (v_s). Tractable approximations of the equations for these quantities are provided. The resulting models provide 'predictions' specific to each seat-winning list, but rely on two input quantities that are realised at election time – the number of candidates in the list c and the number of seats won by the list s – and therefore are not, strictly

[2] To my knowledge, the first to compute effective number of candidates in a PLPR system was Arter (2013).

[3] I refer to this quantity in terms of 'last eligible' rather than 'last elected' because the model aims to apply to all types of PLPR, including flexible-list systems. In this sub-type of PLPR, there is a threshold of preference votes that candidates must meet to be elected on their preference votes, otherwise the allocation of seats defaults to the candidates' position on the list.

speaking, *pre*-dictive. To address this shortcoming, I proceed to show that these equations can be recast in terms of purely institutional input variables, which vary at district level. I therefore derive two separate models for each of the three quantities of interest (v_1, N_c and v_s): (1) a list-level 'post-result' equation, which returns different values for each list in the same district, and (2) a district-level 'results-blind' prediction, which relies uniquely on institutional variables and predicts the value of the quantity in expectation for *any* list in a district. These sets of predictive equations are tested empirically on a sample of open- and flexible-list electoral outcomes, for a total of over 2,600 seat-winning lists. Section 4 describes the data and the institutional characteristics of the countries in the sample. Section 5 outlines the empirical modelling choices, and motivates the indicators chosen to compare and assess model fits. Section 6 presents the results; the performance of the models' estimates and predictions is compared, in turn, with that of existing models of intra-party competition and with that of the more established predictions of inter-party quantities of the seat-product model. I conclude in section 7 highlighting implications and limitations of the analysis.

2 Modelling First-ranked Candidate Preference Shares

2.1 The Shugart-Bergman-Watt (Shugart-Bergman-Watt (SBW)) Model

Applications of the logical modelling approach to intra-party competition under preferential voting rules have already been attempted, most notably in a paper by Shugart, Bergman and Watt (2013), henceforth SBW, in the context of a comparison between open-list and single non-transferable vote systems (a slightly modified version is in Shugart and Taagepera, 2017, 215-235). The basic building block of the SBW model is the formula for the fractional share of preference votes for the first-ranked[4] candidate of a seat-winning list. An appealing feature of this model is its parsimony: in its simplest form, the SBW models employs only one input variable, c, the number of candidates in a list. However, the deriving predictions fall well short of the accuracy of those available for the inter-party dimension, and to improve accuracy SBW often have to rely on constants derived empirically that correct the equations for biases that cannot be accounted for theoretically. Moreover, these models' scope has so far been limited to a relatively narrow set of 'simple' preferential voting systems, which exclude flexible list types or systems that allow multiple preference votes. The wager of this paper is that we can do better at relatively little cost in terms of simplicity.

It is however worth revisiting the derivation of v_1 in the SBW model, as an illustration of how logical

[4]I used the term 'rank' to refer to the ordering of candidates according to their preference votes, and 'position' to denote their order on the ballot paper.

model-building proceeds. The first step consists in identifying the conceptual boundaries of the quantity of interest. At one extreme, v_1 may not exceed 1: at best, the first-placed candidate can get *all* the preference votes cast for a given list. As for the lower bound, v_1 may not be lower than $\frac{1}{c}$: this is the case where all candidates in the list get the same share of votes, so that the list is maximally competitive. A candidate getting less than $\frac{1}{c}$ cannot logically come first in the list. The expectation for v_1 lies between these two bounds and can be approximated as an average of the two. In line with the rest of the literature on logical models of electoral system quantities, the geometric mean is preferred to the arithmetic mean,[5] so that

$$v_1 = \left(1 \times \frac{1}{c}\right)^{\frac{1}{2}} = c^{-\frac{1}{2}} \tag{1}$$

That is, the predicted fractional share for the first candidate in a list with c candidates is the inverse of the square root of c.

2.2 The Revised Model

This SBW model is indeed the 'best guess' for v_1, when this quantity is hypothesised to depend uniquely on c. Moreover, it is derived from 'hard' conceptual boundaries, which are *impossible* for v_1 to cross. Let us now relax these conditions by considering the role of two additional input variables: the number of preference votes and the expected number of seats at stake. Essentially, this new version of the model attempts to incorporate two intuitions in the derivation of v_1:

- The competitiveness of a list should increase in the number of preference votes that each voter can cast. Thus v_1 should decrease as voters are allowed to express more preferences (provided that they are non-cumulative).
- The competitiveness of a list should increase in the expectation of the number of seats attributed to each list. Thus v_1 should decrease as lists are expected to elect more candidates.

To model these assumptions, we proceed by replacing the hard conceptual bounds in equation 1 with some

[5]Logical models effectively approximate the *median* value of a quantity, returning predictions that are in expectation equally likely to be below or above its actual value. Geometric means between conceptually extreme cases express the median of distributions better than arithmetic means when the variable of interest can only take positive values, and therefore a normal distribution cannot be assumed. In particular, the geometric mean is preferable when the assumed distribution of a variable spans different orders of magnitude, as it abides by the principle of equal distortion: when we have no priors over two extreme possible values x_{max} and x_{min}, a 'best guess' that is equally likely to be above or below the true value is that both are off by the same multiplicative factor k, so that $x^* = (x_{max} \times x_{min})^{\frac{1}{2}}$. In the specific case of the coarse model for v_1, the arithmetic mean of conceptual boundaries would predict a first-ranked candidate share of 0.5 in the limit for increasing values of c which is an implausible degree of concentration for an extremely high number of candidates. For more details on the use of geometric means in logical models see Taagepera (2008, pp. 120–124).

'soft' upper and lower bounds, beyond which the quantity is *unlikely* to lie due to the expected effect of the additional input variables. Because these new boundaries are themselves derived from the geometric mean of conceptual bounds, the revised model will therefore be expressed in terms of 'predictions from predictions'.

2.2.1 The Number of Preference Votes (p)

Many preferential-list systems allow voters to express one or more non-transferable and non-cumulative[6] preferences for different candidates: this is the case in a few open-list (Cyprus, Greece, Italy, Kosovo, Peru, Sri Lanka) and many flexible-list systems (Belgium, Bosnia, Czechia, Slovakia). This factor is important because *the number of preferential votes at voters' disposal can be expected to compress the upper bound on the share of preference for the first party.* For instance, consider a case where voters *must* cast two preferences. Even if a candidate gets a vote from *all* electors, her share of preference votes will not cross $\frac{1}{2}$, as voters' second preference is spread across the other candidates in the list. To my knowledge, no preferential voting system employed for a parliamentary election requires voters to cast more than one preference vote; however, many *allow* such an option. In this case, the 'hard' upper bound for v_1 in a system where voters may cast from 1 to p preference votes is still 1: the scenario where all voters cast one preference and they all go to one candidate. But we may identify a more realistic – or at least, more useful – 'soft' upper bound by positing that, in such a system, voters will cast a number of preferential votes comprised between 1 and p. If all voters cast one vote, then the maximum fractional share the first-ranked candidate can obtain is still 1; if all voters cast p votes, then it is $\frac{1}{p}$. Taking the geometric mean of these boundaries, therefore our 'soft' upper bound for v_1 is $\frac{1}{p^{\frac{1}{2}}}$, or $p^{-\frac{1}{2}}$, where p is the maximum number of preference votes at voters' disposal. Of course, the logical standing of this boundary condition is only as good as our assumption of the number of preference votes cast by each voter: if they cast more, it should be lower; if they cast fewer, it should be higher. In this sense, it is a 'prediction from prediction', rather than a conclusion from pure deductive reasoning.

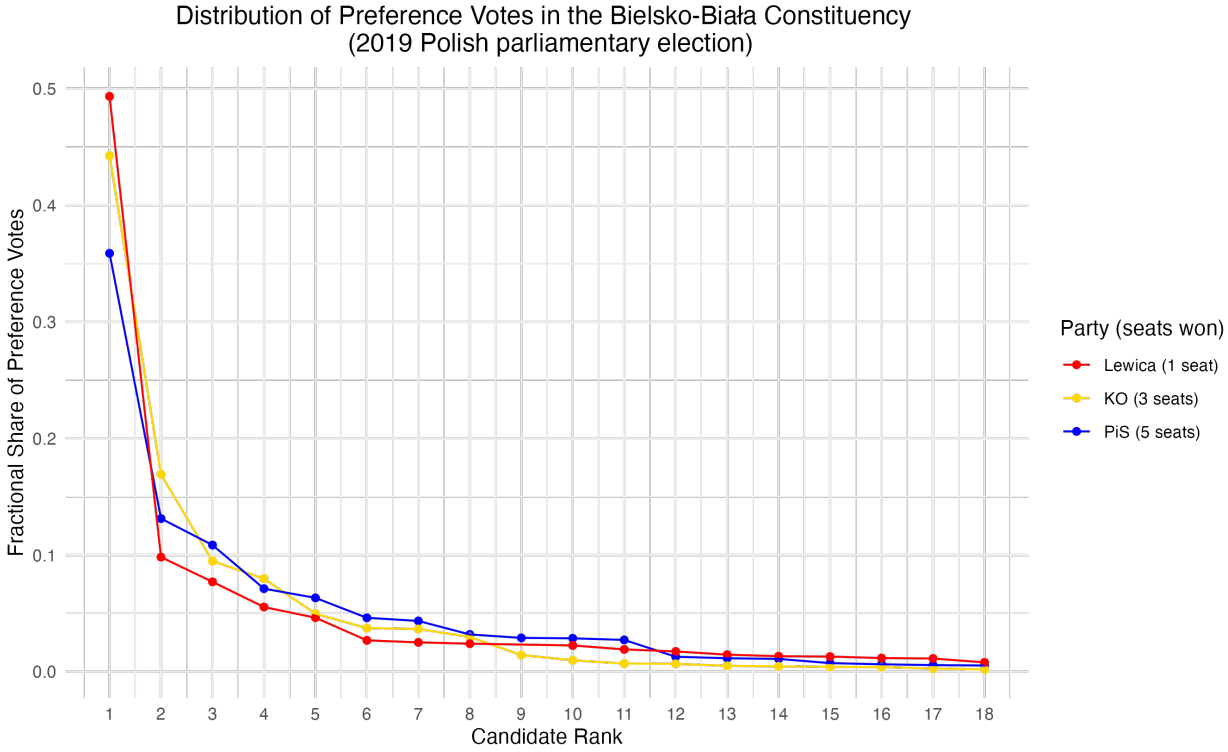
2.2.2 Expectations of Inter-party Performance

The second modification of the SBW model attempts to model the assumption that *intra-party competition in a list is endogenous to actors' expectation of the list's inter-party performance.* In other words, assuming that vote-seeking is costly, the more seats are expected to be assigned to a list, the more candidates will be

[6] There are rare cases that allow voters to express multiple preferences for the same candidate, such as in Switzerland, Luxembourg, and Hamburg's and Bremen's State Parliament electoral systems. Because non-cumulability of the vote is a crucial assumption at this stage, such preferential-list systems fall outside of the scope conditions of the model.

serious about seeking preference votes.[7] In a way, this assumption simply maps Duverger's intuition onto intra-party competition: just as the number of electorally competitive parties decreases with the district magnitude M, the number of electorally competitive candidates decreases in the expectation of the number of seats at stake s.

Figure 1: Preference votes in the Bielsko-Biała constituency (2019 Polish parliamentary election).



This point is perhaps best illustrated by peeking (temporarily) behind the veil of ignorance-based theorising and into real-world results. Figure 1 shows the distribution of preference votes for the three seat-winning parties competing in the Bielsko-Biała (Nr. 27) constituency in the 2019 Polish parliamentary election.[8] The first thing to note from a glance at the plot is that most candidates for all parties get close to zero percent of preference votes (not an uncommon pattern: see, for instance, Cheibub and Sin, 2020). Most names appearing on the ballot are effectively 'list fodder' or 'top-up candidates' (Arter, 2013), with little personal support or hope for election. This observation suggests that c, the key variable in the SBW model, can be a very volatile and imprecise predictor of intra-party competition: if the parties had only fielded 9

[7] Another way of thinking about this is from voters' perspective: the more seats are expected to be assigned to a list, the less voters will be concerned about wasting their preference on non-viable candidates.

[8] It can be shown empirically that v_1 is significantly lower for more successful seat-winning parties, discounting all district-level variation through fixed effects. See Appendix section A.

candidates instead of 18, the SBW model's prediction would change dramatically (the upper bound would double and v_1 would increase by a factor of $\sqrt{2}$), whereas in fact the last-placed 9 candidates account for only 10% of each party's total share of preference votes.

Secondly, the plot shows that the list that came on top in the district – the right-wing *Prawo i Sprawiedliwość* (PiS), which won five seats – has a lower first-candidate preference share of the vote and a flatter distribution of the vote than the other two, the liberal *Koalicja Obywatelska* (KO) and the left-wing *Lewica Demokraticzna* (Lewica). This observation is consistent with the idea that the number of electorally competitive candidates behaves with respects to the expectations of a list's performance in a similar way to how the number of electorally competitive parties behaves with respects to the district magnitude: the fewer spots available, the fewer serious contenders. (And, in turn: the fewer serious contenders, the more concentrated preferential votes will be towards the top-end of the candidate ranking.)

This intuition is perhaps the main innovative aspect of this paper – though Crisp, Jensen and Shomer (2007) made a similar observation – but its formalisation requires something of a leap of faith. How to model the effect of seat expectations on intra-party competition? My proposal is to introduce the concept of 'number of pertinent vote-earning candidates' c^* as a substitute for c, in order to formalise the idea that using simply the list length c will fail to differentiate how lists with higher or lower seat-winning potential will have different levels of internal competitiveness. Conceptually, c^* is an analogue of N_{v_0}, the number of pertinent vote-earning parties in a district (Shugart and Taagepera, 2017, p. 128): a quantity that denotes "how many [parties] are sufficiently important to contribute to our prediction." In the same way, c^* is a 'phantom quantity': it does not refer to an observable, but it helps to derive one, because it incorporates the idea that fewer candidates will seriously compete for votes when fewer seats are expected to be a stake. In this sense, a list will be *de facto* shorter when candidates expect there to be few seats at stake, and *de facto* longer when candidates are competing for more spots. Moreover, c^* is distinct from, say, the effective number of candidates, because we are not mainly interested in the size of their preference vote share: holding c constant, a candidate getting 1% of preference votes may be hopeless in a list competing for one seat, but has a realistic shot at a seat in a list competing for 15 spots in parliament.

Because expectations are crucial to these theoretical steps, the conceptual boundaries of the number pertinent vote-earning candidates c^* are identified in terms of actors' ability to infer list performance. On the one hand, candidates may be entirely in the dark about their electoral potential and the realistic number of seats allocated to their list. In this case, they all vie for votes, so that $c^* = c$, as in the SBW model. At the

other extreme, candidates may have perfect foresight about election results and their vote-winning potential, so that only the candidates who expect to end up in seat-eligible spots in the final rank order campaign and vie for votes.[9] In this case, where actors are perfectly efficient in choosing their level of campaign effort and have perfect priors over the number of seats the party will win, $c^* = s$, where s notates the number of seats won by the party in the district. Hence the expected number of pertinent vote-earning candidates in a list of size c gaining s seats can be computed as the geometric mean of s and c: $c^* = (sc)^{\frac{1}{2}}$.[10] It follows that the revised lower bound for v_1 is the inverse of c^*: considering only pertinent vote-earning candidates, at its lowest v_1 takes the value of $\frac{1}{c^*} = (sc)^{-\frac{1}{2}}$, when all the 'serious' candidates get the same share of the vote.

2.2.3 Bringing It All Together

Substituting the soft upper bound $p^{-\frac{1}{2}}$ and the soft lower bound $(sc)^{-\frac{1}{2}}$ into equation 1, we may therefore derive the revised prediction for the value of v_1:

$$v_1 = \left(\frac{1}{(sc)^{\frac{1}{2}}} \times \frac{1}{p^{\frac{1}{2}}}\right)^{\frac{1}{2}} = (scp)^{-\frac{1}{4}} \tag{2}$$

To recapitulate, equation 2 expresses that the fractional share of preference votes obtained by the first-ranked candidate v_1 is comprised between these two bounds:

— The candidate's share under the least-competitive scenario, where the the first candidate gets one preference from *all* voters, and voters on average cast $p^{\frac{1}{2}}$ preferences each, with p being the maximum number allowed. For instance for $p = 2$, 100 voters can be expected to cast a total of approximately 141 votes, so that the upper bound for v_1 is $100/141 \approx 0.71$ or equivalently $\frac{1}{2^{\frac{1}{2}}} \approx 0.71$.

— The candidate's share under the most-competitive scenario, where all pertinent vote-earning candi-

[9] It is not unrealistic that candidates have decent priors over their list's inter-party competitiveness; as Shugart and Taagepera (1989, p. 215) put it, "in a multi-seat district with a fairly stable voting pattern, the number of seats one particular party can obtain is known ahead of time within plus or minus one seat." Furthermore, candidates may have decent *ex ante* information about *their own* vote-winning potential as well, as their ballot position is strongly (and, in part, causally) associated with their chances (Lutz, 2010; Blom-Hansen et al., 2016; Devroe and Wauters, 2020; Van Erkel and Thijssen, 2016).

[10] To be exact, the model uses s, the observed number of seats won, as a proxy for $E(s)$, the unobservable number of seats actors in a list *expect* to win. This requires us to restrict the analysis to 'seat-winning' parties, discounting unsuccessful lists, where s is zero. c^* may be adjusted to be $[(s+1) \times c]^{1/2}$ – i.e. assuming that lists are at least marginally over-optimistic about their potential, otherwise they would not compete – to predict preference shares of non-seat-winning parties as well. A similar adjustment is proposed by Selb and Lutz (2015, p. 332): "we rely on the assumption that candidates' expectations are correct on average, and use the actual number of seats won by the candidates' list in the current election as a proxy for the expected number of seats. Moreover, to avoid substantial losses of observations due to zero divisions for all the lists that did not gain any seats, we will add one seat to the denominator [...] which implements the reasoning that any list which stands for election does so because its members expect at least one seat." However, to avoid the complication of introducing a further constant, the main model retains the simpler formula, and keeps the focus on seat-winning parties for reason of substantive interest and in line with previous theoretical work. Appendix section B.1 reproduces the analysis in section 6.1 with predictions based on a lower bound for c^* set to be $[c(s+1)]^{1/2}$ (on the same sample of seat-winning lists used the main analysis).

dates, a quantity endogenous to the 'length' of a list and its expected performance, get the same number of votes each. For instance, in a list with 25 candidates expected to gain 6 seats, the number of pertinent vote-earning candidates is $(25 \times 6)^{\frac{1}{2}} \approx 12.25$, and minimum value of v_1 is $\frac{1}{12.25} \approx 0.08$.

Thus, in this example, our best guess for v_1 in a list with 25 candidates winning 6 seats under PLPR rules allowing a maximum of 2 preference votes is $(2^{-\frac{1}{2}} \times (25 \times 6)^{-\frac{1}{2}})^{\frac{1}{2}} \approx (0.71 \times 0.08)^{\frac{1}{2}} \approx 0.24$. Note that, because $s \geq 1$ (the scope conditions are limited to seat-winning parties) and $c \geq p$ (voters cannot cast more votes than there are candidates), it follows that $(sc)^{-\frac{1}{2}} \leq p^{-\frac{1}{2}}$: the 'soft' upper bound may not logically be lower than the 'soft' lower bound.

2.2.4 Unmodelled Variables

Although a three-variable model is perhaps already complex enough to test Taagepera and colleagues' injunction to make logical models "as simple as possible, but no simpler" (Taagepera, Selb and Grofman, 2014, pp. 396-397, attributed to Albert Einstein), it obviously does not exhaust all possible sources of variation in intra-party competitiveness. Ultimately, whether and how well one can really predict average values of v_1 from variation in s, c and p is an empirical question. But it is worth briefly mentioning what has been left out of the picture.

First, the model makes no reference to the type of PLPR it applies to, the relevant distinction here being between open and flexible lists. While in the former all candidates are elected according to the order of preferences received, the latter require candidates to reach some preference threshold to be elected via preferences, while the rest of the seats are filled according to the party list position. In short, we simply have no directional prior as to how list type should matter, let alone a way to quantify such an effect. On the one hand, a flexible list system may decrease first-candidate vote share, if candidates listed in higher positions put less effort in attracting preferential votes, knowing they are likely to be elected regardless. On the other hand, a flexible list system might make the distribution of votes steeper, dissuading lower-positioned candidates from vying for votes, as those may be insufficient to clear the threshold.[11] Secondly, the model does not distinguish between contexts where preferential vote is mandatory and those where it is optional. Once again, there is no clear directional prior as to the effect of this variable, and effectively we should take a leap of faith and assume that the model predictions will apply *on average* across a sample of diverse

[11] It is however worth noting that, as flexible-list systems normally allow multiple preference votes, having included p as an input variable makes it more realistic that the scope conditions can be extended beyond simple open-list PR. The empirical analysis in section 6.1 includes separate tests of the models for open- and flexible-list elections.

institutional contexts. Factors falling outside the realm of electoral institutions – larger parties may be more factionalist, smaller parties may have fewer qualified candidates, outcomes are more easily predictable in stable party systems – are also beyond the reach and remit of a logical model.

3 Extending the Intra-party Model

This section develops the fundamental building block of the model, the equation $v_1 = (scp)^{-\frac{1}{4}}$, in two directions:

— In subsection 3.1, it is shown that two additional quantities of interest describing the degree of intra-party competitiveness of a list can be derived from v_1: the effective number of candidates (N_c) and the vote share for the last eligible candidate (v_s). First, I show how these quantities can be computed via an algorithmic iteration of the same procedure used to derive v_1 that derives predictions for preference shares for *all* candidates in a list. Then, I propose some approximations for these quantities that retain the simple X^k structure of the formula for v_1.

— In subsection 3.2, I recast the formula for v_1 – and the related equations for N_c and v_s – in terms of variables that are purely institutional, and can be gathered independently of outcomes realised on election day. This yields a 'results-blind', institutions-only model that makes predictions for average expected values of these quantities *at district level*, as opposed to the list-level prediction of the model in terms of s, c and p.

3.1 Deriving N_c and v_s

3.1.1 Algorithmic Approach

Just as we did with v_1, we may use the same process of individuating conceptual boundaries and taking their geometric mean to derive predictions for the values of $v_2, v_3, v_4 ... v_c$. These values would allow to derive N_c and v_s from the distribution of expected preference vote shares for *all* candidates. As we will see, this is algebraically messy. However, it is worth outlining how we may obtain these quantities with an iterative algorithm, as these predictions can serve as 'sanity checks' for the more synthetic approximations of N_c and v_s described in the rest of this section.

Let us start from v_2: the fractional share of preferential votes for the second-ranked candidate. Once v_1 is derived, we may use an analogous logic to derive the vote share of the second elected candidate, v_2. We

must, however, distinguish two cases.

- **Case 1** $v_1 \leq 0.5$

If the first-ranked candidate gets less than half of the preference votes, the upper bound for v_2 is v_1: the second candidate may not get more votes than the first candidate. The lower bound corresponds to the case where all the remaining pertinent vote-earning candidates get an equal share of the remaining vote share once v_1 is realised. The lower bound for v_2 is therefore $\frac{1-v_1}{(s\times(c-1))^{\frac{1}{2}}}$. The geometric mean of lower and upper bounds is the expected share of votes for the second candidate elected:

$$v_2 = \left(v_1 \times \frac{1-v_1}{(s\times(c-1))^{\frac{1}{2}}}\right)^{\frac{1}{2}} \tag{3}$$

- **Case 2** $v_1 > 0.5$

We must however consider the case in which the first candidate gets *more* than half of the preference votes. Such circumstance makes it illogical to posit v_1 as the upper bound for the second candidate: the second candidate's share never be as much, as the sum of first and second candidate shares would exceed 1. In this case, we must substitute in equation 3 the *actual* upper bound, which will be $1 - v_1$:

$$v_2 = \left((1-v_1) \times \frac{1-v_1}{(s\times(c-1))^{\frac{1}{2}}}\right)^{\frac{1}{2}} = \frac{1-v_1}{(s\times(c-1))^{\frac{1}{4}}} \tag{4}$$

Equations 3 and 4 can be generalised to the n^{th} candidate, whose share v_n is the geometric mean between the following two conceptual bounds:

— the upper bound is whichever is smaller of v_{n-1} (the share of the $(n-1)^{th}$ candidate) and $1-\sum_{i=1}^{n-1} v_i$ (the remaining share of the vote after v_{n-1} is realised). That is, the share of the vote for the n^{th} candidate is subject to the conditions that it cannot be greater than the share of the vote of the candidate placed above her and it cannot tip the total of preference shares above 1.

— the lower bound is $\frac{1-\sum_{i=1}^{n-1} v_i}{(s(c+1-n))^{1/2}}$, i.e. 1 minus the sum of $v_1, v_2, v_3 ... v_{n-1}$ over the number of pertinent vote-earning candidates $(s(c+1-n))^{\frac{1}{2}}$, where c is reduced by one each time one candidate share is realised. That is, the share of the vote for the n^{th} candidate is smallest when she gets the same preference votes as all the pertinent vote-earning candidates placed below her.[12]

[12]To compute values for candidates beyond c^* and allow the sum of all the fractional shares to converge to 1 in the limit, we need to hold s constant rather than using $(s+1-n)$ as we did with c. In this way, the number of candidates that are relevant to compute for second, third, fourth etc. place is reduced at each iteration – because the list gets shorter once each fractional share is realised – but never falls below zero.

The generalised algorithm to compute v_n $(2 \leq n \leq c)$ is therefore as follows:

$$v_n = \left(min\{v_{n-1}, 1 - \sum_{i=1}^{n-1} v_i\} \times \frac{1 - \sum_{i=1}^{n-1} v_i}{(s(c+1-n))^{1/2}} \right)^{\frac{1}{2}} \tag{5}$$

3.1.2 An Approximation of N_c

Having derived $v_1, v_2, v_3 ... v_c$ with the algorithm in equation 5, we may therefore compute the effective number of candidates N_c from its definition as $N_c \equiv \frac{1}{\sum_1^c ({v_i}^2)}$. This obviously does not simplify to a neat, generalisable expression in the form of X^k, or at least not one with a tractable value of the base. We may therefore attempt to use a shortcut to derive N_c from approximated conceptual boundaries expressed in terms of v_1. To distinguish this 'approximated' prediction for N_c from the value of the effective number of candidates obtained from iterating the algorithm in equation 5, I temporarily note the approximation as $\hat{N}_c$, with a hat.

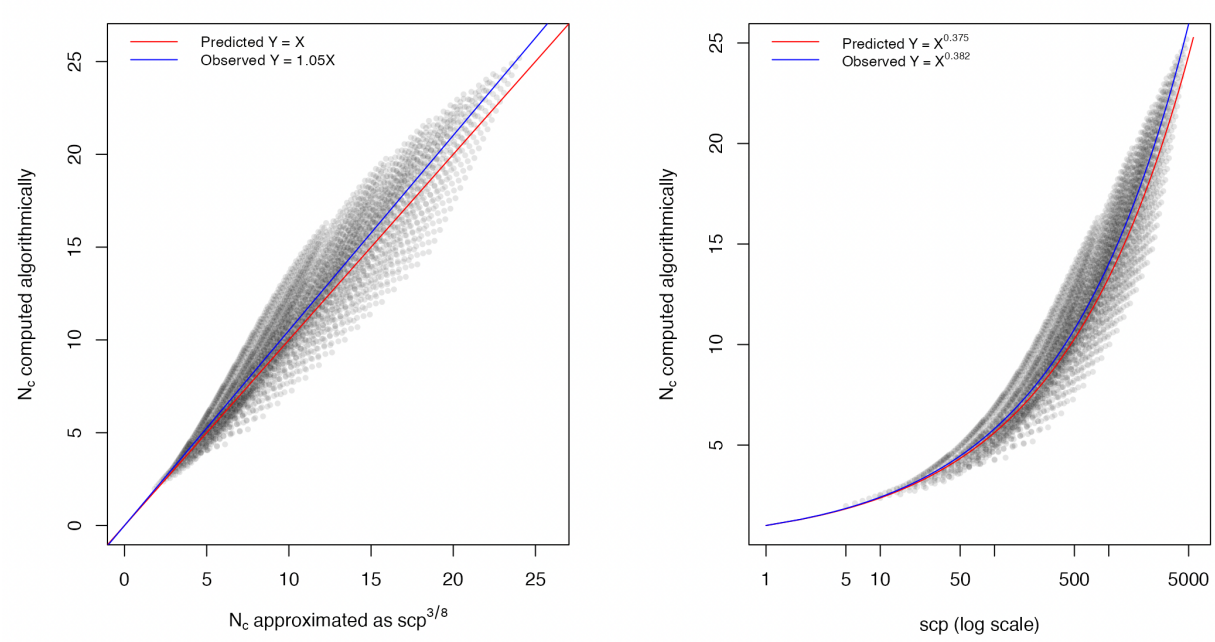
The value of the effective number of candidates N_c is at its greatest when all candidates share the remaining share of the vote after v_1 is realised equally. For a sufficiently large value of c, these small shares, corresponding to $\frac{1-v_1}{c-1}$ each, will become negligible when squared. Hence, the upper bound can be approximated as $\frac{1}{{v_1}^2+0^2+0^2...} = \frac{1}{{v_1}^2}$. The lower bound for v_1 corresponds to the case where the distribution of the vote is as 'compact' as possible: because the maximum a candidate can get is v_1, there will be $\frac{1}{v_1}$ candidates getting v_1 share of the vote each. It follows that the lower bound will be approximately $\frac{1}{\frac{1}{v_1} \times v_1^2} = \frac{1}{v_1}$.[13] Taking the geometric means of these conceptual bounds and substituting our v_1 equivalence from equation 2, we obtain the following approximation:

$$\hat{N}_c = \left(\frac{1}{v_1} \times \frac{1}{{v_1}^2} \right)^{\frac{1}{2}} = {v_1}^{-\frac{3}{2}} = (scp)^{\frac{3}{8}} \tag{6}$$

It can be shown graphically that this estimation approximates quite well the value of the effective number of candidates N_c computed via algorithmic iteration, at least for realistic values of s, c and p. I simulated data with all combinations of c comprised between 5 and 40, s comprised between 1 and 20, and p comprised between 1 and 6. After removing the 'illogical' cases where $s > c$ or $p > c$, I computed N_c with the algorithmic method for all combinations of values and $\hat{N}_c$ with the approximation in equation 6. The graph on the left in figure 2 shows how the two estimates compare. Fitting a fixed-intercept linear regression $(Y = \beta X)$ returns a linear relationship between the two corresponding to $N_c = 1.05(\hat{N}_c)$, which is reasonably close

[13] A similar procedure to approximate the effective number of parties is in Taagepera and Shugart (1993).

Figure 2: Comparison of N_c estimates.



to the expected identity $N_c = \hat{N}_c$. The graph on the right in figure 2 shows the relationship between the 'raw' measure of the *scp* product and the algorithmic estimation of the effective number of candidates N_c. Modelling the relationship as a fixed exponent regression ($Y = X^{\beta}$) returns the function $N_c = scp^{0.382}$, which again is closely in line with our approximation $\hat{N}_c = scp^{\frac{3}{8}}$. To be sure, this does not mean the approximation for the effective number of candidates is 'true': it simply means that it is consistent with the generative process assumed to be behind the estimation of v_1 and extended via algorithmic iteration to $v_n \mid n \geq 2$.

3.1.3 An Approximation of v_s

A free-standing, empirically useful formula for v_s has remained elusive to logical modellers, to the extent that the two existing attempts at deriving this quantity (Shugart, Bergman and Watt, 2013; Shugart and Taagepera, 2017, pp. 226-235) ultimately recur to empirically derived constants to adjust the models. In this section, I propose a new approach to derive a 'shortcut' approximation for v_s. The method proposed is somewhat more convoluted than the approximation of N_c and involves something of a mathematical sleight of hand in positing the mean and median of a quantity to be approximately equal in expectation. Nonetheless, as it will be shown, the deriving formula, when expressed in purely institutional terms without empirical

input, performs respectably well on both simulated and real-world data.

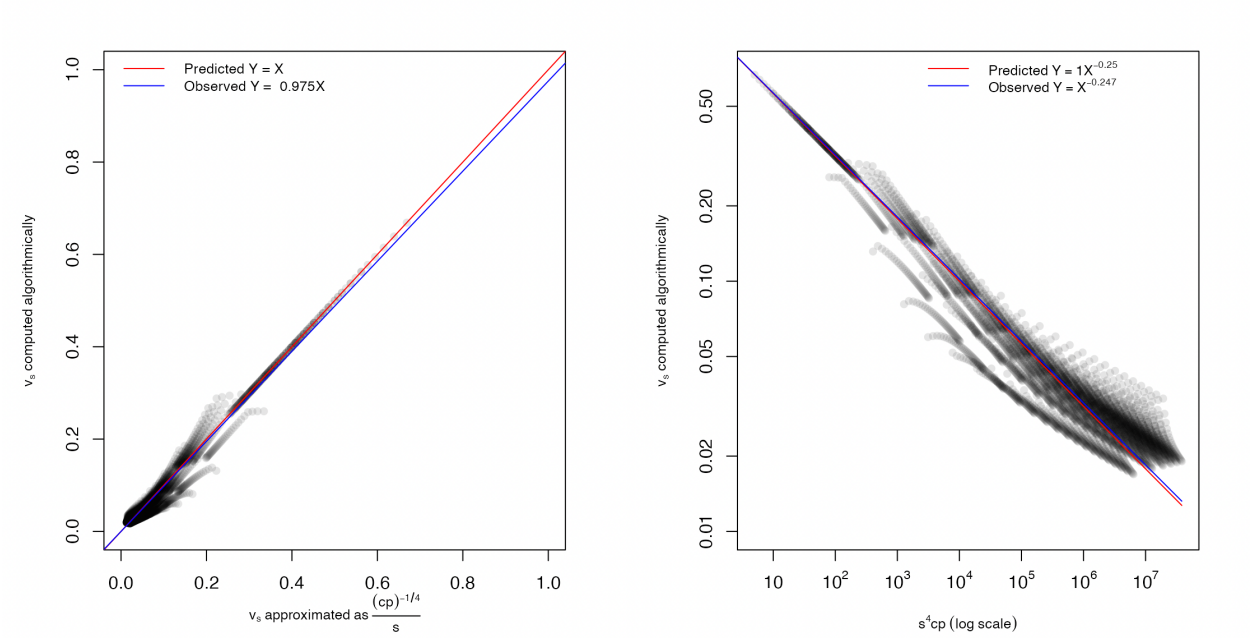
Let us start by considering the *expected share for a seat-eligible candidate*, $v_{i\leq s}$. We can estimate such a quantity as a function of v_1 (in expectation) in two ways: either as the expected median of preference shares of seat-eligible candidates (notated with a tilde $\tilde{v}_{i\leq s}$) or as the expected mean of preference shares of seat-eligible candidates (notated with a bar $\bar{v}_{i\leq s}$). First, relying on Taagepera's argument that the geometric mean of conceptual bounds approximates a median, we can derive the expected median preference vote for a seat-eligible candidate $\tilde{v}_{i\leq s}$ as a quantity comprised between v_1 and v_s, and therefore $\tilde{v}_{i\leq s} = (v_1 \times v_s)^{\frac{1}{2}}$. Let us now repeat the conceptual boundaries logic for the mean preference share for a seat-eligible candidate $\bar{v}_{i\leq s}$, starting from a computation of upper and lower bounds for the mean as functions of the realised value of v_1. The mean preference share is at its highest when $s = 1$, the list only wins one seat and the mean is v_1: hence, $(1 \times c \times p)^{-\frac{1}{4}}$ is the upper bound. The lower bound is realised when all candidates after v_1 up to v_s get as few preference shares as possible: for values of c sufficiently larger than s, the value of each individual one of these $\frac{1-v_1}{c-1}$ shares will tend to zero, so that the expected mean preference vote will be approximately $\frac{v_1}{s}$. We can thus proceed to derive the mean preference vote of seat-eligible candidates in expectation as $\bar{v}_{i\leq s} = ((cp)^{-\frac{1}{4}} \times \frac{v_1}{s})^{\frac{1}{2}}$.

An approximation for v_s, notated as $\hat{v}_s$, can be derived algebraically by positing that the median ($\tilde{v}_{i\leq s}$) and mean ($\bar{v}_{i\leq s}$) share of a seat-eligible candidate's preference votes are approximately equal:

$$\begin{aligned} \tilde{v}_{i\leq s} &\approx \bar{v}_{i\leq s} \\ (v_1 \times v_s)^{\frac{1}{2}} &\approx \left((cp)^{-\frac{1}{4}} \times \frac{v_1}{s}\right)^{\frac{1}{2}} \\ v_s &\approx \frac{(cp)^{-\frac{1}{4}}}{s} \\ \hat{v}_s &= (s^4 cp)^{-\frac{1}{4}} \end{aligned} \tag{7}$$

Let us now repeat the same 'sanity checks' for $\hat{v}_s$ as we did for $\hat{N}_c$ (graphically, in figure 3). Regressing the algorithmically derived value of the share for the last eligible candidate v_s on the approximation $\hat{v}_s$ obtained from equation 7, we obtain $v_s = 0.975\hat{v}_s$, which is close to an identity. The exponential function linking v_s to s^4cp is $v_s = (s^4cp)^{-0.247}$, which is close to the the approximation's prediction $\hat{v}_s = (s^4cp)^{-\frac{1}{4}}$. Even so, the theoretical derivation of this approximation is admittedly not as rigorous or satisfying as we would wish, as it relies on the assumption that two quantities that we *know* to be distinct are approximately

Figure 3: Comparison of v_s estimates.



equal. In particular, the comparability of the two measures of central tendency may be limited when parties gain many seats, and thus the distribution of preferences for seat-eligible candidates becomes more skewed. There are hints of this bias in figure 3, which suggests that for very high values of the base product (low values of v_s), the approximation somewhat over-estimates the quantity relative to the algorithm's prediction.

3.2 Predicting v_1, N_c and v_s from Institutional Variables

The equations $v_1 = (scp)^{-\frac{1}{4}}$, $N_c = (scp)^{\frac{3}{8}}$ and $v_s = (s^4cp)^{-\frac{1}{4}}$ are 'list-level' models insofar as they predict different values for observations in the same district, varying from list to list as a function of their different values of s and c (conversely, p is constant in a district).[14] However, as s and c are only realised at election time, these models do not fully qualify as *pre*-dictive: they can only yield predictions after candidate lists are presented by parties and inter-party seat allocation is realised. To be of use to institutional engineers, we need to recast these models in terms of average expected indicators of intra-party competitiveness that vary as a function of variables that pre-exist the selection and election of candidates. In particular, as it will be argued shortly, both s and c can be expressed as some function of the district magnitude. In doing

[14] I henceforth refer to $\hat{N}_c$ and $\hat{v}_s$, the approximations of the quantities, simply as N_c and v_s, as the algorithm-based prediction for these quantities are not relevant to the rest of the analysis.

so, the models will shift from the list level to the district level: if we do not know the exact values of c and s for each list, we may only express these quantities in terms of their expected value for *any* seat-winning list in a district. In practice, this is the same as deriving the value of a quantity for the expected median seat-winning party, as this prediction would thus be equally likely to fall below or above the real quantity for any party.

First, let us consider how c is related to electoral institutions. How many candidates can a seat-winning list be expected to field in a district where M seats are at stake? Here we need to make a further assumption based on observed empirical regularities: *lists nominate as many candidates as they are legally allowed to*. This is because, unlike under STV and SNTV, preferential-list systems *pool* individual candidates' votes into list totals, which in turn determine the inter-party allocation of seats. Thus, the returns to additional candidate nominations are always positive for parties, as there is no risk of 'wasting' votes through over-nomination error (Shugart, Bergman and Watt, 2013; André, Depauw and Deschouwer, 2014). Legal limits to the number of candidates vary from electoral system to electoral system. In countries like Italy, Cyprus and Belgium, the number of candidates in a party-list is capped to the number of seats at stake in a district, so that $c = M$. Elsewhere, they might nominate up to $M + 2$ (Estonia) or $2M$ (Poland). Therefore, in general, $c = rM$, where r is the ratio between the maximum number of candidates allowed and the district magnitude, varying at district level. Having assumed that seat-winning lists always over-nominate, our 'over-nomination ratio' parameter r is thus a purely institutional measure, albeit one that is often hidden in obscure electoral regulations.[15]

Secondly, let us consider how s may be expressed as a function of institutional quantities. In this case, we once again employ the prediction-from-prediction approach to derive the average (in this case, median) expected number of seat for any seat-winning party in terms of M. The work of Taagepera and Shugart (1993) shows that the number of seat-winning parties N'_0 in a district is $M^{\frac{1}{2}}$: the geometric mean between the minimum of 1 and the maximum of M. From here, they derive the fractional share of seats σ'_1 won by the largest party *in a district* as the geometric mean between the maximum of 1 and the minimum of $\frac{1}{N'_0}$, returning $\sigma'_1 = M^{-\frac{1}{4}}$. The expected number of seats won by the largest party is therefore $M \times \sigma'_1 = M^{\frac{3}{4}}$.[16]

[15] This assumption holds up well empirically for most seat-winning parties (see section 4). Very small lists may be short of personnel and thus be forced to under-nominate; but as our focus is on seat-winning parties, these lists are unlikely to be relevant. A thornier problem is represented by countries with particularly complex over-nomination rules. For instance, in Brazil, each *individual party* in a *coalition* list may nominate up to M candidates. This means not only that r varies as a function of 'political' factors, rather than simply institutional ones, but also that some parties may prefer to under-nominate to concentrate their preference votes within the coalition. Such a case is therefore not tractable in an institutions-only model.

[16] In a more recent version of their predictive model of district-level party shares (Shugart and Taagepera, 2017, 153–180), they introduce a constant k to adjust the exponent of σ'_1 (and the deriving quantities) for the 'embeddedness' of districts in the broader political system. The deriving model predictions for intra-party quantities and their performance are very similar

It follows that the expected number of seats for *any* seat-winning party is the geometric mean between the minimum of 1 and the maximum of $M^{\frac{3}{4}}$, i.e. $M^{\frac{3}{8}}$.

Substituting rM for c and $M^{\frac{3}{8}}$ for s, we obtain district-level predictions for our quantities of interest in terms of M, r, and p. These should predict the expected variation in the dependent variables across *any* seat-winning list.

$$v_1 = (scp)^{-\frac{1}{4}} = (prM^{\frac{11}{8}})^{-\frac{1}{4}} \tag{8}$$

$$N_c = (scp)^{\frac{3}{8}} = (prM^{\frac{11}{8}})^{\frac{3}{8}} \tag{9}$$

$$v_s = (s^4cp)^{-\frac{1}{4}} = (prM^{\frac{5}{2}})^{-\frac{1}{4}} \tag{10}$$

Of course, we need not pass by the list-level model to derive these district-level predictions (although it is a useful illustrative step). For instance, equation 8 can be obtained as the geometric mean of the upper bound $p^{-\frac{1}{2}}$ and the lower bound $(M^{\frac{3}{8}} \times rM)^{-\frac{1}{2}}$. The lower bound is in turn $\frac{1}{c^*}$, i.e. the inverse of the geometric mean between rM, the maximum number of candidates a list can field, and $M^{\frac{3}{8}}$, the expected number of seat-winners for the median seat-winning list, denoting the minimum number of candidates that have an incentive to compete for votes in such a list.

4 Data

To test the models, I collected preference shares for candidates at the district-list-election level for 31 Lower Chamber elections in nine preferential-list PR systems, including both open- and flexible-list systems.[17] Preference shares for each candidate in a list are defined as a candidate's share of all preferences cast for the party in the district they run in. From these, I coded the actual values of v_1, N_c and v_s for each list-district-election observation: these serve as the dependent variables in the empirical test of the models.

All these elections were conducted under PLPR rules, and there was no major institutional change over the period of time considered. Table 1 summarises some key inter-party and intra-party characteristics of

using this more complex value: see Appendix section B.2.

[17] I excluded results from two single-member districts: Aosta, where $c = 1$ and thus is effectively not a preferential-list system; and Åland, where the data available does not allow to isolate which candidates belong to which lists (all are listed as 'Other' as the Swedish-speaking region has its own party system).

Table 1

General Information

Country	Years	Assembly Name	Elections	Districts	Size	list type	Pref. Vote
Belgium	2003-19	Chambre des Représentants	5	11	150	flexible	optional
Cyprus	2011-21	Voulí ton Antiprosópon	3	6	56	open[a]	optional
Czechia	2013-21	Poslanecká Sněmovna	3	14	200	flexible	optional
Estonia	2011-19	Riigikogu	3	12	101	flexible	mandatory
Finland	2011-19	Eduskunta	3	12	200	open	mandatory
Italy	1976-92	Camera dei Deputati	5	31	630	open	optional
Peru	2014-21	Congreso de la República	3	26[b]	130	open	optional
Poland	2011-19	Sejm	3	41	460	open	mandatory
Slovakia	2012-20	Národná Rada	3	1	150	flexible	optional

Seat Distribution Rules

	Inter-party dimension		*Intra-party dimension*		
Country	PR Formula	Party Thresh.	Max. No. Cands.	Preference Thresh.	Maximum Votes
Belgium	D'Hondt	—	M	quota-based[c]	M
Cyprus	Hare	3.6%	M	—	$M/4$ (rounded up)
Czechia	D'Hondt	5%	varies by district	5% of party votes	4
Estonia	D'Hondt[d]	5%	M+2	quota-based[c]	1
Finland	D'Hondt	—	M or 14 (for $M \leq 14$)	—	1
Italy	Imperiali[d]	—	M	—	3 or 4 (for $M \geq 15$)
Peru	D'Hondt	5%	M or 3 (for $M \leq 3$)[e]	—	2
Poland	Sainte-Laguë	5%[f]	$2 \times M$	—	1
Slovakia	Hag.-Bischoff	5%	M	3% of party votes	4

Notes: (a) In Cyprus, the vast majority of MPs are elected via open-list PR. However, party leaders are elected automatically from the list they run in. But this applies only to 5-6 candidates per election-year, so while technically flexible, the system might be considered virtually open (Passarelli, 2020, pp. 92-93). (b) In the 2021 election, there were 27 districts, as the district reserved for voters resident abroad was separated from the Lima district. (c) In Belgium, the threshold is equal to the party's Hagenbach-Bischoff quota; in Estonia, the threshold is 10% of the Hare quota. See Passarelli (2020, pp. 88-9 and 96-97) for details. (d) In Italy and Estonia, compensation mandates are allocated to party and districts via an upper tier. (e) The nomination limit was raised to 4, for districts electing fewer than 4 seats, in 2021. (f) The representation threshold in Poland is 5% for single-party lists, but 8% for coalitions.

each electoral system. As discussed, many institutional differences on either dimension are not explicitly modelled: the assumption, or rather the 'wager', is that actors will behave similarly in list-based systems that are sufficiently proportional and sufficiently preferential. The cases were mainly selected due to considerations of data availability; nonetheless, the sample contains a diverse range of PLPR institutional set-ups:

— Five OLPR systems (Cyprus, Finland, Poland, Italy, Peru) and four FLPR systems (Czechia, Slovakia, Belgium, Estonia).

— Three countries where voters *must* cast a preferential vote (Estonia, Finland, Poland) and six where they may do so, or otherwise cast only a list vote (Belgium, Cyprus, Czechia, Italy, Peru, Slovakia).

— Four countries where parties cannot nominate more candidates than the district magnitude (Belgium, Cyprus, Italy, Slovakia) and five where they can over-nominate (Czechia, Estonia, Finland, Peru, Poland).

— Three simple PR systems on the inter-party dimension (Belgium, Finland, Italy) and six where proportionality is corrected via the introduction of preference thresholds (Cyprus, Czechia, Estonia, Peru, Poland, Slovakia).

— Six countries employing a divisor formula (Belgium, Czechia, Estonia, Finland, Peru), and three using a quota formula (Cyprus, Italy, Slovakia).

— Six countries allowing multiple preference votes (Belgium, Cyprus, Czechia, Italy, Peru, Slovakia) and three restricting voters to one preference vote (Poland, Finland, Estonia).

Table 2: Median values of the variables of interest (maxima and minima in parentheses).

Dependent Variables

Country	Observations	v_1	N_c	v_s
Belgium	270	0.28 (0.12–0.62)	7.87 (2.38–14.99)	0.11 (0.02–0.59)
Cyprus	84	0.25 (0.11–0.65)	6.74 (2.07–15.10)	0.21 (0.04–0.59)
Czechia	241	0.16 (0.09–0.48)	13.18 (4.12–22.58)	0.11 (0.02–0.35)
Estonia	160	0.44 (0.17–0.94)	3.82 (1.13–8.22)	0.32 (0.03–0.85)
Finland	230	0.23 (0.09–0.9)	8.04 (1.22–23.09)	0.14 (0.02–0.90)
Italy	906	0.26 (0.07–0.96)	7.61 (1.09–27.31)	0.18 (0.01–0.96)
Peru	217	0.40 (0.15–0.75)	3.40 (1.69–15.54)	0.36 (0.01–0.75)
Poland	489	0.35 (0.09–0.90)	5.72 (1.23–18.26)	0.17 (0.01–0.73)
Slovakia	20	0.26 (0.17–0.39)	9.39 (5.82–13.02)	0.01 (0.001–0.02)

Independent Variables

Country	c	s	p	M	r
Belgium	16 (4–24)	2 (1–11)	16 (4–24)	15 (3–24)	1 (1–1)
Cyprus	11 (3–20)	1 (1–7)	3 (1–5)	11 (3–20)	1 (1–1)
Czechia	22 (14–36)	2 (1–11)	4 (4–4)	12 (5–26)	1.8 (1.31–2.8)
Estonia	10 (7–17)	2 (1–6)	1 (1–1)	8 (5–15)	1.2 (1.13–1.4)
Finland	17 (2–36)	2 (1–11)	1 (1–1)	17 (6–36)	1 (1–2.33)
Italy	20 (3–54)	2 (1–20)	4 (3–4)	21 (2–55)	1 (1–1)
Peru	5 (3–36)	1 (1–15)	2 (2–2)	5 (1–36)	1 (1–4)
Poland	23 (13–40)	2 (1–12)	1 (1–1)	12 (7–20)	2 (2–2)
Slovakia	150 (148–150)	15.5 (10–83)	4 (4–4)	150 (150–150)	1 (1–1)

Turning now to the observed values of the variables of interest, table 2 reports median, maxima and

minima for the dependent (v_1, N_c, v_s) and independent (c, s, p, r, M) variables, computed at list level and broken down by country. A descriptive analysis of list-level variables is also reassuring with regards to the assumption that parties always over-nominate: the mean ratio of c/M across the whole sample is 1.30, while the mean value of r is 1.31, meaning that the institutional limit to nominations is effectively tantamount to the number of candidates nominated by seat-winning parties. Even in Poland, where r is highest as lists may nominate up to $2M$ candidates (and in theory they may field as few as $M/2$), on average seat-winning lists nominate a number of candidates that is 1.97 the district magnitude.

5 Methodology

Restating the conclusions of the theoretical section, I derived six quantitative predictions in the form $Y = X^k$ linking products of electoral quantities to indicators of intra-party competitiveness: one for each of the three dependent variable in the list-level model, and one for each of the three dependent variable in the district-level institutions-only model. These are effectively our hypotheses: given a product of electoral system quantities as the base X of the function, the exponent k is expected to take a certain value derived theoretically (in practice, $-\frac{1}{4}$ or $\frac{3}{8}$). I notate the value of the exponent expected from theory as $\hat{k}$ and the value obtained from regression as β in the rest of the analysis. Summing up:

Y variable	X (list-level)	X (district-level)	$\hat{k}$ (expected slope)
v_1 (first candidate's share)	scp	$prM^{\frac{11}{8}}$	$-\frac{1}{4}$ or -0.25
N_c (eff. number of candidates)	scp	$prM^{\frac{11}{8}}$	$\frac{3}{8}$ or 0.375
v_s (last eligible cand. share)	s^4cp	$prM^{\frac{5}{2}}$	$-\frac{1}{4}$ or -0.25

The empirical section proceeds in two steps:

— In subsection 6.1, I employ regression analysis to test the bias of the slope predictions $\hat{k}$ in the table above. Moreover, for v_1 and v_s, I compare the revised intra-party models' performance to their equivalents in SBW. The bias of the models is measured here as the discrepancy between the expected and observed slopes, normalised by the standard error.

— In subsection 6.2, I compare the precision of the individual predictions of the models of intra-party quantities with the better established predictions of their inter-party analogues of the seat-product model (SPM).[18] Specifically, the models for the fractional share of first-ranked candidates in a district (v_1) are compared with the SPM predictions for the fractional share of seats for the largest party in an assembly (σ_1); the models for the effective number of candidates (N_c) are compared with the SPM predictions for the effective number of parliamentary parties (N_S). The precision of the predictions is measured by computing the deviation-from-prediction of each observation, using the d index proposed by Nemčok and Šedo (2018) and explained shortly in subsection 5.2. The d index is computed for (1) the values of v_1 and N_c of each list compared to the list-level model's predictions, (2) the median

[18] Bias and precision are here used to denote different aspects of model performance. Bias indicates error in *the extent to which a model describes the overall relationship between variables in the sample*, while precision indicates *the extent to which the model can predict individual observations in the sample*. The latter is therefore a summary measure of bias and variance, and applies to observations rather than samples.

values of v_1 and N_c for each district compared with the district-level model's predictions, and (3) the values of σ_1 and N_S for each assembly election compared with the SPM's predictions.

5.1 Model Bias and Comparison with the SBW Model

In the first empirical section, I present graphical summaries of each of the six hypotheses, where the predicted and observed functional forms of the regressions are plotted against the data. Moreover, I report the exponential slope coefficients obtained by performing regression analysis on the full sample, only on the sample of elections contested under OLPR rules, and only on the sample of elections contested under FLPR rules. As a measure of coefficient bias, I compute for each of these the absolute difference between the observed slope β and predicted estimate $\hat{k}$ normalised by the standard error of the estimate (standard errors are clustered at the election-district level).

To test the models, I follow the methodological recommendations in Taagepera (2008) and employ fixed-intercept exponential regressions where the relationship between variables is modelled as $Y = X^{\beta}$.[19] This is equivalent to a log-log regression where the intercept is set to be zero, so that fitted values may not exceed $X^0 = 1$. Regression parameters are thus constrained to predict positive values of the dependent variable, lying between 0 and 1 when k is negative and larger than 1 when k is positive. Moreover, the fixed-intercept serves an 'anchor point' (Taagepera, 2008, pp. 44-45) that prevents us from making illogical prediction: we *know* a priori that in the limit where a PLPR contest collapses into a single-member district plurality race, $c = s = p = scp = 1$ and all three quantities v_1, N_c and v_s will be 1. In other words, the seat-winning party will field only one candidate, who will come 'both first and last' with 100% of the votes. Deriving an empirical intercept other than 1 would predict an absurdity.

In the presentation of the results for v_1 and v_s, I also compare the model performance against the existing predictions for these quantities in SBW.[20] Their estimation of v_1 in open-list systems, as discussed, is

$$v_1 = c^{-\frac{1}{2}} \tag{11}$$

As for v_s, the SBW model's expectation is that in an open-list PR system the share of preference votes of the last elected candidate is comprised between v_1 and the last elected candidate's share under SNTV rules, which is separately estimated as c^{-1}. Hence, in principle,

[19] The estimator used is simple OLS. In the appendix section B.3, I present the results of the same tests employing 'symmetric regression', a different estimator favoured by Taagepera (2008). Moreover, I present slope estimates dropping one country from the sample, to show that the predictive accuracy is not primarily driven by any one institutional set-up (section B.4).

[20] To my knowledge, there are no existing logical models for N_c.

$$v_s = (c^{-\frac{1}{2}} \times c^{-1})^{\frac{1}{2}} = c^{-\frac{3}{4}} \tag{12}$$

However, SBW prefer to employ the *observed* value of the exponent for v_1 instead of the 'ignorance-based' value of $-\frac{1}{2}$, so that after observing the empirical value of the slope β, where $v_1 = c^{\beta}$, v_s is estimated as

$$v_s = (c^{\beta} \times c^{-1})^{\frac{1}{2}} = c^{\frac{-1+\beta}{2}} \tag{13}$$

Both versions – which I call respectively the 'uncorrected' and 'corrected' SBW models – are computed and compared to my model's predictions for v_1. The extent to which the observed relationship between c and v_s deviates from these predictions is again calculated by taking the absolute value of the difference divided by the standard error.

5.2 Model Precision Compared to Inter-Party Models

The final empirical section addresses the question posed at the very beginning of the paper: *to what extent can we make predictions about intra-party competition in the same way as the seat-product model does with respect to inter-party competition?* To do so, I compare the list- and district-levels models for v_1 and N_c with their inter-party analogues:[21] respectively, the fractional share of seats of the largest party (σ_1) and the effective number of parties in an assembly (N_S). As noted at the start of the paper, the inter-party quantity predictions of the seat-product model consist of functions of the product of district magnitude (M) and assembly size (S). The formulas for σ_1 and N_S given in Taagepera (2007) and Shugart and Taagepera (2017) are, respectively:

$$\sigma_1 = (MS)^{-\frac{1}{8}} \tag{14}$$

$$N_S = (MS)^{\frac{1}{6}} \tag{15}$$

Following Nemčok and Šedo (2018), in this part of the analysis I compute for each observation a measure of discrepancy d as $\log_{10}(\frac{y}{\hat{y}})$, where y is the observed value and $\hat{y}$ is the prediction. The logarithmic function of the ratio between observed and predicted values takes the value of 0 when the observation perfectly mirrors

[21] There is no obvious inter-party counterpart to v_s. An analysis of deviation-from-prediction of the v_s model analogous to the one conducted for v_1 and N_c is presented in the appendix, without comparison to inter-party quantities.

theoretical expectations, positive values when the observed quantity is larger than expected and negative values when it is smaller. More specifically, d expresses the factor by which the prediction is off: if the observed value is twice the predicted value, d will be $\log_{10}(2) \approx 0.3$; if it is half the predicted value, d will be $\log_{10}(0.5) \approx -0.3$. Again following Nemčok and Šedo (2018), I use these two values as the bounds of 'tolerable' deviation, and compute the percentage of cases that fall in between, and thus are predictable from input variables within a factor of 2. I also report measures of central tendency for $|d|$, from which it is possible to extrapolate the average factor of error of the model predictions, as $(10^{|d|} - 1) \times 100\%$. For instance, predictions that are twice or half the observed value will both have $|d| \approx 0.3$, and the factor of error will be $(10^{0.3} - 1) \times 100\% = (2 - 1) \times 100\% = 100\%$.

For the intra-party model, I present two sets of results: first, I compute d for v_1 and N_c for each of the 2,617 list-in-district observations in the intra-party dataset described in section 4, using the predictions of the list-level model. Then, I compute d for the *median values* of v_1 and N_c in each of the 549 districts in my data, using the predictions of the district-level model.[22] As discussed, the district-level predictions will take the same value for each list, and they are meant to capture the level of competitiveness for any list in expectation. For the inter-party analysis, I replicate and extend the analysis in Nemčok and Šedo (2018), using the country-election level data they gathered on institutional characteristics and election results in 560 democratic elections.[23]

Therefore, I compute the index d of six quantities: the list-level observed intra-party quantities v_1 and N_c; the district-level median intra-party quantities $\tilde{v}_1$ and $\tilde{N}_c$; and the election-level inter-party quantities σ_1 and N_S. By the definition of d, the formulas are as follows:

Intra-Party Quantity (List-Level)	d index	Intra-Party Quantity (District-Level)	d index	Inter-Party Quantity (Election-Level)	d index
v_1	$\log_{10}\left(\frac{v_1}{(scp)^{-\frac{1}{4}}}\right)$	$\tilde{v}_1$	$\log_{10}\left(\frac{\tilde{v}_1}{(prM^{\frac{11}{8}})^{-\frac{1}{4}}}\right)$	σ_1	$\log_{10}\left(\frac{\sigma_1}{(MS)^{-\frac{1}{8}}}\right)$
N_c	$\log_{10}\left(\frac{v_1}{(scp)^{\frac{3}{8}}}\right)$	$\tilde{N}_c$	$\log_{10}\left(\frac{\tilde{N}_c}{(prM^{\frac{11}{8}})^{\frac{3}{8}}}\right)$	N_S	$\log_{10}\left(\frac{N_S}{(MS)^{\frac{1}{6}}}\right)$

[22] In section 6.2, I use the median values, as logical models are meant to yield predictions that are equally likely to be above or below the real values. In section E of the Appendix, I repeat this exercise with the mean values of the variables. Medians of v_1 and N_c are notated as $\tilde{v}_1$ and $\tilde{N}_c$, while means are notated as $\bar{v}_1$ and $\bar{N}_c$.

[23] The Nemčok-Šedo dataset of electoral quantities is described in the Appendix, section C.

6 Empirical Tests

6.1 Model Bias and Comparison with the SBW Model

6.1.1 First-ranked Candidate Share (v_1)

Regressing the observed values of v_1 on *scp* returns the fixed-intercept exponential function of $Y = X^{-0.262}$ with the coefficient having a 95% confidence interval $(-0.254, -0.270)$. The list-level prediction $\hat{k} = -0.25$ therefore falls just outside the confidence interval, narrowly over-predicting first-candidate share. As for institutions-only model, the regression of v_1 on the product $prM^{\frac{11}{8}}$ returns the fixed-intercept exponential function $Y = X^{-0.253}$ with the coefficient having a 95% confidence interval (-0.260, -0.246), therefore including the prediction $\hat{k} = -0.25$.

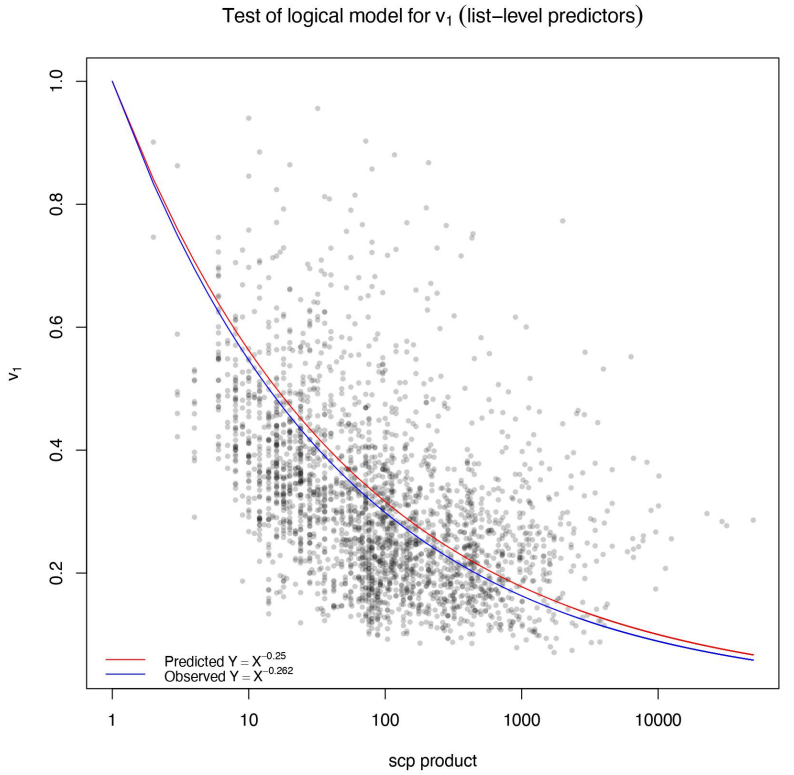
Figure 4 plots the predicted and observed exponential functions for the two models. In a pattern that we will observe across model tests of v_1 and N_c, the Slovak lists on the far right of the plots, which take the highest values of the product as they refer to parties competing in a district of magnitude 150, have substantially higher values than predicted. While the institutions-only prediction presents even less bias than the already rather accurate list-level model, it is clear from the plot there is much more scatter around it, as the independent variable takes the same value for all lists-in-district regardless of their actual inter-party competitiveness.

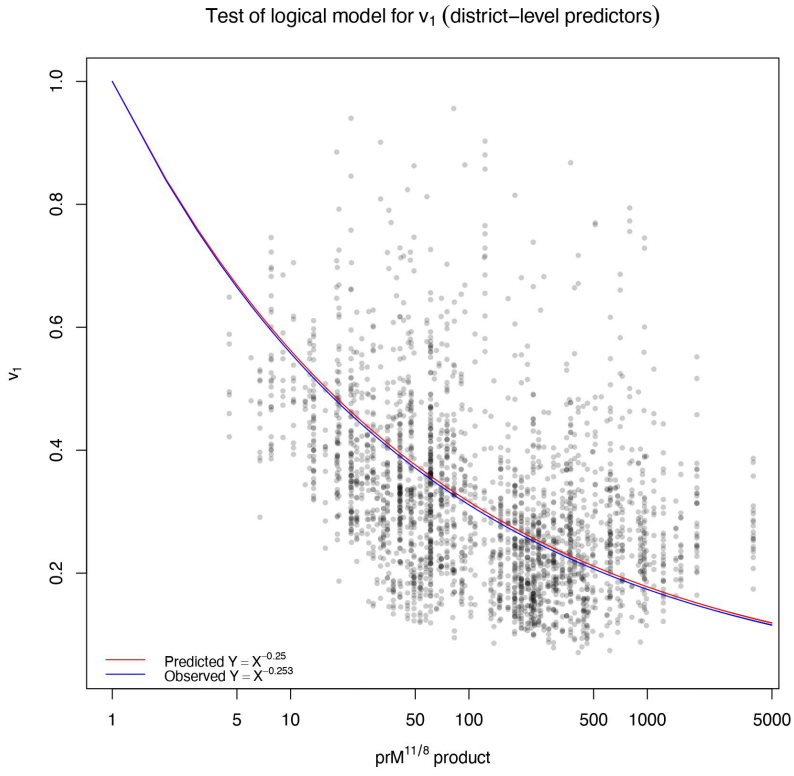
Table 3: Comparison of model fits for predictive models of the first-ranked candidate's fractional share of preference votes (v_1).

		Full Sample		Open List Only		Flexible List Only	
	$\hat{k}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$
List-Level Model	-0.250	-0.263	3.088	-0.272	5.753	-0.243	0.806
District-Level Model	-0.250	-0.253	0.992	-0.259	2.264	-0.243	1.018
SBW Model	-0.500	-0.435	11.574	-0.423	13.456	-0.468	2.270

Table 3 reports a comparison of the slope coefficients obtained from regressing v_1 on the products of list-level and district-level quantities, against the performance of the SBW model, where v_1 is regressed on c. Absolute values of the discrepancy between predicted and observed values of the coefficients normalised by the standard error are reported across the full, open-list and flexible-list samples. While both the refined models outperform the coarse model, the district-level model's predictions are remarkably accurate, with $\hat{k}$ falling within one or two standard errors from the prediction, depending on the sample specification. In the full sample, the SBW's slope falls over 11 standard errors away from the observed slope: a much larger bias

Figure 4: Test of list- and district-level predictions for the first-ranked candidate's share of preference votes: predicted and observed slope of v_1 regressed on scp and $prM^{\frac{11}{8}}$.





than the 3 standard errors of the list-level 'revised' model and the 1 standard error bias of the district-level 'revised model'. The SBW model for v_1 performs 'best' in the flexible-list sub-sample: this is very surprising as it was meant for and tested on open-list systems only.

6.1.2 Effective Number of Candidates (N_c)

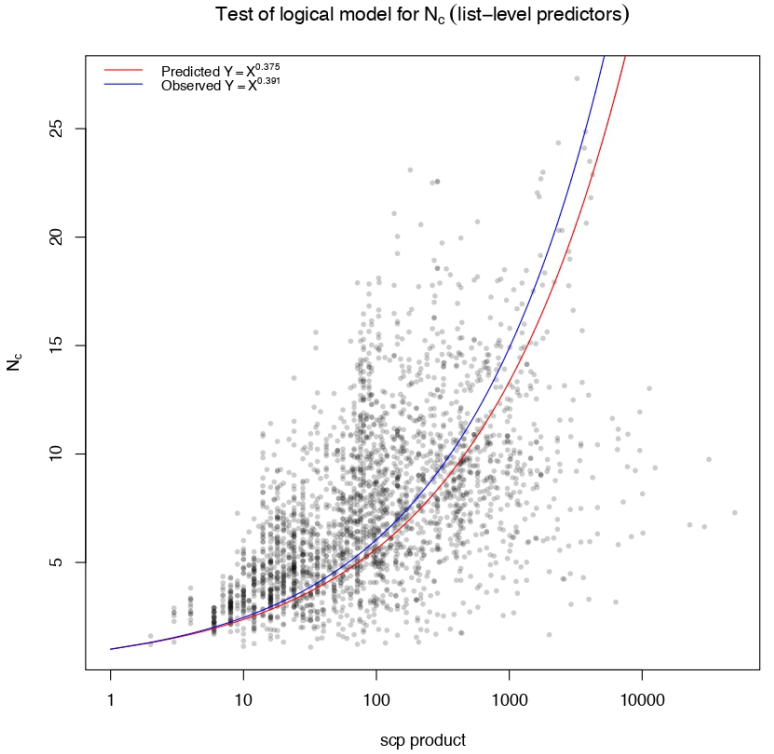
Regressing the observed distribution of the effective number of candidates N_c on *scp* returns a fixed-intercept exponential function of $Y = X^{0.391}$ with the coefficient having a 95% confidence interval $(0.381, 0.40)$. Again, the prediction $\hat{k} = 0.375$ comes close to the observed value of the slope but somewhat understates intra-party competitiveness, in this case by narrowly under-predicting N_c. Regressing N_c on $prM^{\frac{11}{8}}$ returns a fixed-intercept exponential function of $Y = X^{0.379}$ with the coefficient having a 95% confidence interval (0.367, 0.390), therefore including the predicted value of $k = 0.375$. Figure 5 plots the predicted and observed exponential functions for the two models, and table 4 reports the observed slope coefficients across different specifications of the sample. As in the case of v_1, the institutions-only model is noticeably less biased than the list-level model and its performance presents little variance due to sample specification. Again, the graphs show clear outliers occurring when the *scp* product and its institution-only equivalent $prM^{\frac{11}{8}}$ take the highest observed values. These occur in Slovakia's nationwide district.[24]

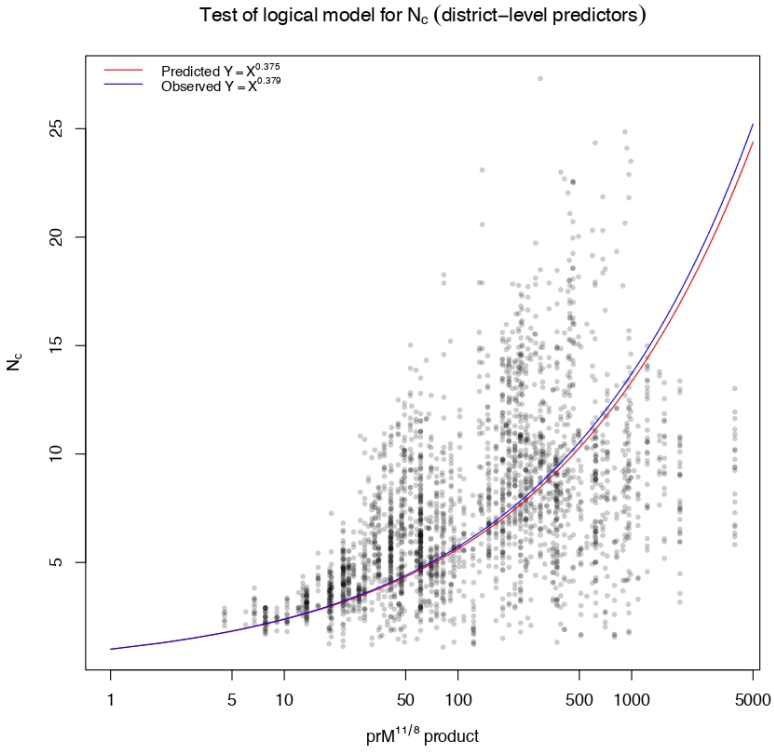
Table 4: Model fits for predictive models of the effective number of candidates.

		Full Sample		Open List Only		Flexible List Only	
	$\hat{k}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$
List-Level Model	0.375	0.391	3.136	0.402	5.414	0.369	0.582
District-Level Model	0.375	0.379	0.888	0.385	1.878	0.367	0.928

[24]For further discussion of these outliers, see section 7.3.

Figure 5: Test of list- and district-level predictions for the effective number of candidates: predicted and observed slope of N_c regressed on scp and $prM^{\frac{11}{8}}$.





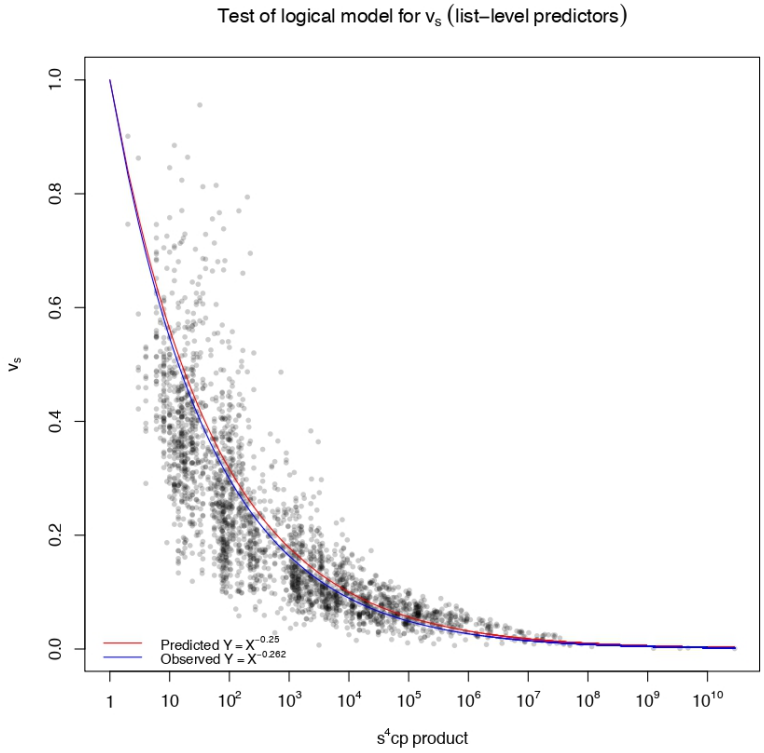
6.1.3 Last Eligible Candidate's Share of Preferential Votes

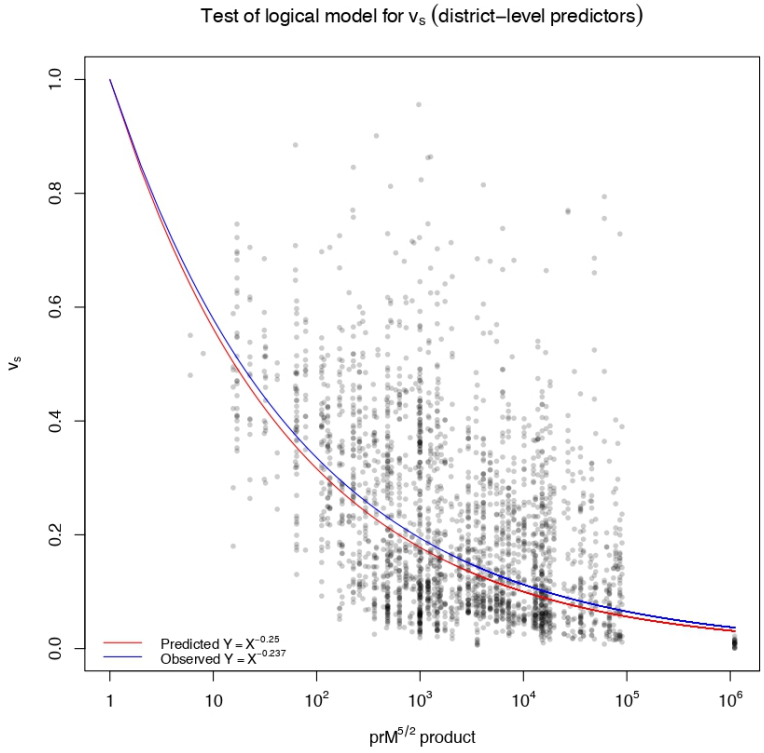
Regressing v_s on s^4cp returns a fixed-intercept exponential function of $Y = X^{-0.262}$ with the coefficient having a 95% confidence interval (-0.266, -0.258). Regressing v_s on $prM^{\frac{5}{2}}$ returns a fixed-intercept exponential function of $Y = X^{-0.237}$ with the coefficient having a 95% confidence interval (-0.244, -0.230). As shown in figure 6, in this case we find that the list-level model somewhat under-predicts v_s, while the district-level model somewhat over-predicts it. Table 5 reports coefficients across specifications of the sample for the list- and district-level models, alongside the normalised discrepancy from the expected value of $\hat{k}$. In this case, the SBW model – especially in its 'corrected' version where $\hat{k}$ depends on the empirical slope of v_1 as a function of c – comes close to the model performance of the two 'refined' models. Indeed, in the sub-sample of flexible-list observations, it outperforms them. However, there is a large amount of variation in performance of the SBW model fit across specifications of the sample: something that we do not observe to the same extent for the 'refined' list- and district-level models. In the full sample of observations, the list- and district-level models for v_s present more bias than in the models for v_1 and N_c described above, but nonetheless they still both outperform the SBW model.

Table 5: Comparison of model fits for predictive models of the last eligible candidate's preference share.

		Full Sample		Open List Only		Flexible List Only	
	$\hat{k}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	β	$\frac{\lvert\beta-\hat{k}\rvert}{se}$
List-Level Model	-0.250	-0.262	6.928	-0.259	4.206	-0.270	6.971
District-Level Model	-0.250	-0.237	4.636	-0.229	7.800	-0.256	1.130
SBW Model (uncorrected)	-0.750	-0.655	13.572	-0.622	24.492	-0.750	0.018
SBW Model (corrected)	-0.717	-0.655	8.897				
	-0.712			-0.622	12.798		
	-0.734					-0.750	1.008

Figure 6: Test of list- and district-level predictions for the last eligible candidate's share of preferences: predicted and observed slope of v_s regressed on s^4cp and $prM^{\frac{5}{2}}$.





6.2 Model Precision and Comparison with Inter-Party Models

6.2.1 First-ranked candidate (v_1) and largest party (σ_1) predictions compared

Figure 7 plots on the y-axis the values of the discrepancy index d – which correspond to the log-transformed ratio of observed and predicted values – and on the x-axis is the base product of each of the three predictive models under examination. Specifically, the top panel visualises the logged ratio of observed values of v_1 and the predicted quantity $(scp)^{-\frac{1}{4}}$ for all seat-winning lists in the dataset described in section 4. The middle panel represents the logged ratio of observed median values of v_1 in each district and the prediction $(prM^{\frac{11}{8}})^{-\frac{1}{4}}$, drawing on the same data. The bottom panel visualises the logged ratio of observed values of σ_1 (fractional share of the largest party in an assembly) and the SPM predicted values of this quantity, computed as $(MS)^{-\frac{1}{8}}$ on the Nemčok-Šedo dataset of elections.

It is evident from the plots that list-level predictions of v_1 tend to fall farther from zero than those of the other models. However, when it comes to predicting *median* intra-party competition, the panel plot for the district-level model is visually very similar to that of the SPM's predictions. Not only do the vast majority of the values of d fall within -0.3 and 0.3, indicating that a prediction is within a factor of 2 from the observed value, but the shapes of the distributions of d in the second and third panels are also quite similar, even though they represent distinct quantities computed on distinct datasets. Specifically, both predictions of first-ranked candidate shares and of largest party shares tend to 'miss the mark' most clearly when the models' base products take the highest values – i.e. for those sets of electoral institutions that in theory should be most conducive to competition – and in both cases the prediction overstates competitiveness relative to reality. This suggests that there are some upper constraints of political nature to the fragmentation of a system can handle when it is least constrained: under highly permissive rules 'on paper', party and preference votes will tend to concentrate in ways that cannot be accounted for simply by institutional factors (or at least not the parsimonious set of institutional factors that are sufficient to predict competition in other contexts).[25]

Table 6: Summary indicators of deviation from prediction: models for v_1, $\tilde{v}_1$ and σ_1 compared

	median of $\lvert d \rvert$		mean of $\lvert d \rvert$		share $d < log_{10}(2)$
	value $\lvert d \rvert$	% error	value $\lvert d \rvert$	% error	and $d > log_{10}(0.5)$
list-level model (v_1)	0.119	31.6%	0.149	40.9%	87.9%
district-level model ($\tilde{v}_1$)	0.083	21.1%	0.103	26.8%	97.1%
seat-product model (σ_1)	0.091	23.4%	0.105	27.4%	96.2%

[25] Interestingly, Slovakia is a case of both phenomena occurring at the same time: a highly permissive nationwide PR system that should have relatively small largest parties and low first-ranked candidate shares, but presents moderate levels of vote concentration on both dimensions, at least in some elections.

Figure 7: Comparison of deviation from prediction for v_1 (list-level model), median $\tilde{v}_1$ (district-level model) and σ_1 (seat-product model). Dashed lines represent values of d corresponding to values where the observed value is either twice or half the prediction.

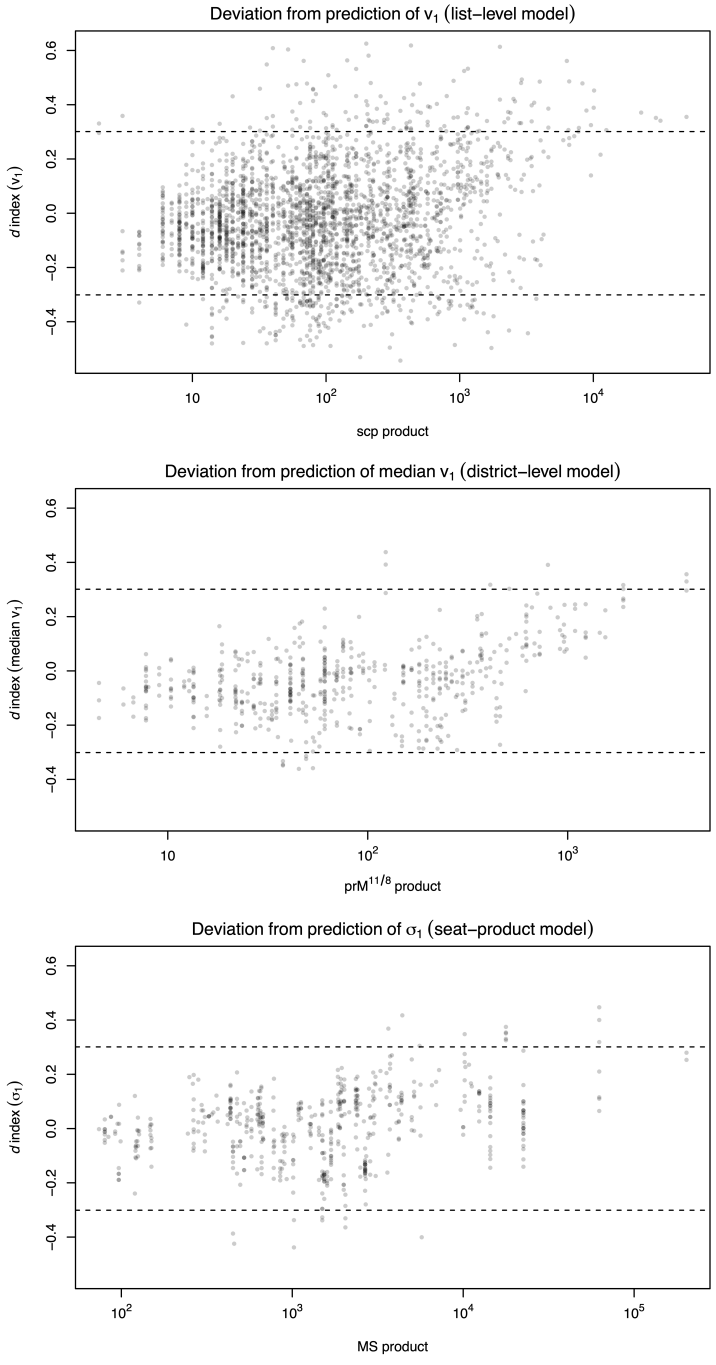


Table 6 confirms the conclusions drawn from the visual presentation of the data. The average absolute values of d are highest for the list-level model of v_1, and similarly low for the district-level model of $\tilde{v}_1$

and the seat-product model for σ_1. Alongside the measures of central tendency, the table also reports the associated error factors: the median (mean) prediction of the list-level model is off by 32% (41%); the median (mean) prediction of the district-level model is off by 21% (27%); the median (mean) prediction of the seat product model is off by 23% (27%). The logical models under consideration can predict the largest party's fractional share of seats and the median fractional share of first candidate's preference votes from institutional quantities for almost all observations (96–97%) in the samples within a factor of 2. However, list-specific predictions are clearly less precise, with 12% of the model expectations being more than double or less than half the actual values.

6.2.2 Effective Number of Candidates (N_c) and Parties (N_S) Predictions Compared

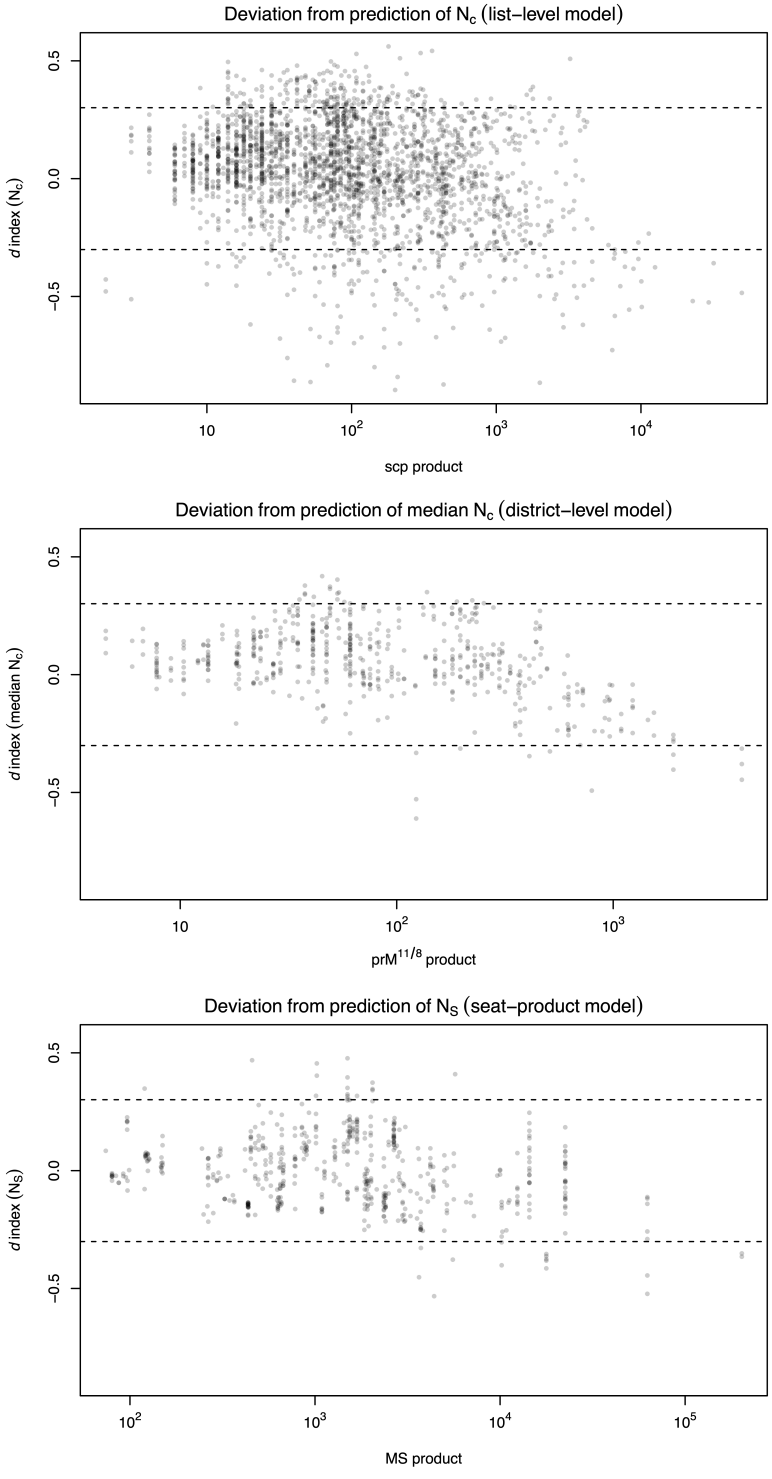
This subsection reproduces the same analysis of 'deviation from prediction' but for a different set of dependent variables: the effective number of candidates of a list (N_c), the median effective number of candidates for a list in a district ($\tilde{N}_c$), and the effective number of parties in an assembly (N_S).

Figure 8 plots the values of d against the models' base products. In keeping with the observations of the previous subsection, the visualisations show that the predictions of the list-level model are much more widely scattered, and hence less precise, than those of the district-level model and the SPM. And once again we find that the distribution of the deviation-from-prediction values for median effective number of candidates in a district and the effective number of parties in an assembly are strikingly similar. In both cases, the vast majority (94.5%) of values lie within the bounds of 'tolerable' discrepancy comprised between -0.3 ($\log_{10} 0.5$) and 0.3 ($\log_{10} 2$); and in both cases, the most notable deviations are instances of *over-prediction* of competitiveness observed where the base products take the highest values observed. Table 7 reports measures of central tendency for the distribution of d across models, which confirm that the district-level model and the seat-product model are about as precise in predicting the effective number of parties/candidates in their respective samples.

Table 7: Summary indicators of deviation from prediction: models for N_c, $\tilde{N}_c$ and N_S compared

	median of $\lvert d \rvert$		mean of $\lvert d \rvert$		share $d < log_{10}(2)$
	value $\lvert d \rvert$	% error	value $\lvert d \rvert$	% error	and $d > log_{10}(0.5)$
list-level model (N_c)	0.142	38.6%	0.169	47.6%	84.3%
district-level model ($\tilde{N}_c$)	0.106	27.8%	0.126	33.8%	94.5%
seat-product model (N_S)	0.113	29.6%	0.124	33%	94.5%

Figure 8: Comparison of deviation from prediction for N_c (list-level model), median $\tilde{N}_c$ (district-level model) and N_S (seat-product model) . Dashed lines represent values of d corresponding to values where the observed value is either twice or half the prediction.



Summing up, the answer to the guiding question of whether we can predict intra-party quantities from institutional variables in the same way as we can predict inter-party is thus arguably both a 'no' and a 'yes'. On the one hand, the list-level model presented fails to estimate first-ranked candidate's preference share *for individual lists* with the same precision of the SPM's prediction for the fractional share of seats of the largest party in an assembly. On the other hand, at district level, we can derive a prediction of the median value of v_1 and N_c for seat-winning parties from purely institutional variables that is just as precise as the SPM's prediction for σ_1 and N_S. Of course, one may retort that even in this case the terms of the comparisons presented are not entirely 'fair': although the number of districts used to test district-level predictions (549) and the number of elections used to test the SPM predictions (560) are almost identical, the SPM draws on data from 40 countries, while the intra-party model only on nine. At the same time, to compare 'like with like', we ideally may want to compare *district-level* inter-party predictions with *district-level* intra-party prediction. These are both valid caveats to the comparison presented, which will be addressed in future research through further data collection of district-level and list-level results.

7 Discussion and Conclusion

Intra-party competition has often been thought of as unpredictable and idiosyncratic in comparison to inter-party competition, to the extent that most works on PLPR begin by commenting on how poorly understood and understudied this aspect of electoral institutions are.[26] This paper aims to join a small but growing group of theoretical (Shugart, Bergman and Watt, 2013; Buisseret and Prato, 2020; Buisseret et al., 2022) and empirical (Selb and Lutz, 2015; Renwick and Pilet, 2016; Blom-Hansen et al., 2016; Däubler and Hix, 2018; Passarelli, 2020; Cheibub and Sin, 2020; Dodeigne and Pilet, 2021) works aiming to identify order in the apparent chaos of intra-party competition in candidate-centred electoral systems. To conclude the paper, I present a possible application of the framework outlined to political practice, discuss the implications of my analysis for further study of PLPR systems, and note some shortcomings of the models.

7.1 An Application to Institutional Choice

What practical use is it to know how preference votes 'should', in expectation, be distributed within a PLPR list? As an illustration of the potential relevance of the models presented to institutional design, let us consider how an institutional engineer may decide to fix the preference threshold in a FLPR system, so as to achieve a desired balance-of-power between voters and parties. Consider a FLPR system where a candidate must receive at least a fractional share t of preference votes to be elected on the basis of preferential votes; if the number of seats allocated to a list exceeds the number of candidates that meet this threshold, the remaining seats are allocated according to the ballot position determined by the party. Thus, t effectively indicates how much control parties and voters have on the intra-party allocation of seats: the higher the fractional value of t, the more candidates will be elected from their ballot position; the lower the fractional value of t, the more candidates will be elected from preference votes.

From the theoretical considerations above, we can identify three key values of t expressed as a fractional share of preference votes: (1) the value of t that is equally likely to produce one or no candidates elected on preferences in a district (t_1), (2) the value of t that is equally likely to produce M or $M-1$ candidates elected on preferences in a district (t_M); and (3) the value of t whereby in expectation 50% of MPs are elected on preference vote ranking and 50% of MPs are elected due to ballot position ($t_{\frac{M}{2}}$). t_1 is realised when the threshold equals the expected fractional share of preference votes of the relative top-performing seat-eligible

[26] For instance, for Buisseret and Prato (2018, p.1) "[i]n spite of their widespread use, open- and flexible-list proportional representation [...] systems have received little attention from empirical scholars and almost none from theoretical scholars. In large part, this is due to the fact that these systems vary tremendously in their operation across countries, bedeviling attempts at classification and limiting scholarly efforts to the analysis of specific cases."

candidate. As, in expectation, such candidate is the first-ranked candidate (who gets a preference share of v_1) of the smallest party (which expects to win one seat), it follows that

$$t_1 = (p \times rM \times 1)^{-\frac{1}{4}} = (prM)^{-\frac{1}{4}} \tag{16}$$

As for t_M, the institutional engineer would set the threshold to equal the expected preference share of the seat-eligible candidate gaining the smallest relative preference share in the district. As, in expectation, such candidate is the last-ranked candidate (who gets a preference share of v_s) of the largest party (gaining $M^{\frac{3}{4}}$ seats), it follows that

$$t_M = \left(p \times rM \times (M^{\frac{3}{4}})^4\right)^{-\frac{1}{4}} = (prM^4)^{-\frac{1}{4}} \tag{17}$$

I conjecture that $t_{\frac{M}{2}}$ may be approximated as the geometric mean of t_1 and t_M, so that

$$t_{\frac{M}{2}} = \left((prM)^{-\frac{1}{4}} \times (prM^4)^{-\frac{1}{4}}\right)^{\frac{1}{2}} = (prM^{\frac{5}{2}})^{-\frac{1}{4}} \tag{18}$$

Which version of t should the institutional engineer choose? t_1 is of no use: a candidate popular enough to get that high a share of preference votes is likely to be already in a seat-eligible position. Effectively, t_1 produces what Däubler and Hix (2018) term 'weakly flexible' lists, where it is unlikely that any candidate beyond the top-ranked gather enough preference votes to be elected on the basis of their personal support. t_M, conversely, would be an apt choice if the institutional engineer wanted to have a permissive 'quasi-open' list system, which only prevents candidates with very little personal appeal from lucking into a parliamentary seat on a sparse number of preference votes. Under t_M, we are guaranteed to observe what Däubler and Hix (2018) term 'strongly flexible' lists. In fact, small parties are effectively competing under open-list rules, because the threshold is too low to ever make a difference, while large parties retain occasionally some degree of control over the intra-party allocation but only for the last few seats. The choice of $t_{\frac{M}{2}}$ would be a compromise between the two. On the one hand, parties retain a substantial deal of control over election outcomes: though half of the candidates in a district are elected on preferences, most of them would presumably be in seat-eligible ballot positions anyway. However, such a system is still flexible enough to reward – occasionally – strong performances from down-ballot candidates, making preferential voting meaningful. Table 8 reports the share of candidates who would have been elected on preference votes under the different district-level specifications of the threshold derived above, *if these elections in the sample had be*

conducted under such FLPR rules. Albeit rudimentary, this analysis is encouraging: in no case the thresholds are so extreme as to become trivial (some, but not all, candidates are elected on preferences in all cases) and $t_{\frac{M}{2}}$ is on average quite close to the speculative prior that it might produce a 50-50 split between candidates elected on preference ranking and candidates elected on ballot position. However, cross-country variation is substantial.

Table 8: Simulation of PLPR outcomes under different values of hypothetical preference thresholds.

	Share of candidates elected on preferences under		
Country/Sample	t_1	$t_{\frac{M}{2}}$	t_M
Belgium	0.18	0.65	0.98
Cyprus	0.02	0.57	0.99
Czechia	0.02	0.39	0.95
Estonia	0.10	0.58	0.84
Finland	0.03	0.36	0.97
Italy	0.08	0.49	0.99
Peru	0.05	0.55	0.92
Poland	0.06	0.39	0.75
Slovakia	0.04	0.27	0.65
Country Average	0.07	0.47	0.89
Full Sample	0.07	0.46	0.92

Of course, in most existing FLPR systems, thresholds are set as a function of *list* rather than *preference* votes, so that the values of t derived above are directly applicable to real-world electoral systems only in cases where $p = 1$ and preferential vote is mandatory.[27] If $p > 1$, preference votes exceed list votes, and thus t would have to be adjusted upwards by a factor of $p^{\frac{1}{2}}$, the expected number of preferences-per-voter. If preferential voting is optional, then list votes exceed preference votes, and t would have to be adjusted downwards by the expected fractional share of voters who cast a preference votes. This is a less readily predictable factor, not just because it may largely depend on political rather than institutional factors (party-voter linkages, voter engagement, democratic experience, party system and personnel stability etc.), but also because the extent to which voters will make use of a preference vote is likely endogenous to list flexibility itself (Däubler, 2020; Renwick and Pilet, 2016, pp. 217–248).

7.2 Implications for the Study of PLPR

The most novel aspect of the model presented is the introduction of proxies for expectations of inter-party performance as an input variable, in the form of s (the number of seats effectively gained by a list) for the

[27] An even less immediately tractable case is when t is expressed as a function of the electoral quota. However, this case could be tackled drawing on recent advances in modelling party *vote* share (Shugart and Taagepera, 2017, pp. 125-138), and thereby expressing expected preference shares as fractions of expected quotas attributed to each party.

list-level model, and $M^{\frac{3}{8}}$ (the expected number of seats for the median list) for the district-level model. The role of actors' expectations in shaping electoral system outcomes has long been studied as a key mechanism shaping inter-party competition. This is effectively what Duverger (1959) termed the 'psychological effect' of district magnitude, and scholars working in his tradition expanded to encompass parties', as well as voters', strategic behaviour (Cox, 1997). Yet, the recognition of a similar mechanism being at play at the intra-party level departs from the standard theoretical assumption, for instance made by Shugart, Bergman and Watt (2013), that district magnitude affects list competitiveness only insofar as it poses constaints on parties' nominating behaviour. The revised model, conversely, takes into account two avenues through which M is connected to intra-party competition: by constraining candidate nominations *and* by providing candidates (and voters) with priors on the viability of their candidacy. Future empirical research might aim to document the extent of actors' anticipations about list performance and its incidence over aspects of candidate behaviour with respects to intra-party competition.

The second innovative aspect of the paper lies in the consideration of open- and flexible-list PR systems, as well as single- and multiple-preference forms of PLPR, within the same theoretical framework. As discussed, the prospect of broadening the scope of application of logical models beyond simple single-vote OLPR by taking into account multiple preference votes is something of a 'wager'. That is, we posited that – once we account for the different number of preferential votes used across systems – the intra-party allocation rules distinguishing OLPR from FLPR would not make a substantial difference. As the analysis presented in section 6.1 shows that the revised model works reasonably well across the FLPR and OLPR sub-samples (unlike the SBW model), we can tentatively conclude that the wager paid off. There is an important pragmatic rationale for extending the comparative analysis of intra-party competition to complex types of PLPR: these constitute the overwhelming majority of empirical cases. In Passarelli's (2020) review of PLPR systems, out of 29 countries for which this information is available, only seven employ single-preference OLPR (Finland, Kosovo, Poland, Brazil, Colombia, Indonesia, Lebanon), two of which have only recently switched from more complex systems (Indonesia from FLPR, and Kosovo from multiple-preference OLPR) and one of which allows parties to field closed lists if they wish (Colombia). In contrast, thirteen countries have $p > 1$ and fourteen use some form of FLPR for electing their national parliaments. In line with empirical research showing that prospects of advancement to electable positions produce a substantial degree of personal vote-seeking effort even under relatively restrictive FLPR rules (André et al., 2017), the findings of this paper should therefore encourage researchers to treat FLPR as a cognate of OLPR rather than 'closed lists in

disguise', at least as far as the aggregate outcomes of intra-party competition are concerned.

7.3 Limitations

All models are wrong, but some are useful. The reader hopefully will agree that the ones presented in this paper may belong to the latter category, insofar as they illustrate non-obvious relationships between electoral system quantities, have respectable predictive power, and – as discussed in this latter section – may have practical applications to institutional choice. Nonetheless, there are some areas of concern and related room for improvement; in particular, there are three outstanding issues with the present attempt to model intra-party quantities, which may be addressed in future research.

First, as noted in various parts of the paper, the models perform particularly poorly in the case of Slovakia, which employs a nationwide district of magnitude 150 (district magnitude M is equal to assembly size S). My sense is that this is due to a 'big fish' effect: politicians who are popular nationwide, like party leaders and prominent frontbenchers, can concentrate preference votes to an exceptional degree on their candidacy. This is of course happens beyond cases where $M = S$: the model over-predicts competitiveness just as poorly for the lists including incumbent Prime Ministers like Poland's Donald Tusk (75% of preference votes) or Belgium's Charles Michel (51%), as well as charismatic party leaders like Estonia's Martin Helme (94%) and Czechia's Tomio Okamura (48%). (And, given what a logical model is for, the model would have no business predicting these cases.) However, where $M \ll S$, these 'big fish' lists are only a fraction of the observations; whereas if $M = S$ then *all* lists will have at least a party leader on the ballot. The resulting bias of the prediction is thus substantially overstating competitiveness. Future iterations of a model for PLPR may thus attempt to account for the 'embeddedness' of districts in national politics: i.e. as the $\frac{M}{S}$ ratio tends to one, our prediction for v_1 should be adjusted upwards, and that for N_c should be adjusted downwards. At present, I am unable to justify a quantitative formalisation of such an adjustment.

Secondly, the theoretical derivation of v_s rests on the heroic assumption that mean and median of preference shares of seat eligible candidates are approximately equal. The empirical performance of the deriving models, both in terms of bias and precision, is respectable; but perhaps it is not respectable enough to warrant such a mathematical heresy, especially as the models for v_1 and N_c tend to do better than those for v_s, and to do so more consistently across specifications of the model and the sample. In sum, v_s remains the weakest link of the interlocked set of equations describing expected quantities of intra-party competition.

A final limitation of the argument presented is a clear definition of its scope conditions. As discussed,

there is value in extending the analysis to as broad a set of PLPR institutional set-ups as feasible, given the diversity of real-world variants of this electoral system family. In section 4, the reference to 'systems that are sufficiently proportional and sufficiently preferential' makes for an informal and intuitive way of defining the scope conditions of the theory, but I recognise that it is not sufficiently precise. A PLPR system with a very high representation threshold or a very small divisor formula might constrain inter-party competition enough to alter significantly actors' expectations over the number of seats at stake for an individual list. On the intra-party dimension, a FLPR system with an unattainably high preference threshold may make intra-party competition meaningless, and hence more random than the predictable patterns a logical model might aim for. Some of these factors may be accounted for by introducing further terms to the formulas, and I have made the case for sacrificing a degree of parsimony in favour of wider applicability to the complexity of real-world cases. But, for some extreme cases, we might simply have to conclude that we are dealing with something other than a *preferential* list *proportional* representation system.

7.4 Conclusion

In spite of these shortcomings, the empirical tests presented in the second part of the paper are overall rather encouraging for our quest towards a model of intra-party competition to match the seat-product model. Summing up, I have shown that we can predict from electoral and institutional quantities (1) what percentage of preference votes the top candidate will get, (2) how many candidates will effectively emerge within a list, and (3) what is the minimum share of preferences a candidate should get to be eligible for a seat. In the sample considered, these predictions are about as precise as those available for similar indicators of inter-party competition, they are less biased than those of existing intra-party models, and they perform consistently across sub-types of PLPR systems. Hopefully, as well as providing a stepping stone towards a more systematic and comprehensive research agenda on intra-party competition in PLPR, these results may also serve as testimony to the power of quantitative predictive logical models as a versatile and parsimonious theory-building tool.

References

André, Audrey, Sam Depauw and Kris Deschouwer. 2014. "Legislators' local roots: Disentangling the effect of district magnitude." *Party Politics* 20(6):904–917.

André, Audrey, Sam Depauw, Matthew S Shugart and Roman Chytilek. 2017. "Party nomination strategies in flexible-list systems: Do preference votes matter?" *Party Politics* 23(5):589–600.

Arter, David. 2013. "The 'Hows', not the 'Whys' or the 'Wherefores': The role of intra-party competition in the 2011 breakthrough of the True Finns." *Scandinavian Political Studies* 36(2):99–120.

Blom-Hansen, Jens, Jørgen Elklit, Søren Serritzlew and Louise Riis Villadsen. 2016. "Ballot position and election results: Evidence from a natural experiment." *Electoral Studies* 44(1):172–183.

Buisseret, Peter and Carlo Prato. 2018. "Legislative representation in flexible-list electoral systems." *Unpublished mansucript* .

Buisseret, Peter and Carlo Prato. 2020. "Voting behavior under proportional representation." *Journal of Theoretical Politics* 32(1):96–111.

Buisseret, Peter, Olle Folke, Carlo Prato and Johanna Rickne. 2022. "Party nomination strategies in list proportional representation systems." *American Journal of Political Science* 66(3):714–729.

Cheibub, José Antonio and Gisela Sin. 2020. "Preference vote and intra-party competition in open list PR systems." *Journal of Theoretical Politics* 32(1):70–95.

Cox, Gary W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge University Press.

Crisp, Brian F, Kathryn M Jensen and Yael Shomer. 2007. "Magnitude and vote seeking." *Electoral Studies* 26(4):727–734.

Däubler, Thomas. 2020. "Do more flexible lists increase the take-up of preference voting?" *Electoral Studies* 68:102232.

Däubler, Thomas and Simon Hix. 2018. "Ballot structure, list flexibility and policy representation." *Journal of European Public Policy* 25(12):1798–1816.

Devroe, Robin and Bram Wauters. 2020. "Does high on the ballot means highly competent? Explaining the ballot position effect in list-PR systems." *Acta Politica* 55(3):454–471.

Dodeigne, Jérémy and Jean-Benoit Pilet. 2021. "Centralized or decentralized personalization? Measuring intra-party competition in open and flexible list PR systems." *Party Politics* 27(2):234–245.

Duverger, Maurice. 1959. *Political parties: Their Organization and Activity in the Modern State*. Metheun & Co. Ltd.

Karvonen, Lauri. 2004. "Preferential voting: Incidence and effects." *International Political Science Review* 25(2):203–226.

Kimura, Daniel K. 1992. "Symmetry and scale dependence in functional relationship regression." *Systematic Biology* 41(2):233–241.

Laakso, Markku and Rein Taagepera. 1979. ""Effective" number of parties: a measure with application to West Europe." *Comparative Political Studies* 12(1):3–27.

Li, Yuhui and Matthew S Shugart. 2016. "The seat product model of the effective number of parties: A case for applied political science." *Electoral Studies* 41(1):23–34.

Lutz, Georg. 2010. "First come, first served: The effect of ballot position on electoral success in open ballot PR elections." *Representation* 46(2):167–181.

Nemčok, Miroslav and Jakub Šedo. 2018. "Theoretical expectations and actual outcomes of electoral systems: how to measure the size of the deviation?" *World Political Science* 14(2):189–212.

Passarelli, Gianluca. 2020. *Preferential Voting Systems*. Springer.

Renwick, Alan and Jean-Benoit Pilet. 2016. *Faces on the Ballot: The Personalization of Electoral Systems in Europe*. Oxford University Press.

Selb, Peter and Georg Lutz. 2015. "Lone fighters: Intraparty competition, interparty competition, and candidates' vote seeking efforts in open-ballot PR elections." *Electoral Studies* 39(1):329–337.

Shugart, Matthew S. 2005. Comparative electoral systems research: The maturation of a field and new challenges ahead. In *The Politics of Electoral Systems*, ed. Michael Gallagher and Paul Mitchell. Oxford University Press pp. 25–56.

Shugart, Matthew S, Matthew E Bergman and Kevin A Watt. 2013. "Patterns of intraparty competition in open-list & SNTV systems." *Electoral Studies* 32(2):321–333.

Shugart, Matthew S and Rein Taagepera. 1989. *Seats and Votes: The Effects & Determinants of Electoral Systems*. Yale University Press.

Shugart, Matthew S and Rein Taagepera. 2017. *Votes from Seats: Logical Models of Electoral Systems*. Cambridge University Press.

Sikk, Allan and Rein Taagepera. 2014. "How population size affects party systems and cabinet duration." *Party Politics* 20(4):591–603.

Taagepera, Rein. 2007. *Predicting party sizes: The logic of simple electoral systems*. OUP Oxford.

Taagepera, Rein. 2008. *Making social sciences more scientific: The need for predictive models*. Oxford University Press.

Taagepera, Rein and Allan Sikk. 2010. "Parsimonious model for predicting mean cabinet duration on the basis of electoral system." *Party Politics* 16(2):261–281.

Taagepera, Rein and Bernard Grofman. 1985. "Rethinking Duverger's law: predicting the effective number of parties in plurality and PR systems–parties minus issues equals one." *European Journal of Political Research* 13(4):341–352.

Taagepera, Rein and Matthew S Shugart. 1993. "Predicting the number of parties: A quantitative model of Duverger's mechanical effect." *American Political Science Review* 87(2):455–464.

Taagepera, Rein, Peter Selb and Bernard Grofman. 2014. "How turnout depends on the number of parties: A logical model." *Journal of Elections, Public Opinion & Parties* 24(4):393–413.

Van Erkel, Patrick FA and Peter Thijssen. 2016. "The first one wins: Distilling the primacy effect." *Electoral Studies* 44(1):245–254.

Appendix

A Within-District relationship between s, v_1 and N_c

In section 2.2.2, I motivated the assumption that v_1 is endogenous to expectations of seat gains by illustrating how the distribution of preference shares in a Polish district is 'flatter' for more successful parties and 'steeper' for smaller ones. The empirical plausibility of this assumption can be more rigorously assessed by considering the wider universe of districts considered in the analysis (see footnote 8). Tables 1 and 2 report the results of OLS regressions *with district-election fixed effects*, where list-level values of v_1 and N_c are regressed on s (the raw number of seats gained) and on $\frac{s}{M}$, the share of seats gained. Linear and log-log specifications are presented. Consistently with the assumption, larger list have lower values of v_1 and higher values of N_c. The coefficients are virtually identical when controlling for the number of candidates fielded c (tables 3 and 4), confirming that the relationship is largely independent from parties' nominating behaviour.

Table 1: District-election fixed effects model: relationship between the number of seats and intra-party quantities.

	Dependent variable:			
	v_1	N_c	$\log(v_1)$	$\log(N_c)$
	(1)	(2)	(3)	(4)
s	-0.007^{***}	0.295^{***}		
	(0.001)	(0.017)		
$\log(s)$			-0.161^{***}	0.151^{***}
			(0.009)	(0.010)
District-Election F.E.	✓	✓	✓	✓
Observations	2,617	2,617	2,617	2,617
R^2	0.051	0.127	0.128	0.103
Adjusted R^2	−0.201	−0.105	−0.103	−0.136
F Statistic (df = 1; 2067)	110.923^{***}	300.391^{***}	304.531^{***}	236.575^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: District-election fixed effects model: relationship between the share of seats and intra-party quantities.

	Dependent variable:			
	v_1	N_c	$\log(v_1)$	$\log(N_c)$
	(1)	(2)	(3)	(4)
s/M	-0.200^{***}	6.004^{***}		
	(0.015)	(0.376)		
$\log(s/M)$			-0.161^{***}	0.151^{***}
			(0.009)	(0.010)
District-Election F.E.	✓	✓	✓	✓
Observations	2,617	2,617	2,617	2,617
R^2	0.077	0.110	0.128	0.103
Adjusted R^2	-0.168	-0.126	-0.103	-0.136
F Statistic (df = 1; 2067)	171.882^{***}	255.481^{***}	304.531^{***}	236.575^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: District-election fixed effects model: relationship between the share of seats and intra-party quantities, controlling for list length.

	Dependent variable:			
	v_1	N_c	$\log(v_1)$	$\log(N_c)$
	(1)	(2)	(3)	(4)
s	-0.007^{***}	0.284^{***}		
	(0.001)	(0.017)		
c	-0.013^{***}	0.196^{***}		
	(0.001)	(0.031)		
$\log(s)$			-0.153^{***}	0.139^{***}
			(0.009)	(0.010)
$\log(c)$			-0.558^{***}	0.816^{***}
			(0.069)	(0.072)
District-Election F.E.	✓	✓	✓	✓
Observations	2,617	2,617	2,617	2,617
R^2	0.096	0.144	0.155	0.155
Adjusted R^2	−0.145	−0.084	−0.070	−0.070
F Statistic (df = 2; 2066)	109.844^{***}	173.252^{***}	189.714^{***}	189.377^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: District-election fixed effects model: relationship between the share of seats and intra-party quantities, controlling for list length.

	Dependent variable:			
	v_1	N_c	$\log(v_1)$	$\log(N_c)$
	(1)	(2)	(3)	(4)
s/M	-0.174^{***}	5.614^{***}		
	(0.015)	(0.377)		
c/M	-0.317^{***}	4.757^{***}		
	(0.030)	(0.744)		
$\log(s/M)$			-0.153^{***}	0.139^{***}
			(0.009)	(0.010)
$\log(c/M)$			-0.558^{***}	0.816^{***}
			(0.069)	(0.072)
District-Election F.E.	✓	✓	✓	✓
Observations	2,617	2,617	2,617	2,617
R^2	0.125	0.127	0.155	0.155
Adjusted R^2	-0.108	-0.105	-0.070	-0.070
F Statistic (df = 2; 2066)	147.542^{***}	150.622^{***}	189.714^{***}	189.377^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

B Bias analysis under different model specifications

Throughout the paper, I have noted a series of possible concerns with the validity of the model tests and assumptions. In this section, the analysis in section 6.1 is replicated with some tweaks of the theoretical assumptions, the modelling choices and the definition of the sample. The subsections below detail the possible alternatives to the choices behind the main results, and present the values of the slope $\hat{k}$ and the bias $\frac{|\beta-\hat{k}|}{se}$ when the alternative choices are implemented. To be sure, most of these are not, strictly speaking, 'robustness checks': they are tests of different models or of different dimensions of model performance. Future developments in formal analysis of PLPR may make the case for their usage over the one presented.

B.1 Alternative Specification of c^*

In section 2.2.2 footnote 10, it was noted that the list-level (and the related district-level) formulas cannot predict the values of v_1 (and the related value of N_c) for non-seat-winning parties. This is because of the assumption that the number of pertinent vote-earning candidates is endogenous to the expectations of the number of seats a list will win, and the further assumption that parties are 'correct' on average about their performance, so that $E(s) = s$, and $c^* = (sc)^{\frac{1}{2}}$. But what of lists that gain no seats? I suggested a possible avenue to generalise the model to lists where $s = 0$ by setting $c^* = [(s+1)c]^{\frac{1}{2}}$. This model effectively formalises a potentially realistic assumption that, even in the least-competitive scenario, there will be at least one non-elected candidate who had a shot at election. Furthermore, it is appealing because c^* retains the 'plus one' element of its inter-party analogue, the number of pertinent vote-earning parties, which is – both at district level and nationwide – modelled as $N_s + 1$ (Shugart and Taagepera, 2017, 128; see also the close cognate notion of 'viable candidates' modelled as $M + 1$ in Cox, 1997).

The deriving list-level equations from this tweak to the assumptions are $v_1 = [(s+1)cp]^{-\frac{1}{4}}$ and $N_c = [(s+1)cp]^{-\frac{3}{8}}$. The district-level equations are $v_1 = [prM(M^{\frac{3}{8}}+1)]^{-\frac{1}{4}}$ and $N_c = [prM(M^{\frac{3}{8}}+1)]^{\frac{3}{8}}$. Note that v_s remains unchanged as (1) algebraically, the s in the formula is derived from the average of seat-winning candidate shares, not from v_1, and (2) conceptually, it is illogical to derive v_s when $s = 0$, i.e. there is no 'last-eligible' candidate in a list that gains zero seat. Relatedly, a potential drawback of this model specification is that the predictions for v_1 and v_s when $s = 1$ differ in expectation. The table below reports the observed slopes β and the bias relative to the expected values $\hat{k}$ for the modified list- and district-level models and the two dependent variables of interest, tested over the same sample as the main analysis. Of course, ideally we would want to test this model on *all* lists, but this is not possible with the current data, which were only collected for seat-winning lists.

Table 1: Bias diagnostics under different specifications of c^*.

	List-level models		*District-level models*	
	v_1	N_c	v_1	N_c
base X	$(s+1)cp$	$(s+1)cp$	$prM(M^{\frac{3}{8}}+1)$	$prM(M^{\frac{3}{8}}+1)$
expected slope ($\hat{k}$)	-0.250	0.375	-0.250	0.375
observed slope (β)	-0.245	0.365	-0.240	0.358
$\frac{\lvert\beta-\hat{k}\rvert}{se}$	1.592	2.508	3.274	4.237

B.2 Embeddedness-adjusted district-level model

In section 3.2 footnote 16, I noted that Shugart and Taagepera (2017) revised their expected seat share for the first party in a district σ_1' to account for the 'embeddedness' of a district in the wider political system. They do so by introducing a k parameter which formalises the intuition that district-level competitiveness increases when a district is part of a larger political system. Consider for instance Barbados and the UK: both have a single-member seat plurality system, but the former has an assembly size S of 30 and the latter has an assembly size of 650. The latter will therefore produce parties that have only realistic chances in some districts, but nonetheless run throughout the country. This will result in higher inter-party competition in districts embedded in larger systems than districts embedded in smaller ones: for instance, the effective number of electoral parties in a British district is close to 3, while in Barbados is almost exactly 2. In multi-member seats, this 'embeddedness' effect is consequential because it means that as the ratio between magnitude and assembly size decreases, the number of small parties expected to win seats increases, and therefore – for instance – the expected share of seats for the first party is smaller.

Under this revised derivation, the expected value of σ_1' is $M^{-\frac{k}{2}}$. The value of k is district specific and set to equal $k = 0.5 + \frac{0.2076\ log_{10}(\frac{S}{M})}{M^{\frac{1}{4}}}$. In practice, k tends to 1 as M becomes infinitesimally small relative to the assembly size, and is exactly 0.5 when $M = S$. Details of its derivation are in Shugart and Taagepera (2017, pp. 174-177). For our purposes, what matters is that from the new formula for σ_1' it follows that the expected number of seats for the first party is $M \times s_1 = M^{1-\frac{k}{2}}$, and the expected number of seats for any party is $M^{\frac{2-k}{4}}$. As argued in section 3.2, this is the value of s expressed as a function of institutional variable, so that the resulting embeddedness-adjust district-level predictions are:

$$v_1 = \left(prM^{\frac{6-k}{4}}\right)^{-\frac{1}{4}} \qquad N_c = \left(prM^{\frac{6-k}{4}}\right)^{\frac{3}{8}} \qquad v_s = \left(prM^{3-k}\right)^{-\frac{1}{4}}$$

However, as discussed in section 7, I think that there is a more significant 'embeddedness' effect of relevance for intra-party competition – i.e. the fact that lists including 'big fish' politicians become more common as $\frac{M}{S}$ decreases – that is not captured by k. Thus, for the sake of simplicity, I chose to present a model that does not take into account the (presumably much smaller) bias engendered by the fact that as $\frac{M}{S}$ increases, more parties compete, and therefore individual parties have lower seat expectations. It is clear that the ratio between district magnitude and assembly size matters, but the way in which the embeddedness parameter k operationalises this effect is insufficient. In any case, the table below reproduces the model bias analysis using these revised predictions of the values of the dependent variables of interest. Note that, confusingly, $\hat{k}$ (k-hat) refers to the expected intercept of the whole base, while the k noted as the dependent variable X refers to the embeddedness parameter.

Table 2: Bias diagnostics of embeddedness-adjusted model.

	District-level models		
	v_1	N_c	v_s
base X	$prM^{\frac{6-k}{4}}$	$prM^{\frac{6-k}{4}}$	prM^{3-k}
expected slope ($\hat{k}$)	-0.250	0.375	-0.250
observed slope (β)	-0.258	0.385	-0.248
$\frac{\lvert\beta-\hat{k}\rvert}{se}$	2.134	2.261	0.707

B.3 Symmetric Regression

As noted in section 5.1 footnote 19, in early tests of logical models, Taagepera and Shugart (Taagepera, 2007; Shugart, Bergman and Watt, 2013) occasionally employed 'symmetric regression', an estimator distinct from OLS insofar as it accounts for error-in-variables on the independent variable (see Taagepera, 2008, pp. 154-175). In a univariate regression, this is equivalent to a Simple Major Axis regression (Kimura, 1992). The estimator has enjoyed virtually no take-up in the wider literature, and presents econometric issues in generalising beyond two dimensions and computing clustered standard errors. In more recent work, Taagepera and Shugart themselves reverted back to log-log OLS to test both inter- (Li and Shugart, 2016; Shugart and Taagepera, 2017, pp. 109–113) and intra-party (Shugart and Taagepera, 2017, pp. 215–235) models of electoral systems. The table below presents model slopes and associated measures of bias for the symmetric regression models, computed on unclustered standard errors as per SBW replication files.

Table 3: Bias diagnostics, symmetric regression used instead of OLS

	List-level			*List-level*			*SBW*	
	v_1	N_c	v_s	v_1	N_c	v_s	v_1	v_s
base X	scp	scp	s^4cp	$prM^{\frac{11}{8}}$	$prM^{\frac{11}{8}}$	$prM^{\frac{5}{2}}$	c	c
exp. slope	-0.250	0.375	-0.250	-0.250	0.375	-0.250	-0.500	-0.750
obs. slope	-0.278	0.407	-0.268	-0.268	0.392	-0.256	-0.459	-0.705
$\frac{\lvert\beta-\hat{k}\rvert}{se}$	5.460	4.110	3.799	3.590	2.547	0.373	8.603	3.711

B.4 Drop-One-Country Subsamples

As discussed in section 5.1 footnote 19, I repeated the model bias analysis on reduced samples obtained dropping alternately one of the nine countries under consideration to assuage concerns that the results are driven by country-specific factors. Sample dependency is not a major concern for the v_s and N_c models: especially in their district-level version they produce consistently coefficients close to the expectation $\hat{k}$. The models for v_s – which, as noted in section 3.1.3 rest on shakier theoretical ground – do however present on occasion fairly large values of the bias indicator.

		List-level			*District-level*		
Country Dropped	variable	v_1	N_c	v_s	v_1	N_c	v_s
	base X	scp	scp	s^4cp	$prM^{\frac{11}{8}}$	$prM^{\frac{11}{8}}$	$prM^{\frac{5}{2}}$
	exp. slope $\hat{k}$	-0.250	0.375	-0.250	-0.250	0.375	-0.250
Belgium	obs. slope	-0.276	0.407	-0.265	-0.264	0.392	-0.238
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	5.668	5.508	7.307	3.650	3.361	3.981
Cyprus	obs. slope	-0.261	0.389	-0.262	-0.252	0.378	-0.237
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	2.621	2.773	6.721	0.521	0.563	4.500
Czechia	obs. slope	-0.254	0.381	-0.259	-0.245	0.369	-0.234
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	1.048	1.142	4.982	1.359	1.294	5.367
Estonia	obs. slope	-0.262	0.390	-0.261	-0.253	0.378	-0.237
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	2.944	2.881	6.428	0.915	0.691	4.607
Finland	obs. slope	-0.256	0.382	-0.259	-0.247	0.371	-0.235
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	1.468	1.475	5.240	0.770	0.918	5.352
Italy	obs. slope	-0.271	0.407	-0.283	-0.266	0.400	-0.262
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	3.132	3.842	15.117	3.022	3.928	3.682
Peru	obs. slope	-0.261	0.389	-0.262	-0.252	0.378	-0.237
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	2.609	2.766	6.390	0.509	0.555	4.729
Poland	obs. slope	-0.261	0.385	-0.255	-0.252	0.373	-0.231
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	2.483	1.852	3.005	0.460	0.521	6.284
Slovakia	obs. slope	-0.266	0.396	-0.261	-0.256	0.381	-0.234
	$\frac{\lvert\beta-\hat{k}\rvert}{se}$	4.407	4.526	6.479	1.622	1.486	7.077

C The Nemčok-Šedo Dataset

I am grateful to Dr Miroslav Nemčok for sharing with me a dataset of electoral system quantities for 560 elections in 40 countries that employ simple electoral systems, which expands on that used by Taagepera (2007) for testing the SPM. Table 1 details the number of election and time range covered, as well as minima and maxima of M (mean district magnitude), S (assembly size), N_S (effective number of parties) and σ_1 (seat share of the largest party), grouped by country.

Table 1: The Nemčok-Šedo dataset of inter-party quantities: descriptive statistics.

Country	# elections	Time Range	M (min-max)	S (min-max)	N_S (min-max)	σ_1 (min-max)
Armenia	2	2007–2012	3.2	131	2.7–3.4	0.49–0.53
Australia	27	1946–2013	1	74–150	2.2–3.2	0.39–0.6
Austria	20	1949–2013	4.3–20.3	165–183	2.1–4.6	0.28–0.52
Belgium	22	1946–2014	6.7–13.6	150–212	2.5–10.1	0.15–0.51
Bulgaria	9	1991–2017	7.7	240	2.4–5.1	0.34–0.57
Canada	23	1945–2015	1	245–338	1.5–3.2	0.4–0.78
Croatia	6	2000–2016	12.6–12.8	151–153	2.7–3.5	0.39–0.53
Cyprus	5	1996–2016	9.3	56	3.5–4.5	0.32–0.36
Czech Republic	8	1990–2013	14.3–25	200	2.3–5.6	0.25–0.62
Denmark	27	1945–2015	5.9–14.9	148–179	3.7–7.2	0.26–0.42
Estonia	7	1992–2015	8.4–9.2	101	3.8–5.9	0.28–0.41
Finland	20	1945–2015	12.5–15.4	200	4.6–5.8	0.22–0.32
France	17	1946–2012	1–4.4	475–618	1.8–6.2	0.2–0.75
Iceland	22	1946–2016	1.9–10.5	52–63	3.2–5.3	0.29–0.42
Ireland	20	1948–2016	3.4–4	144–166	2.4–4.6	0.32–0.57
Israel	20	1949–2015	120	120	3.1–8.7	0.22–0.47
Italy	12	1946–1992	17.4–19.7	556–630	2.6–5.7	0.33–0.53
Japan	12	1960–1993	3.9–4	467–512	2–4.1	0.44–0.64
Latvia	6	1998–2014	20	100	3.9–6	0.23–0.33
Luxembourg	16	1945–2013	12.8–16	51–64	2.7–4.3	0.31–0.5
Macedonia	8	1990–2016	1–20	120–123	1.9–6	0.32–0.72
Malta	6	1996–2017	5–5.3	65–69	2–2	0.51–0.57
Moldova	7	1994–2010	101–104	101–104	1.8–3.4	0.4–0.7
Montenegro	9	1990–2012	5.1–85	71–125	2.1–3.2	0.47–0.66
Netherlands	22	1946–2017	100–150	100–150	3.5–8.1	0.21–0.36
New Zealand	17	1946–1993	1	80–99	1.8–2.2	0.51–0.69
Norway	18	1945–2013	5.2–8.9	150–169	2.7–5.4	0.26–0.57
Poland	8	1991–2015	8.8–12.4	460	2.7–10.9	0.13–0.51
Portugal	15	1975–2015	10–11.9	230–263	2.2–4.2	0.35–0.59
Romania	5	1990–2004	7.9–9.7	332–396	2.2–4.8	0.34–0.66
Serbia	11	1990–2016	1–250	250	1.6–4.9	0.29–0.78
Slovakia	9	1990–2016	37.5–150	150	2.9–6.1	0.24–0.55
Slovenia	8	1990–2014	5.7–11.2	80–90	4.1–8.2	0.17–0.4
South Africa	5	1994–2014	44.4	400	2–2.3	0.62–0.7
Spain	13	1977–2016	6.7	350	2.3–4.1	0.35–0.58
Sweden	21	1948–2014	8.2–12.5	230–350	2.8–5	0.32–0.54
Switzerland	18	1947–2015	7.7–8	194–200	4.7–6.8	0.22–0.32
Turkey	18	1950–2015	4.2–9.1	400–610	1.2–4.9	0.25–0.92
Ukraine	3	1994–2007	1–450	450	3.1–3.4	0.19–0.41
United Kingdom	20	1945–2017	1	625–659	2–2.6	0.47–0.63
United States	18	1948–2016	1	435–437	1.8–2	0.51–0.68

D Precision analysis for v_s predictions

As noted in the discussion of methodological choices(footnote 22 in section 5.2), the analysis presented in section 6.2 is restricted to v_1 and N_c insofar as these quantities have intuitive inter-party analogues in σ_1 and N_S. Diagnostics of deviation-from-prediction can however be conducted on the v_s list-level and district-level models as well. Consistently with the main analysis, the index d is computed as $\log_{10}\left(\frac{v_s}{(s^4cp)^{-1/4}}\right)$ for the list-level model and as $\log_{10}\left(\frac{\tilde{v}_s}{(prM^{\frac{5}{2}})^{-1/4}}\right)$ for the district-level model tested on district medians. Table 1 presents median and mean values of absolute discrepancy, with the associated factors of error, and the share of observations that fall withing the 'tolerable error' band. Figure 1 plots the values of the index d against the base products $(s^4cp)^{-\frac{1}{4}}$ and $(prM^{\frac{5}{2}})^{-\frac{1}{4}}$.

Figure 1: Comparison of deviation from prediction of the district-level model for the means values of the dependent variables: $\bar{v}_1$, $\bar{N}_c$, and $\bar{v}_s$. Dashed lines represent values of d corresponding to values where the observed value is either twice or half the prediction.

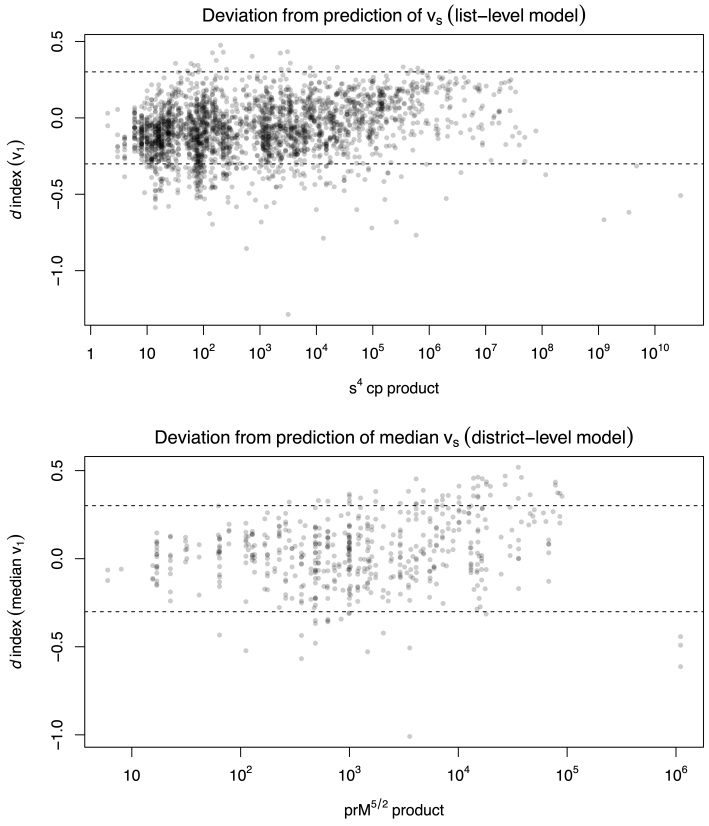
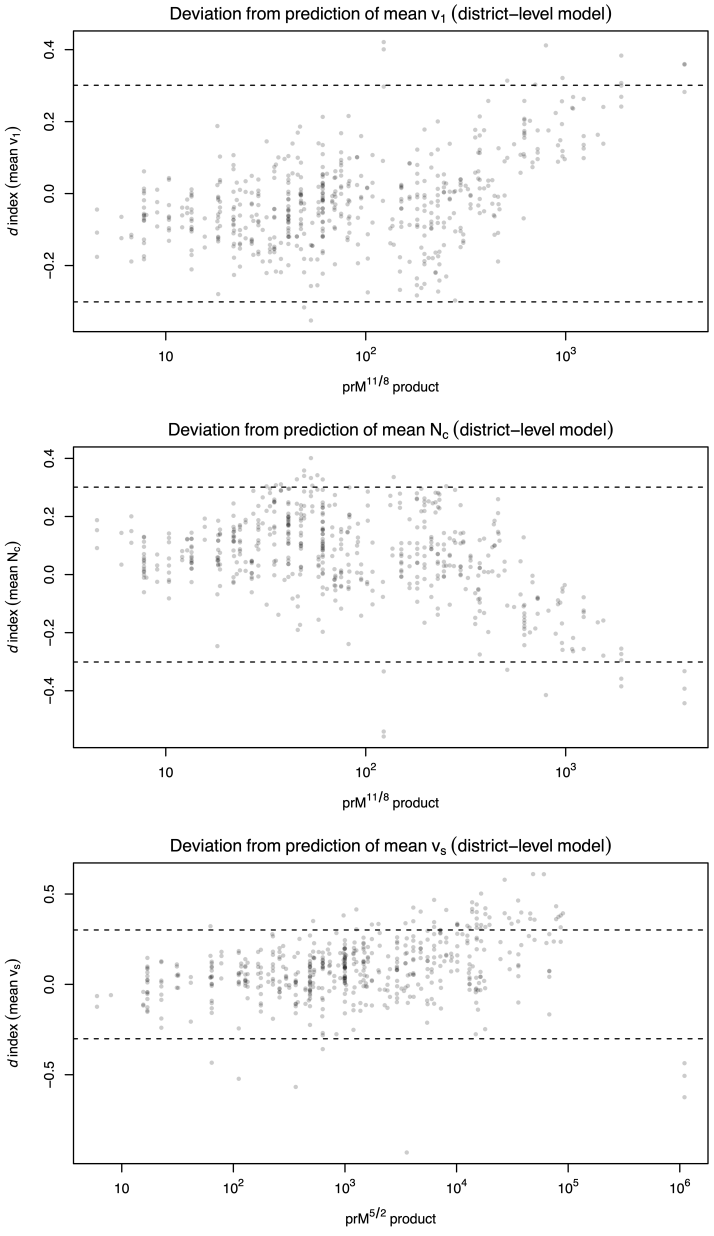


Table 1: Summary indicators of deviation from prediction: models for v_s and $\tilde{v}_s$.

	median of $\lvert d\rvert$		mean of $\lvert d\rvert$		share $d < log_{10}(2)$
	value $\lvert d\rvert$	% error	value $\lvert d\rvert$	% error	and $d > log_{10}(0.5)$
List-level model	0.122	32.5%	0.147	40.2%	90.6%
District-level model	0.121	32.2%	0.152	42%	87.3%

E Precision analysis for district means

Figure 1: Comparison of deviation from prediction of the district-level model for the means values of the dependent variables: $\bar{v}_1$, $\bar{N}_c$, and $\bar{v}_s$. Dashed lines represent values of d corresponding to values where the observed value is either twice or half the prediction.



Precision diagnostics for the district-level model may also be computed on district-level means of v_1, N_c and v_s, as opposed to the median values presented in the main analysis and in section D of the Appendix. Figure 1 plots the distribution of the values of d against the base products $prM^{\frac{11}{8}}$ and $prM^{\frac{5}{2}}$. Diagnostics analogous to those presented in the main analysis are presented in table 1.

Table 1: Summary indicators of deviation from prediction: models for $\bar{v}_1$, $\bar{N}_c$ and $\bar{v}_s$.

	median of $\|d\|$		mean of $\|d\|$		share $d < log_{10}(2)$
	value $\|d\|$	% error	value $\|d\|$	% error	and $d > log_{10}(0.5)$
$\tilde{v}_1$	0.081	20.4%	0.098	25.3%	97.8%
$\tilde{N}_c$	0.109	28.7%	0.126	33.6%	95.8%
$\tilde{v}_s$	0.113	29.6%	0.142	38.8%	88.6%