Journal of East Asian Studies 15 (2015), 199–226

Interpreting Ethnicity and Urbanization in Malaysia's 2013 General Election

Thomas B. Pepinsky

In this article I reinterpret Ng, Rangel, Vaithilingam, and Pillay's analysis in this issue of pro-BN voting in Peninsular Malaysia in Malaysia's 2013 general election. I show that the authors' statistical methods are inappropriate for testing whether district ethnicity predicts district-level BN vote share, and that their modeling choices result in tests of hypotheses that do not exist and cannot be derived from standard theoretical approaches to ethnic voting in Malaysia. I then provide a range of statistical evidence that supports three main conclusions: (1) ethnicity and district area (a proxy for urbanization) both predict BN vote shares at the district level, (2) neither the effect of ethnicity nor of district area can be reduced to the other, and (3) there is no interactive effect between ethnicity and urbanization. These results are in direct contradiction with the authors' results, and apply equally in Peninsular Malaysia and the entire country. I also discuss the broader issues that emerge when testing competing theories of BN vote share. **KEYWORDS:** *ethnicity, urbanization, elections, authoritarianism, Malaysia, statistics*

NG, RANGEL, VAITHILINGAM, AND PILLAY'S ANALYSIS (THIS ISSUE) OF ETHNICITY, urbanization, and pro-regime voting in Malaysia's 2013 general election is an important contribution to contemporary Malaysian political studies. The authors (hereafter NRVP) use advanced statistical techniques to estimate the relationships between ethnic population totals, urbanization, and constituency-level votes for the Barisan Nasional (BN) coalition in Peninsular Malaysia. By interacting a measure of ethnic composition with a measure of district area, they purport "to identify which of . . . two factors, ethnicity or urbanization, provides a stronger explanation for the erosion of the BN's popular votes in GE13 [the 2013 general election]." They conclude that "the *Chinese-Urbanization* factor is having the most dominant influence on the proportion of votes garnered by

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

BN," and also that "whether the constituency is an urban or rural region, an increase in the number of Bumiputera voters in that constituency, ceteris paribus, does not alter the level of support for the ruling coalition."

NRVP's article raises important questions about Malaysian politics, and the way that the authors tackle them has implications for the comparative study of ethnic politics. In the case of Malaysia, ethnicity has been the dominant framework for interpreting Malaysian politics since independence, and the durability of the BN regime has always depended on its ability to amass Bumiputera votes and, in particular, on its ability to mobilize Malay voters in Peninsular Malaysia. One consequence of the BN's strategy is that the percentage of a district's population that is Malay is a powerful predictor of the share of the vote in that district going to the BN. Recently, a wealth of qualitative data—including my own subjective impressions—suggests that urbanized Malays are no longer as closely aligned with the United Malays National Organization (UMNO) and the BN as they once were. If it could be shown that there is no longer a correlation between district-level ethnic composition and BN vote shares, and that some other factor—perhaps modernization, perhaps urbanization, perhaps some other form of social change—had replaced it, then this would be powerful evidence that the customary logic of Malaysian politics had changed in a fundamental way, with implications for the durability of the BN regime and for opposition party strategy.

This is why NRVP's analysis, which emphasizes the importance of urbanization over ethnicity, is so important to our understanding of Malaysian politics. I join NRVP in emphasizing that a comprehensive treatment of the data is necessary, but the details of that analysis matter, and unavoidably involve technical discussions of statistical specification. We must also understand the conceptual issues with "causes of effects" research designs (see Gelman 2011) that aim to adjudicate among different explanations for BN vote share. As I argue below, "horserace" approaches that pit one explanation against another by including both and their interaction in a regression model are not proper tests of competing hypotheses.

In this comment, I present a simpler analysis, one guided by the substantive problem and attentive to the complexity of making inferences from massively interactive models with highly correlated predictors. Some of the discussion below is technical in nature, but this is both unavoidable and essential to understanding how the statistical models relate to substantive questions. Taken together, the evidence supports three main conclusions.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

1. Both district-level ethnic structure and district land area (a proxy for urbanization) predict BN vote shares at the district level.
2. Neither the effect of ethnicity nor that of urbanization can be reduced to the other.
3. There is no interactive effect between ethnicity and urbanization.

These results are in direct contradiction with the authors' results, and apply equally in Peninsular Malaysia and the entire country.

My analysis sounds a note of skepticism that urbanization has moderated—much less superseded—the relationship between district ethnic composition and BN vote share. Instead, it confirms that both ethnicity and urbanization are excellent predictors of BN vote share, which suggests that it would be misleading to select only ethnicity or urbanization for analysis, or to argue that only one and not the other matters. However, if we follow the authors' lead in asking which variable—ethnicity or urbanization—"provides a stronger explanation" for BN vote share, using appropriate tests for competing hypotheses, then ethnicity wins. Every model, every time.

Background

Most of the pertinent details about Malaysia's 2013 general election can be found in NRVP, so I do not repeat them again here.[1] The centerpiece of their analysis is a statistical analysis of the relationship between urbanization, ethnicity, and district-level vote returns. To my knowledge, the first peer-reviewed article in English that used regression analysis to understand ethnicity and vote returns is my own 2009 article in this journal (see Pepinsky 2009). That analysis did not consider urbanization as a competing explanation for patterns of vote returns, so it is imperative to recognize that NRVP's consideration of the competing dynamics of urbanization is an important, necessary step forward. It helps to build a more sophisticated, more nuanced characterization of district-level vote returns than one that can be achieved by looking at ethnicity in isolation.

NRVP's article—in particular, the working paper version[2]—was also part of a lively debate, during and after the election, about urbanization in Peninsular Malaysia and the declining support for the BN. Analysts in the run-up to GE13 emphasized the importance of the UMNO machine in rural areas (see, e.g., Aspinall 2013), and afterward argued that the conduct of the election and its results reflected an urban-rural divide in the Malay electorate (e.g., Aljunied 2013). Given that the BN won the election with a minority of the popular vote, emphasis naturally turned to ger-

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

rymandering, in particular to the rural bias in constituency delineation that tended to favor the BN (e.g., Lee 2013; Ostwald 2013). Nevertheless, there were other voices, such as Kessler (2013), who argued that

> UMNO/BN saw, as some who were not part of its campaign also understood, that the key to the election was the Malay votes. . . . It was conducted in Malay terms and directed to a Malay audience. . . . It was a campaign conducted for the votes of Malays, mainly for those of the great bulk of the more "traditionally-minded" Malays, in the Malay rural heartland areas.

But Kessler's formulation is instructive. Even after decades of urbanization, Malay voters still tend to be rural voters, and the Malay constituencies in which UMNO and the BN needed to win were therefore rural constituencies.

The observation that ethnicity and urbanization covary has profound implications for our ability to disentangle conceptually which one drives support for the BN. Whether using qualitative evidence or statistical modeling, we cannot simply look at rural areas and their tendency to vote BN and conclude that they do so because they are rural, rather than because they are predominantly Malay. This observation also helps to put GE13 in its proper historical political context, for ethnicity and urbanization covary in Malaysia for reasons that are critical for understanding Malaysian party politics—that is, the perceived social and economic hierarchy in colonial Malaya, which featured a largely (but not exclusively) urban Chinese population and a largely rural Malay population. The fact that the Malays were largely rural, and hence "backward," was considered part of the justification for why Malays needed a party like UMNO that would advocate in favor of their interests. It would not have made sense to separate UMNO's rural focus from its Malay focus, for historically they were one and the same, and one justified the other.

This dynamic has not much changed. A party campaigning for Malay votes in a rural district will need to emphasize rural issues. In rural areas, therefore, rural issues happen to also be Malay issues. This is not to ignore the other resources that UMNO and the BN have in rural areas. UMNO is a finely tuned machine with deep reach into rural communities. But of course, these are also Malay communities. We must be careful not to ignore the substantive weight of ethnicity when a party named the United Malays National Organization, founded to represent Malay interests, with a successful and widely known history of campaigning on—and governing on behalf of—Malay interests, campaigns for Malay votes in Malay areas.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Altogether, NRVP's analysis of ethnicity against urbanization is an important addition to the literature on Malaysian voting. But even if it is possible to distinguish between them statistically, in reality, ethnicity and urbanization are part of a single, larger political dynamic in Malaysian politics. With this in mind, I turn now to NRVP's statistical methods.

Statistical Issues

Two particular features of the data guide NRVP's statistical analysis. The first is the limited range of the dependent variable (*BN Vote Share*), which is the ratio of votes obtained by the BN to total votes cast. This variable may logically range from 0 (no votes to the BN) to 1 (all votes to the BN). There are two related issues here. The first is statistical: a linear regression may generate illogical predicted values of the dependent variable that lie outside of the feasible interval of [0,1]. The second is theoretical: it is reasonable to expect that the effect of an increase in Bumiputera population share is different for districts that are 20 percent Bumiputera versus 80 percent Bumiputera. NRVP confront both of these issues using a fractional logistic regression approach (Papke and Wooldridge 1996), which both accounts for the bounded nature of the dependent variable and uses the logit link function to structure the analysis around one natural form of nonlinearity in the effects of independent variables.[3]

There is no doubt that the limited range of the dependent variable could in principle affect inferences. However, I will demonstrate that simple ordinary least squares (OLS) regression performs extremely well in modeling the relationships among ethnicity, urbanization, and vote share, such that employing the fractional logit approach makes no substantive difference to the inferences we draw from the analysis. It is a nice application of generalized linear modeling, but it does not require us to rethink any conclusions that we might have drawn from a simple OLS analysis. One reason that most political scientists use OLS to model vote shares is that fractional regression methods rarely change substantive conclusions unless vote shares of zero appear frequently in the data (see, e.g., the discussion in Gardeazabal 2010).

The second troublesome feature of the data is the nature of district ethnic structure. For each district, there is a breakdown of ethnicity population shares F for each of four key ethnic categories: (F_{Bumi}, F_{Chinese}, F_{Indian}, F_{Other}). This type of data is known as compositional data (Aitchison 1986), and it raises a thorny problem for statistical analysis. Because $F_{\text{Bumi}} + F_{\text{Chinese}} + F_{\text{Indian}} + F_{\text{Other}} = 1$, it must be the case that increasing the

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

share of one group corresponds to a decrease in the share of at least one other group. But when we include each of the four terms as predictors in a regression-type analysis, interpreting coefficients requires a counterfactual statement of the type "an increase in F_i holding all $F_{\sim i}$ constant." We thus have a contradiction, because we cannot logically increase, say, Bumiputera population share while holding other population shares constant.

NRVP confront this challenge by making a substantively important change in how they measure ethnicity. Rather than use F_i, they use the total ethnic population per district, T_i, which they estimate by multiplying F_i by the total number of voters in a district. Because the sums of the total ethnic populations are not constrained to add up to 1, T_i is free from the interpretation challenges associated with ethnic population shares.

The decision to replace F_i with T_i is driven entirely by the problems of using compositional data in regression-type analyses. NRVP note, appropriately, that standard solutions for compositional data involve complex transformations of the problematic independent variables that are both uninterpretable in substantive terms and still more confusing in interaction models. But their solution has the effect of changing the research question at hand from the analysis of the effect of *ethnic composition* to *ethnic population totals*. I am aware of no theory of why districts with higher raw numbers of Bumiputeras, Chinese, Indians, or others in a district would be more likely to vote one way or another, whereas a long line of research and even the most cursory observation of Malaysian politics over the past half century would suggest that the higher the Bumiputera population share, the higher the BN vote share. By measuring ethnic population totals rather than population shares, NRVP predict that Bukit Mertajam constituency in Penang (18.9 percent Bumiputera) would be comparable to Putrajaya (95.5 percent Bumiputera) simply because the total number of Bumiputera voters in each is approximately 15,000! As it turns out, the BN received 18.7 percent of the vote in Bukit Mertajam, and 69.3 percent of the vote in Putrajaya.

My prediction, moreover, emerges logically from a microfounded theory of ethnicity and partisanship in Malaysia. If (a) Bumiputera are more likely to vote for the BN than non-Bumiputera, then (b) ceteris paribus, the higher the proportion of voters in a district that are Bumiputera, the higher the BN vote share. The same prediction does not hold for population totals: even if (a) holds, then *it does not follow* that more voters in a district are Bumiputera, the higher the BN vote share.[4] Replacing F_i with T_i, then, results in a test of a theory that has not been

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

articulated, that does not accord with the realities of Malaysian politics, and cannot even be derived from assumptions about ethnicity and voting behavior at the individual level.

Unnoticed by NRVP is an alternative way forward. There is a simple, theoretically appropriate, and statistically sound modeling strategy for testing the effects of ethnic population shares on BN vote shares. There is no need to enter (F_{Bumi}, F_{Chinese}, F_{Indian}, F_{Other}) into the same regression. When doing so—and for now ignoring the compositional data problem—the result is a test of the effect of, for example, Bumiputera population share relative to other population share, holding Chinese and Indian population shares constant. (This is because one of the four categories will form a reference category, and will be dropped from the regression.) To test the effects of Bumiputeras relative to all others, however, we can simply enter F_{Bumi} alone into a regression. The reference category, now dropped from the analysis, will be all non-Bumiputeras (that is, Chinese, Indians, and others together). We can repeat this for each of the other three categories to produce four regressions, each of which tests whether there is a correlation between one ethnic group's population share and the percentage of votes received by the BN. Doing so preserves the substantive hypothesis about the predictive effects of ethnicity on BN votes, violates no assumptions about coefficient interpretability due to compositional data problems, and can be extended in a straightforward manner to interaction models. The cost is only several milliseconds of computing time.

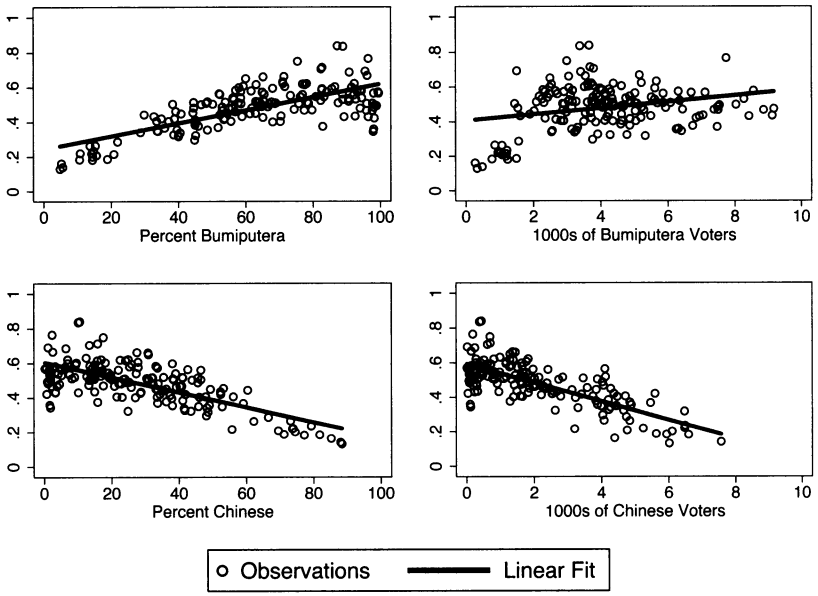
Visualizing Election Results

Before showing those regression results, it is helpful to look directly at the data. In Figure 1, I plot the correlations between BN vote share and percent Bumiputera and percent Chinese (left side), and estimated number of Bumiputera and Chinese voters (right side), using NRVP's own data, which they generously shared with me.

The correlations between percentage BN vote share and percent Bumiputera and percent Chinese are strong and obvious. *No amount of statistical modeling in the rest of this comment will overturn these findings.* However, the correlations between total number of Bumiputera and BN vote share are not as strong. In fact, without the cluster of districts that have both small numbers and small proportions of Bumiputeras, total Bumiputera population would have no predictive power at all over BN vote shares. Note, however, the strong negative correlation between numbers of Chinese voters and BN vote share.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 1 Ethnicity and BN Vote Shares



Source: Data are from NRVP.

Note: This figure displays district-level data from Peninsular Malaysian parliamentary districts that compare Bumiputera and Chinese population shares (left two plots) to BN vote shares and estimated total numbers of Bumiputera and Chinese voters to BN vote shares (right two plots).

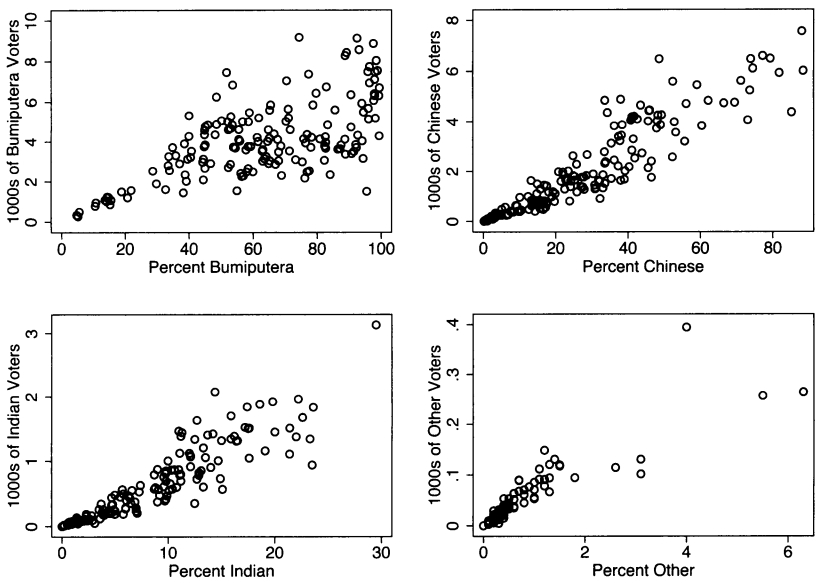
This suggests a strong correlation between population shares and population totals for Chinese, and that is exactly what the data show. In Figure 2, I plot percentages versus population totals for all four ethnic groups.

There is always a correlation between population shares and population totals, but that in the case of Bumiputera, the variance is much larger. This has implications for statistical analysis. When predicting BN vote shares, population totals will be reasonable—albeit imperfect—proxies for the actual theoretical variable, ethnic population share. But it turns out that when using interactive multivariate models, in which eyeballing the data across multiple dimensions is not possible, imperfect proxies will generate misleading inferences.

Before proceeding to the multivariate analysis, I can also examine the relationship between population shares and urbanization. As a proxy for urbanization at the electoral district level, I use district size. It turns out that district size is highly skewed, as Figure 3 shows.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 2 Percent Bumiputera Versus Total Bumiputera



Source: Data are from NRVP.

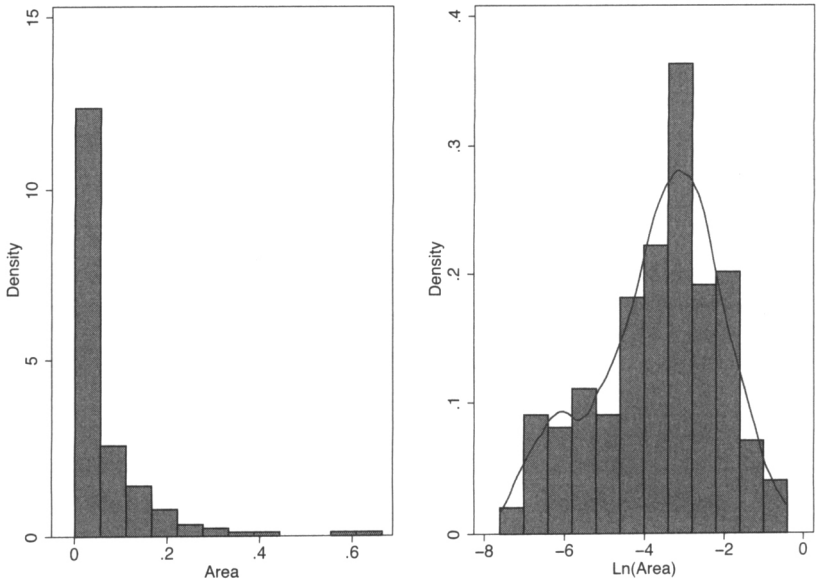
Note: This figure displays district-level data from Peninsular Malaysian parliamentary districts that compare ethnic population shares to estimated total numbers of voters by ethnicity.

However, Figure 3 also shows that the natural logarithm of district size is closer to being normally distributed. I therefore use the natural logarithm of district size as my key measure of how urban or rural an electoral district is.

In Figure 4, I provide scatterplots of ethnic population share for Bumiputera and Chinese and the log of district area. We see that on average, larger (i.e., more rural) districts tend to be more heavily Bumiputera than smaller districts. The reverse is true for Chinese, who tend to be the predominant ethnic group in smaller, more urban districts. The correlations are not perfect, of course. If they were, it would be impossible to distinguish empirically between the effects of ethnicity and urbanization, and all comparisons of the predictive effects of ethnicity versus urbanization are identified statistically by the variation in urbanization that exists for any given ethnic structure. Yet examining the raw data in this way reveals—in a way that regression analysis cannot—that urbanization and ethnicity are highly correlated, and both predict BN vote share.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 3 The Distrubution of District Area



Source: Data are from Greenberg and Pepinsky (2013).
Note: This figure displays district area and the natural logarithm of district area for Peninsular Malaysian parliamentary districts.

Modeling

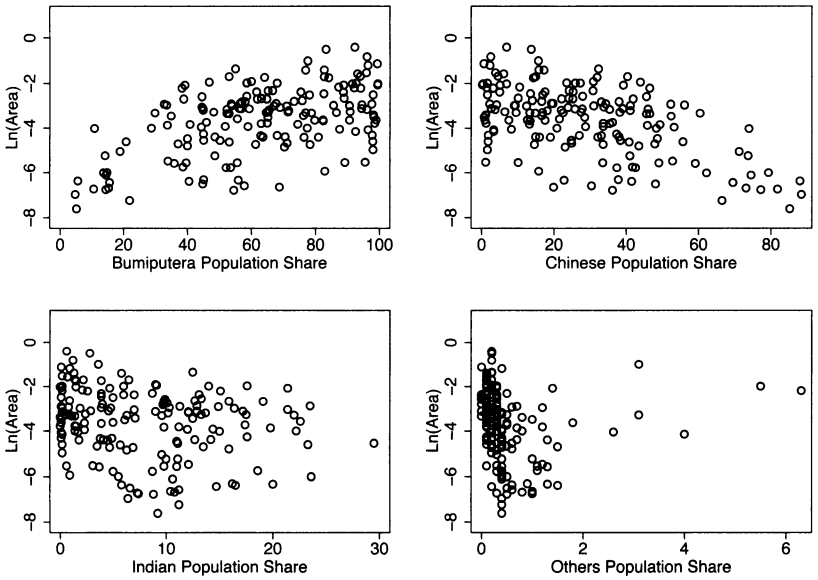
With these visual results in hand, I turn now to a formal statistical analysis. The dependent variable is *BN Vote Share* described above. The central independent variables are *Ln(Area)* to proxy for urbanization and *% Ethnicity*$_i$ (denoted F_i above) for each of the four main ethnic groups to capture district ethnic structure. I examine a series of models that include the urbanization and ethnicity variable independently, additively, and interactively. The full model with interactions, then, is

$$BN\ Share = \beta_0 + \beta_1\ \%\ Ethnicity_i + \beta_2\ Ln(Area) + \beta_3\ \%\ Ethnicity_i \times Ln(Area) + \delta D + \varepsilon$$

Here, D is a vector of state fixed effects, and ε is an error term. I note here that I depart from NRVP by estimating robust standard errors clustered by state (rather than simple robust standard errors) throughout, although this has no substantive impact on the inferences that I draw from the results. More substantively, the state effects D capture any differences

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 4 Ethnic Groups by District Area



Source: Data are from NRVP and Greenberg and Pepinsky (2013).
Note: This figure displays district-level data from Peninsular Malaysian parliamentary districts that compare ethnic population shares to the natural logarithm of district size.

across states that might affect BN vote share. Given that states in the northern "Malay belt," especially Kelantan, have historically been centers of opposition to the BN, and that there is variation by state both in the distribution of district areas and of ethnic composition, including state effects will absorb any state-level factors that threaten my inferences about how ethnic structure and urbanization affect BN vote choice.

I begin by estimating models with only ethnicity and state fixed effects as the independent variables. The results appear as models 1–3 in Table 1.

As expected, ethnic population shares for Chinese and Bumiputera are excellent predictors of BN vote share. Indeed, together with state fixed effects, they alone explain most of the variation in BN vote share in Peninsular Malaysia. Results for Indian population share are markedly less strong, which is consistent with the relatively weak political position of Indian Malaysians. In model 4, I enter *Ln(Area)* as the sole predictor of BN vote share aside from the state dummies. This result too is very strong: larger (more rural) districts yield higher BN vote shares. In mod-

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Table 1 Baseline Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
% Bumiputera	0.01***				0.00***		
	(11.22)				(7.29)		
% Chinese		—0.01***				–0.00***	
		(–13.14)				(–8.40)	
% Indian			–0.01*				–0.01*
			(–2.76)				(–2.73)
Ln(Area)				0.06***	0.03***	0.02**	0.06***
				(11.22)	(4.38)	(3.51)	11.94)
N	165	165	165	165	165	165	165
Adjusted R^2	0.84	0.84	0.39	0.59	0.87	0.86	0.62
AIC	–514.84	–519.13	–297.82	–361.48	–551.12	–545.77	–375.88
BIC	–511.74	–516.02	–294.71	–358.37	–544.91	–539.56	–369.67

Notes: Each model is an ordinary least squares regression with BN vote share as the dependent variable. Each model includes state fixed effects (not reported), and standard errors are clustered by state. T-statistics in parentheses. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

els 5–7, I enter each ethnicity variable together with *Ln(Area)* to test whether the effect of one absorbs the effect of another. The results of these three models are the central findings in this analysis: the strong positive (negative) correlation between Bumiputera (Chinese) population share and BN vote share remains highly statistically significant even when controlling for district area. And the reverse is true as well, with the strong positive relationship between district area and BN vote share remaining highly statistically significant after controlling for each ethnic group's population share.

To summarize the first set of results, a simple analysis of effects of ethnicity and urbanization shows that both are excellent predictors of BN vote share, in ways that are consistent with a commonsense interpretation of Malaysian politics.

At this point the analysis might stop. However, NRVP's preferred approach to modeling the relationship between urbanization, ethnicity, and BN vote share is to interact the predictors, rather than simply entering their effects additively. Why do this? The intuition is that the effects of ethnicity might themselves depend on the level of urbanization. Uncovering these kinds of effects requires interactive models. Note, however, that the nature of the data will make it hard to test every interactive hypothesis. There are no large rural districts that are overwhelmingly Chinese, so while it is possible to calculate predicted BN vote share for a district that is both rural and overwhelmingly Chinese, such a district

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

does not exist (see King and Zeng 2006 for a discussion). These possibilities necessitate care in interpreting the results that we obtain from interactive models, for these calculations may be performed even if they do not make substantive sense.[5]

In Table 2, I show the results of interactive models. Models 1, 3, and 5 are identical to models 5, 6, and 7 in Table 1, and are included in Table 2 again as a reference against which to compare the interactive models.

The results are interesting. When interacting Bumiputera population share with district area, the interactive effect is miniscule and imprecisely estimated. Moreover, the standard errors on the main effect for district area rise substantially. The same nonresults for interactive effects obtain for the other two ethnic population shares, although the main effect for population size remains highly statistically significant. Yet the main effects for ethnic population share remain large and highly statistically sig-

Table 2 Interaction Models

	(1)	(2)	(3)	(4)	(5)	(6)
% Bumiputera	0.00***	0.00**				
	(7.29)	(3.80)				
% Chinese			–0.00***	–0.01***		
			(–8.40)	(–4.14)		
% Indian					–0.01*	–0.01
					(–2.73)	(–1.84)
Ln(Area)	0.03***	0.02	0.02***	0.03***	0.06**	0.07***
	(4.38)	(1.84)	(3.51)	(3.64)	(11.94)	(5.39)
% Bumiputera x *Ln(Area)*		0.00				
		(0.08)				
% Chinese x *Ln(Area)*				–0.00		
				(–1.12)		
% Indian x *Ln(Area)*						–0.00
						(–0.73)
Constant	0.35***	0.34**	0.75***	0.77***	0.78**	0.79***
	(6.31)	(3.91)	(25.37)	(20.62)	(22.30)	(17.23)
N	165	165	165	165	165	165
Adjusted R^2	0.87	0.87	0.86	0.87	0.62	0.62
AIC	–551.12	–549.15	–545.77	–548.78	–375.88	–375.90
BIC	–544.91	–539.83	–539.56	–539.47	–369.67	–366.58

Notes: Each model is an ordinary least squares regression with BN vote share as the dependent variable. Each model includes state fixed effects (not reported), and standard errors are clustered by state. T-statistics in parentheses. *$p < 0.05$; **p < 0.01; ***$p < 0.001$.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

nificant. In short, these results show no evidence whatsoever of an interactive effect of ethnicity and urbanization. Viewed next to the simpler analyses in models 1, 3, and 5, it is clear that the effects of urbanization and ethnicity are better captured as additive effects.

Why are my results so different than those of NRVP? NRVP devote considerable attention to the functional form assumptions and the logical limits on the range of the dependent variable. Is it possible that my use of OLS regression explains my different results? In Table 3, I check by estimating fractional logit equivalents for every OLS model in Table 2.

The fractional logit estimates are substantively identical to OLS estimates. We can also check to see if I obtain massively different—or illogical—predicted values from the OLS models. In Figure 5, I compare the predicted values from model 2 in Table 2 (OLS) and model 2 in Table 3 (fractional logit).

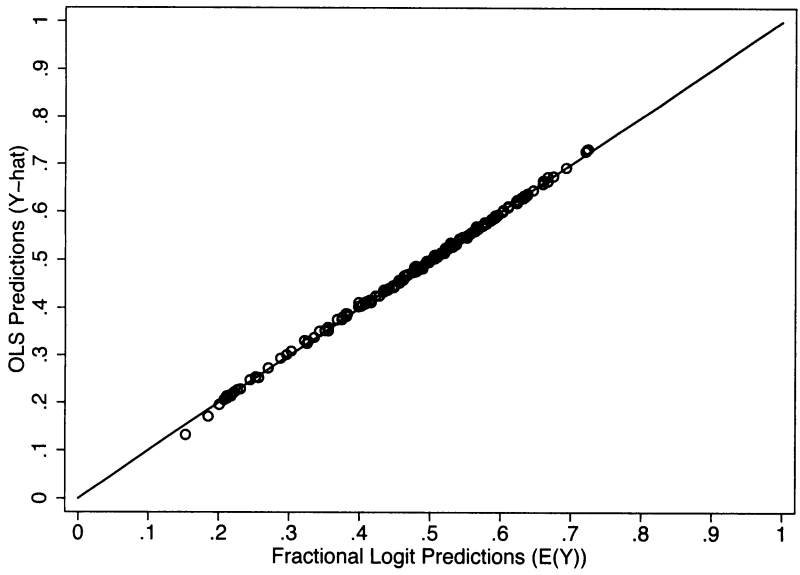
Table 3 Interaction Models, Fractional Logit Estimation

	(1)	(2)	(3)	(4)	(5)	(6)
% Bumiputera	0.02***	0.02**				
	(6.91)	(3.32)				
% Chinese			—0.02***	–0.02***		
			(–7.93)	(–3.65)		
% Indian					–0.02**	–0.03
					(–2.79)	(–1.70)
Ln(Area)	0.11***	0.14*	0.09***	0.10***	0.24**	0.27***
	(4.38)	(2.45)	(3.70)	(2.97)	(10.91)	(5.25)
% Bumiputera x *Ln(Area)*		0.00				
		(–0.58)				
% Chinese x *Ln(Area)*				–0.00		
				(–0.36)		
% Indian x *Ln(Area)*						–0.00
						(–0.62)
Constant	–0.66**	–0.49	1.06***	1.09***	1.17***	1.22***
	(–2.77)	(–1.30)	(9.27)	(6.69)	(7.35)	(6.26)
N	165	165	165	165	165	165
AIC	147.93	149.92	147.90	149.90	150.69	152.65
BIC	154.15	159.24	154.12	159.22	156.90	161.97

Notes: Each model is an ordinary least squares regression with BN vote share as the dependent variable. Each model includes state fixed effects (not reported), and standard errors are clustered by state. T-statistics in parentheses. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 5 Comparing Predictions from OLS and Fractional Logit



Notes: This figure compares OLS predicted values from model 2 in Table 2 to fractional logit expected values from model 2 in Table 3. The forty-five-degree reference line represents the point of equivalence between the two. The figure demonstrates that OLS and fractional logit predictions are nearly identical for nearly every district, and no OLS predicted value lies beyond the logical interval of [0,1].

The predictions are essentially the same, and no OLS predicted values are anywhere close to 0 or 1. There are no grounds to worry that the functional form assumptions of OLS are generating faulty inferences.

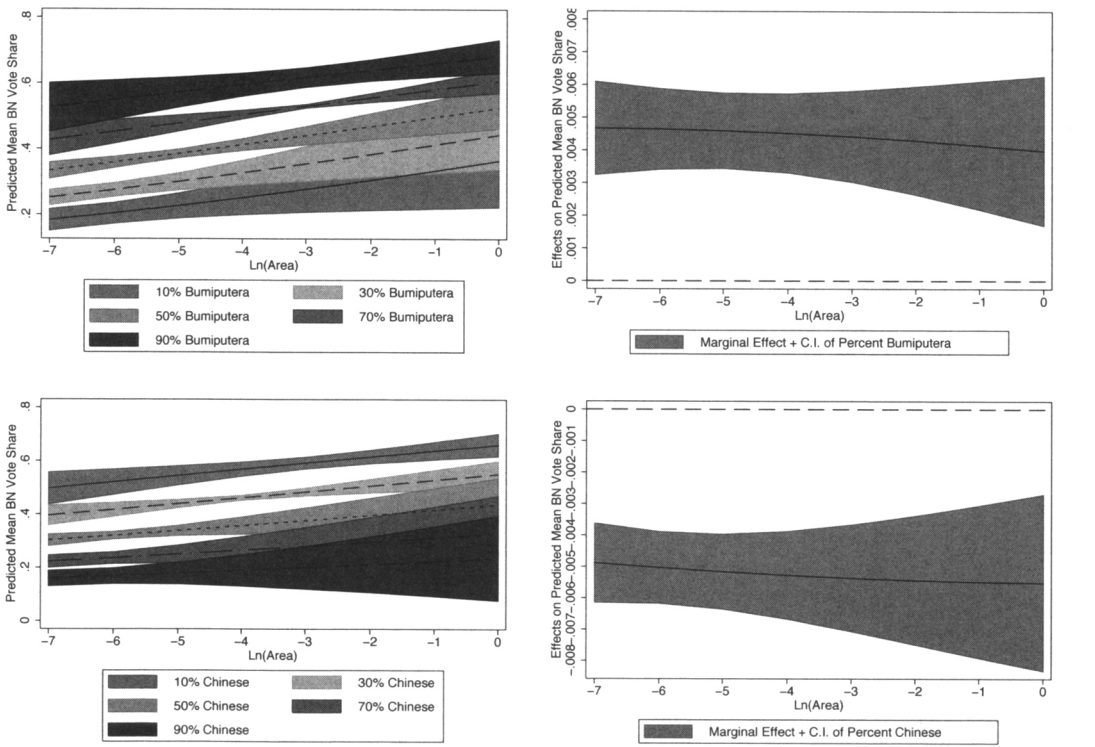
Could it be that I have misinterpreted the results by focusing on regression coefficients? Brambor, Clark, and Golder (2006) remind us that coefficients and standard errors in tabular regression outputs are not easy to interpret. So in Figure 6, I plot both expected values and marginal effects from models 2 and 4 in Table 3, alongside their 95 percent confidence intervals.

Look first at the top two plots. The top left figure plots the predicted BN vote share across the range of values of *Ln(Area)* for different levels of Bumiputera population share. Consistent with the interpretation above, the larger the area, the higher the predicted BN vote share—this is what the upward-sloping lines convey. Furthermore, the higher the Bumiputera population share, the higher the predicted BN population share—this

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 6 Predicted Values and Marginal Effects



Notes: These figures display predicted BN vote shares by district for different Bumiputera and Chinese population shares (left two plots) and the marginal effects of Bumiputera and Chinese population shares (right two plots). Both predicted vote shares and marginal effects are calculated across the range of values of *Ln*(*Area*). The predictions were derived from models 2 and 4 in Table 3.

is what the five separate shaded regions show. More important, the five lines all rise in parallel, which indicates that the effect of urbanization is roughly the same regardless of the value of Bumiputera population share. This conclusion can also be drawn from the top right plot, which shows the marginal effect of an increase in Bumiputera population share across levels of *Ln(Area)*. The line slopes downward a bit, but the range of the predicted marginal effects is always far smaller than the 95 percent confidence band. And the marginal effect of Bumiputera population share is always positive. There is no evidence that the effects of Bumiputera population share depend in any way on district size.

The results for Chinese population share are exactly the reverse. The higher the Chinese population share, the lower the predicted BN vote share, even allowing for the finding that the larger the district area, the higher the predicted BN vote share. Moreover, the marginal effect of Chinese population share is always negative, and while the magnitude increases slightly in larger districts, the range of the predicted marginal effects always lies well within the 95 percent confidence band. Note further the wide confidence intervals around the darkest line, corresponding to the predicted BN vote shares for a 90 percent Chinese district, in large districts. This reminds us that any predictions about the effects of Chinese ethnicity in rural districts should be treated with caution. In sum, the findings from Figure 6 demonstrate once again that both ethnicity and urbanization are strong predictors of vote share, and that there is no evidence of any interactions between the two.

If neither functional form assumptions nor interpretation issues explain the difference between my results and those of NRVP, what does? There are two answers: my use of a more theoretically appropriate and substantively interpretable measure of ethnicity,[6] and my inclusion of state fixed effects *D*. I have already shown that ethnic population shares are more appropriate than ethnic totals, but before proceeding I discuss the importance of accounting for state-specific effects.

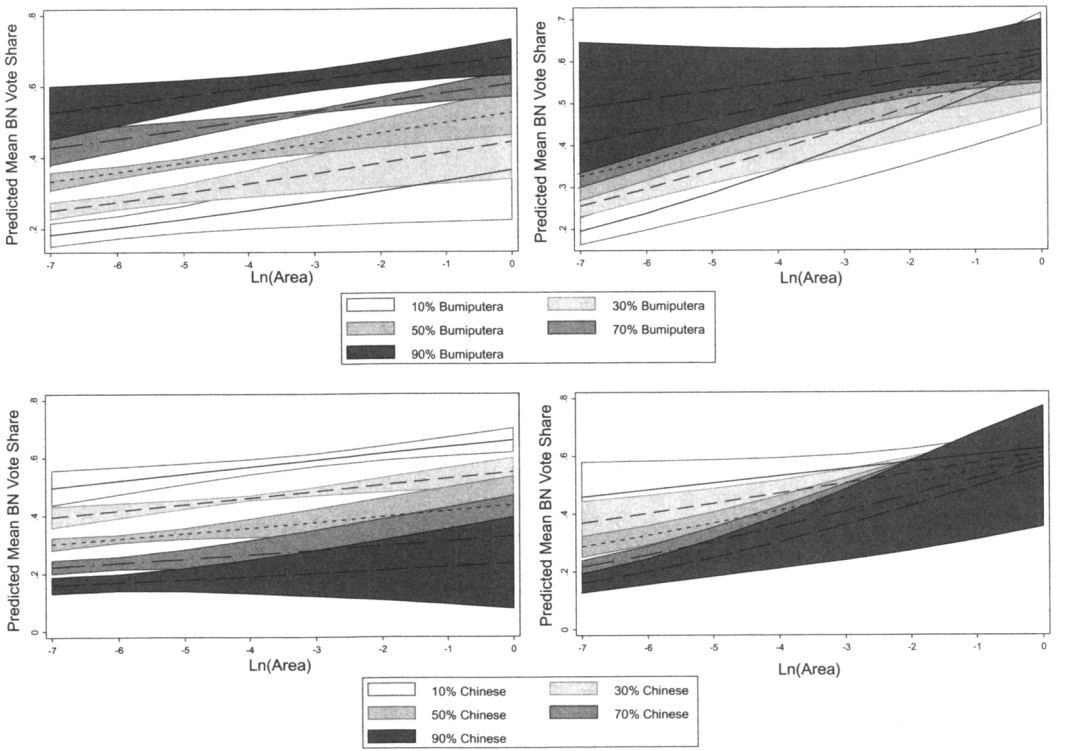
State fixed effects have important consequences for how we interpret the interactive effects of ethnicity and urbanization. In Figure 7, I compare the predicted BN vote shares from models 2 and 4 in Table 3 with the same results obtained from fractional logit models without state effects.

The differences between fixed effects and nonfixed effects models are quite apparent. The effects of ethnicity on BN vote share disappear in larger districts when we ignore state fixed effects, and, furthermore, there is no evidence of an effect of Bumiputera population share on BN vote share for any level of urbanization. Such results might be interpreted

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

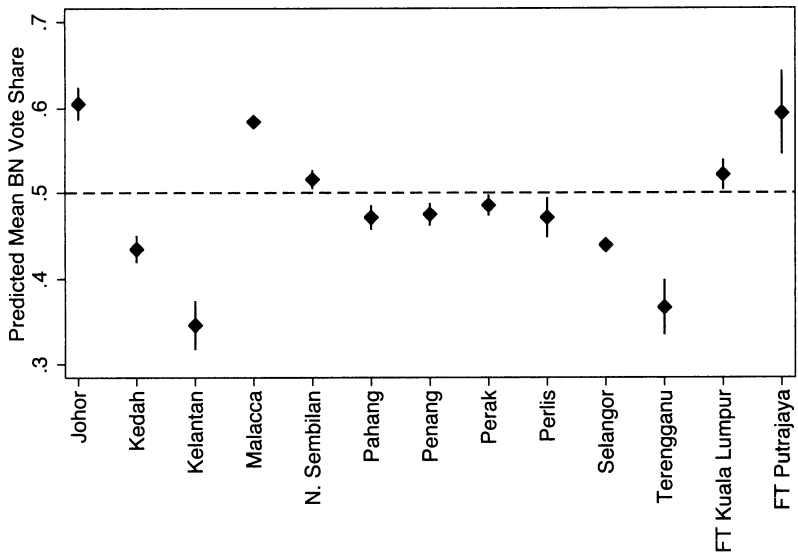
https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 7 Interactive Results, with or Without State Effects



Notes: These figures display predicted BN vote shares by district for different Bumiputera and Chinese population shares across the range of values of *Ln(Area)* for fractional logit models including fixed effects (models 2 and 4 in Table 3, left two plots) and otherwise identical models without fixed effects (right two plots).

Figure 8 Heterogeneity in BN Support by State (Peninsular Malaysia)



Notes: This figure plots predicted BN vote share across states, net of the effects of Bumiputera population share, district area, and their interaction. The predictions were derived from model 2 in Table 3.

as evidence that urbanization matters and ethnicity only affects BN vote share among ethnic Chinese in urban areas, which is broadly consistent with NRVP's results.

However, ignoring state effects deliberately obscures the obvious variation across Peninsular Malaysia in support for the BN. The predicted BN vote shares in 2013 differ dramatically across states, as shown in Figure 8.

And because states differ in their ethnic compositions, we risk attributing the effects of state-specific histories and political conditions to our observed theoretical variables. Large rural districts in Kelantan and Terengganu differ from large rural districts in other states, even if they are all heavily Bumiputera, and accounting for these state-level differences enables a more precise analysis of how ethnicity and urbanization shape BN vote shares.

Horseracing

The analyses shown thus far demonstrate that ethnicity and urbanization both predict vote choice extremely well.[7] This has an effects of

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

causes approach rather than a causes of effects approach (see Gelman 2011), for I have only sought thus far to characterize the predictive power of ethnicity and district area, not to select a cause of the distribution of BN vote shares across Peninsular Malaysian districts. Yet NRVP have a different aim: "the aim of this study is to identify which of the two factors, ethnicity or urbanization, provides a stronger explanation for the erosion of BN's popular votes in GE13." Theirs is a causes of effects approach.

I am sympathetic to NRVP's interest in knowing whether urbanization or ethnicity is a stronger explanation for why Malaysian electoral returns are the way that they are. My personal view, as an observer of Malaysian politics, is that ethnicity is an essential, fundamental factor in Malaysian politics. Yet realism tempers my sympathy for their instinct to view ethnicity and urbanization as competing explanations for Malaysian politics. There is no objective reason to believe that either ethnicity or urbanization is *the* essential driver of Malaysian politics. Instead, I suspect that the instinct to look for effects of urbanization that supersede those of ethnicity is driven by the hope among many Malaysians and political observers for a shift toward a postethnic Malaysian politics, and the belief that statistical analysis of the electoral results might provide evidence that this has taken place.[8]

For an effects of causes research design, multiple regression—when viewed as a way to illustrate causal relationships instead of just as a way to summarize partial correlations—assumes that one set of outcomes can have multiple causes. There is much less agreement about how to formally compare or adjudicate among different causes of effects. For some, the entire endeavor is ill-posed: what does it mean to assert that some explanation is "the cause of" some effect (Gelman and Imbens 2013)? One way to do this is to compare the extent to which two independent variables explain the variation in a dependent variable—in this case, do rural/urban differences explain more about the electoral results than ethnicity does? Unfortunately, in the present application, both explain a lot of variation in BN vote shares.

There are various other kinds of model selection procedures that can be used to select which model does "better" according to some metric, such as comparing R^2 as a measure of fit, comparing Akaike and Bayes information criteria, and the J and Cox-Pesaran tests. Recently, Imai and Tingley (2012) provided a very different way to think about this problem. We have two theories of what determines BN votes at the district level: ethnicity and urbanization. These two theories imply two different hypotheses. The hypotheses are non-nested: ethnicity is not a subset of ur-

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

banization, nor the other way around. Imai and Tingley propose that we can compare any set of theories using finite mixture models to compare the proportion of the cases being analyzed that are "statistically significantly consistent" with one theory versus the other.

So despite my own belief that both ethnicity and urbanization are good explanations for BN vote shares in Peninsular Malaysia, it is possible to follow NRVP, assume that explanations based on ethnicity and urbanization really are mutually exclusive explanations for BN vote share, and then consider the various methods for adjudicating between them. To repeat, this assumption that the two theories compete with one another is a theoretical assumption rather than an empirical result—it also ignores the more comprehensive additive or interactive models—yet in what follows, I proceed under this maintained assumption to see what happens. Unlike NRVP, though, my strategy does not rely on interaction terms,[9] but instead draws on established approaches to model selection and the testing of non-nested hypotheses.

The very simplest way to compare models is to compare the adjusted R^2, or the percentage of the total variation in the dependent variable that is explained by the independent variables (with a penalty applied for complex models that might overfit the data). It is worth pausing to emphasize that comparing R^2 is *very bad statistical practice*, especially from an effects of causes perspective. However, if we interpret the task of comparing theories as measuring the proportion of variance in BN vote shares explained by the different models, adjusted R^2 does this (King 1986, 677–678). We see that in Table 1, adjusted R^2 is higher for model 1 and model 2 (ethnicity) than for model 4 (district area). In a head-to-head contest between ethnicity and urbanization, score one for ethnicity.

More sophisticated model selection procedures for non-nested hypotheses include comparisons of Information Criteria, the J test, and the Cox-Pesaran test. The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are lower in models 1 and 2 and 4. Score one more for ethnicity. The J test and Cox-Pesaran tests, interestingly, are uninformative because *each test rejects both models*.[10] This can happen when both models fit the data well, as is the case here. While this is not a victory for ethnicity over urbanization per se, it does raise another red flag about the wisdom of conceiving of these two theories as mutually exclusive.

Finally, consider the mixture modeling approach proposed by Imai and Tingley. Table 4 displays two quantities from each of two mixture models, one using Bumiputera population share and district area (equiv-

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Table 4 Mixture Model Results

Model	Prior Probability	Number of Observations
Model 1 (Bumiputera)	0.871	143
Model 4 (*Ln(Area)*)	0.129	22
Model 2 (Chinese)	0.854	144
Model 4 (*Ln(Area)*)	0.146	21

Notes: The second column displays the mean of the estimated prior probabilities that each observation is consistent with each model. The third column displays the number of observations that are statistically significantly consistent with each model.

alent to comparing model 1 with model 4 from Table 1), the other using Chinese population share and district area (equivalent to comparing model 2 with model 4).

The second column displays the mean of the estimated prior probabilities that each observation is consistent with models 1, 2, or 4. The third column displays the number of observations that are statistically significantly consistent with model 1, 2, or 4. Together, the results are unambiguous evidence that more district election results are consistent with an explanation based on ethnicity than one based on urbanization. Score these results as the final piece of evidence in favor of ethnicity over urbanization.

I conclude this discussion by emphasizing one more time that *every piece of data that we have* indicates that it is misleading to ask whether *either* ethnicity *or* urbanization explains BN vote shares in Peninsular Malaysia: not just the results from multivariate analyses, which show that both are strong predictors even when in the same model, or a historical perspective that shows how the two variables are conceptually linked, but also additional statistical results comparing multivariate models to the single-explanation models. Additive and interactive models of BN vote share have higher adjusted R^2 and lower AIC and BIC scores than either single explanation model (see the last rows in Table 1 and Table 2). Likelihood ratio tests easily reject both individual models in favor of the additive model (they also fail to reject the additive model in favor of the interactive model). The mixture modeling approach overwhelmingly selects the additive model over either individual model (and also over the interactive model).[11] These results are strong evidence that both ethnicity and urbanization matter; the effects of neither urbanization nor ethnicity can be reduced to the other.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Extending the Analysis Throughout Malaysia

Finally, I extend this analysis to cover all of Malaysia, including the states of Sabah and Sarawak and the Federal Territory of Labuan in East Malaysia in addition to Peninsular Malaysia. To do this, I augment the data on Bumiputera and Chinese population shares and BN vote share from NVRP with data scraped from the website http://undi.info in 2013 (Greenberg and Pepinsky 2013). I then rerun the previous analyses, presenting the key results in Table 5 and Figure 9.

Begin first with Table 5. Comparing models 1 and 2 (Peninsular Malaysia only) to models 3 and 4 (all of Malaysia, identical to models 5 and 6 in Table 1) reveals that Bumiputera and Chinese population shares continue to be strong predictors of BN vote share, net of state effects, when we expand the sample to include all of Malaysia. However, the same is not true for *Ln(Area)*, where the coefficient estimate is not significant at conventional levels. Models 5 and 6 confirm that the same result holds when using fractional logit instead of OLS.

Table 5 Results for All of Malaysia

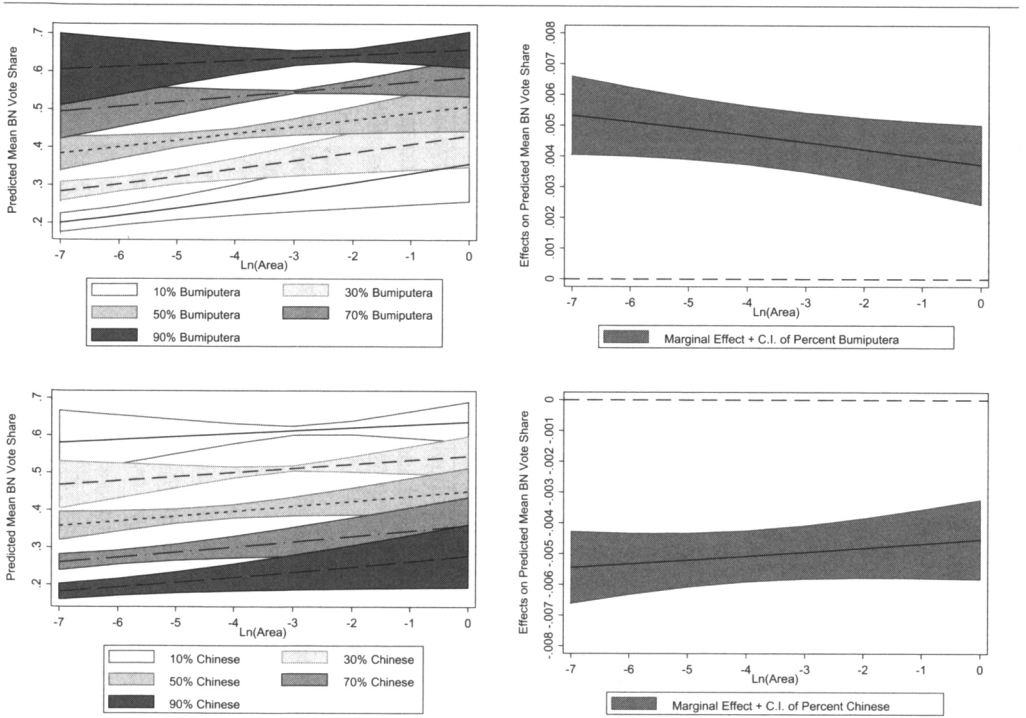
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
% Bumiputera	0.00***		0.00***		0.02***		0.02***	
	(7.29)		(9.28)		(8.92)		(5.52)	
% Chinese		—0.00***		–0.01***		–0.02***		–0.02***
		(–8.40)		(–11.42)		(–11.04)		(–6.50)
Ln(Area)	0.03***	0.02**	0.01	0.01	0.05	0.04	0.13***	0.03
	(4.38)	(3.51)	(1.42)	(1.17)	(1.50)	(1.19)	(3.48)	(0.60)
% Bumiputera x *Ln(Area)*							–0.00	
							(–1.82)	
% Chinese x *Ln(Area)*								0.00
								(0.93)
Constant	0.35***	0.75***	0.26**	0.67***	–1.04***	0.73***	–0.69*	0.71***
	(6.31)	(25.37)	(3.57)	(16.02)	(–3.38)	(4.03)	(–2.43)	(3.42)
N	165	165	222	222	222	222	222	222
Adjusted R^2	0.87	0.86	0.82	0.82				
AIC	–551.12	–545.77	–595.56	–593.19	195.97	195.94	197.89	197.92
BIC	–544.91	–539.56	–588.76	–586.38	202.77	202.74	208.10	208.13

Notes: This model compares results for Peninsular Malaysia only (models 1 and 2) with results from all of Malaysia (models 3–8). Each model uses BN vote share as the dependent variable. Models 1–6 are ordinary least squares regressions, and models 7 and 8 are fractional logit regressions. Each model includes state fixed effects (not reported), and standard errors are clustered by state. T-statistics in parentheses. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Figure 9 Interactive Results, All of Malaysia



Notes: These figures display predicted BN vote shares by district for different Bumiputera and Chinese population shares (left two plots) and the marginal effects of Bumiputera and Chinese population shares (right two plots). Both predicted vote shares and marginal effects are calculated across the range of values of *Ln*(*Area*). The predictions were derived from models 7 and 8 in Table 3, and cover all 222 parliamentary districts in Malaysia.

Models 7 and 8 test the interactive hypotheses, with predicted values and marginal effects displayed in Figure 9. Interestingly, it is only in *these* models where we uncover limited evidence of an interactive effect of urbanization and ethnicity. Specifically, the top right panel demonstrates that while the marginal effect of Bumiputera population share on BN vote share is always positive and statistically significant, there is evidence that the magnitude of this effect decreases when comparing the smallest to the largest districts. This difference is statistically significant at the $p < 0.1$ level. Of course, this interaction does not eliminate the predictive effects of ethnicity on vote share, but it does modestly attenuate the size of that effect in the largest districts.

Conclusion

This article has shown that NRVP's substantive conclusions about ethnicity and urbanization are incorrect, driven by statistical modeling choices that are not appropriate for analyzing the additive and interactive effects of the two explanations for district vote returns. A simpler yet more theoretically precise statistical analysis yields a wealth of findings, but together they point to three conclusions: (1) ethnicity and urbanization both predict BN vote shares at the district level, (2) neither the predictive effects of ethnicity nor those of urbanization can be reduced to the other, and (3) there is no evidence of an interactive effect between ethnicity and urbanization. These results hold both for Peninsular Malaysia and the entire country.

Thomas Pepinsky is associate professor in the Department of Government and associate director of the Modern Indonesia Project at Cornell University. He is author of *Economic Crises and the Breakdown of Authoritarian Regimes: Indonesia and Malaysia in Comparative Perspective* (2009), as well as articles in the *American Journal of Political Science, British Journal of Political Science, Economics and Politics, International Studies Quarterly, World Development, World Politics*, and other venues.

Notes

1. This section draws on an earlier post on my blog, http://tompepinsky.com/2013/05/16/rural-or-malay-contending-perspectives-on-ge13-1/.

2. That working paper version is available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2395091. Its conclusions were more pointed than the

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

current version. It argued that "for any given parliamentary constituency classified as either rural, semi-urban or urban, voters have a similar voting pattern regardless of ethnicity. Therefore, the differences in the voting patterns for BN stems from the urbanisation factor instead" (p. 16).

3. The logit link function imposes a particular nonlinear functional form on the effects of predictor variables. Some readers may not be aware that it, too, is an assumption like any other, made for convenience and interpretability rather than explicitly grounded in a theory. Thus NRVP's observation that an OLS regression assumes linear effects is true, but it is not an argument *tout court* against using OLS rather than fractional logit, which replaces this linearity assumption with a different assumption about the form that nonlinearity takes. See Aldrich and Nelson (1988, 24–37) for a full discussion.

4. It is also not the case that (b) logically entails (a). It is possible that districts with higher Bumiputera population shares have higher BN vote shares for reasons other than a pro-BN bias among Bumiputeras. It could be, for example, that non-Bumiputera voters unanimously vote for the BN *only if* they are small minorities. Or it could be that Bumiputeras happen to live in rural areas, and rural voters vote for the BN. The district-level aggregate patterns cannot resolve these competing theories. This problem of uncovering individual behavior from collective behavior is known as the ecological inference problem, and has been the subject of intense study for decades (Kousser 2001). For one provisional attempt to solve the ecological inference problem in the context of Malaysia's 2008 election, see Pepinsky (2009).

5. Of course, the same is true for additive models as well, but the subtleties of interpreting interactive models appear to generate particular challenges in interpretation.

6. One might still wonder about the correlations between district population totals (which is one component of NRVP's measures of ethnic population totals) and BN vote share. In separate results, available upon request, I can demonstrate that accounting for district population total (either alone or in a triple interaction with both ethnicity and district area) has no substantive consequences for inferences about ethnicity and urbanization.

7. This section draws on an earlier post on my blog, http://tompepinsky.com/2013/05/18/rural-or-malay-contending-perspectives-on-ge13-2/.

8. Eric Thompson (2013) uses the term "urban chauvinism" to describe some of the interpretations of the results of GE13 that emphasize an urban-rural divide. I highlight this here as a reminder that nonethnic explanations for GE13 results are no less subject to normative biases than are explanations that highlight patterns in district ethnicity and BN vote share.

9. Indeed, while NRVP explicitly state that they wish to "identify which of the two factors, ethnicity or urbanization, provides a stronger explanation for the erosion of BN's popular votes in GE13," it is not immediately clear how any of their statistical analyses actually answer that question.

10. Results are available from the author upon request.

11. Results for mixture models and likelihood ratio tests are available upon request from the author.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. New York: Chapman and Hall.

Aldrich, John H., and Forrest D. Nelson. 1988. *Linear Probability, Logit, and Probit Models*. Newbury Park: Sage Publications.

Aljunied, Khairun. 2013. "How Malays Voted at GE13." New Mandala, May 9. http://asiapacific.anu.edu.au/newmandala/2013/05/09/how-malays-voted-at-ge13.

Aspinall, Edward. 2013. "Triumph of the Machine." Inside Story, May 9. http://insidestory.org.au/triumph-of-the-machine/.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14, 1: 63–82.

Gardeazabal, Javier. 2010. "Vote Shares in Spanish General Elections as a Fractional Response to the Economy and Conflict." Economics of Security Working Paper 33. www.diw.de/documents/publikationen/73/diw_01.c.356877.de /diw_econsec0033.pdf.

Gelman, Andrew. 2011. "Causality and Statistical Learning." *American Journal of Sociology* 117, 3: 955–966.

Gelman, Andrew, and Guido Imbens. 2013. "Why Ask Why? Forward Causal Inference and Reverse Causal Questions." National Bureau of Economic Research (NBER) Working Paper 19614.

Greenberg, Sarah, and Thomas B. Pepinsky. 2013. "Data and Maps for the 2013 Malaysian General Elections." Working Paper, Department of Government, Cornell University.

Imai, Kosuke, and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56, 1: 218–236.

Kessler, Clive J. 2013. "Malaysia's GE13: What Happened, What Now? (Part 1)." New Mandala, June 12. http://asiapacific.anu.edu.au/newmandala/2013/06/12/malaysias-ge13-what-happened-what-now-part-1/.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30, 3: 666–687.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14, 2: 131–159.

Kousser, J. Morgan. 2001. "Ecological Inference from Goodman to King." *Historical Methods* 34, 3: 101–126.

Lee, Hock Guan. 2013. "Steadily Amplified Rural Votes Decide Malaysian Elections." *ISEAS Perspective*, no. 34.

Ng, Jason Wei Jian, Gary John Rangel, Santha Vaithilingam, and Subramaniam S. Pillay. "2013 Malaysian Elections: Ethnic Politics or Urban Wave?" *Journal of East Asian Studies* (this volume).

Ostwald, Kai. 2013. "How to Win a Lost Election: Malapportionment and Malaysia's 2013 General Election." *The Round Table* 102, 6: 521–532.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press

Papke, Leslie E., and Jeffrey M. Wooldridge. 1996. "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates." *Journal of Applied Econometrics* 11, 6: 619–632.

Pepinsky, Thomas B. 2009. "The 2008 Malaysian Elections: An End to Ethnic Politics?" *Journal of East Asian Studies* 9, 1: 87–120.

Thompson, Eric C. 2013. "GE13 and the Politics of Urban Chauvinism." New Mandala, May 14. http://asiapacific.anu.edu.au/newmandala/2013/05/14/ge13-and-the-politics-of-urban-chauvinism.

https://doi.org/10.1017/S1598240800009346 Published online by Cambridge University Press