Using Qualitative Information to Improve Causal Inference*

Adam N. Glynn[†] Nahomi Ichino[‡]

March 25, 2013

Abstract

We show that with the Rosenbaum (2002, 2009) approach to observational studies, we can formally incorporate qualitative information into quantitative analyses to improve causal inference in three ways. First, we can ameliorate the effects of poorly measured outcomes by including qualitative information on outcomes within matched sets, sometimes reducing p-values. Second, additional qualitative information across matched sets enables the construction of qualitative confidence intervals on effect size. Third, qualitative information on unmeasured confounders within matched sets reduces the conservativeness of Rosenbaum-style sensitivity analysis. This approach accommodates small to medium sample sizes in a nonparametric framework, and may therefore be particularly useful for analyses of the effects of institutions in a given set of countries or subnational units. We illustrate these methods by examining the effect of using plurality rules in transitional presidential elections on opposition harassment in 1990s sub-Saharan Africa.

*Earlier versions of this paper were presented at the Institute for Qualitative and Multimethod Research Workshop at Syracuse University, June 23–24, 2012, and the 29th Annual Meeting of the Society for Political Methodology, July 19–21, 2012. Julie Faller, Amanda Pinkston, and Jisu Yoo provided able research assistance. We thank Konstantin Kashin, Matthew Blackwell, Nick Weller, Alberto Abadie, and Teppei Yamamoto for helpful comments.

†Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@fas.harvard.edu

‡Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. nichino@gov.harvard.edu

1 Introduction

Observational studies in political science are often beset by problems that can lead to fragile and biased estimates of causal effects. Most fundamentally, when treatment cannot be randomly assigned, important confounding variables that affect both the treatment variable and the outcome variable may be unmeasured. Even measured confounding and outcome variables may only be poorly measured. Moreover, many of these observational studies are also "medium-n," having fewer observations than is needed for large-sample techniques to provide accurate approximations.

This sample size problem also afflicts more large-n studies than is generally recognized. Large-n datasets often contain units that are incomparable on measured confounding variables, and this lack of overlap between treatment and control units results in analyses that rely upon extrapolation for causal inference. We may guard against this by restricting a study to a smaller set of similar observations (Brady & Collier 2004) or remove these incomparable observations from the dataset by pre-processing the data through matching (Ho, Imai, King & Stuart 2007). But what often remains after limiting the scope of the analysis in this way is a medium-n study.

This paper presents a set of methods to improve causal inferences in medium-n studies through a formal synthesis of qualitative information and quantitative analysis. We adopt the Rosenbaum (2002, 2009) randomization inference-based approach to observational studies which enables non-parametric inference with small sample sizes. We initially employ the basic version of this approach which uses pairs of units that have been matched on measured confounders, as this simplifies the presentation and allows for an analogy to a repeated use of the comparative method (Lijphart 1975). We subsequently demonstrate that these techniques can be extended to some of the more complicated matching strategies in Rosenbaum (2002, 2009).

This approach can integrate qualitative information obtained from case studies with a quantitative analysis to improve causal inference in three ways. First, we can ameliorate the effects of poorly measured outcomes with the inclusion of ordinal qualitative information on outcomes within the matched sets. We demonstrate that this can reduce p-values. Second, additional qualitative information on the ranks of the sizes of the within-set differences allows us to present *qualitative confidence intervals* – that is, qualitative descriptions of effect sizes that have the same properties as conventional confidence intervals. We note that nonmetric scaling is not feasible with this amount of qualitative information (Kruskal 1964). Third, qualitative information on unmeasured confounders within matched sets enables a sensitivity analysis that is less conservative than the typical Rosenbaum-style sensitivity analysis.

In showing how to incorporate qualitative information on the outcome and confounders to improve causal inference, we focus on treatment effects for a binary treatment. However it is straightforward to adopt our methods for other causal questions such as treatment effects of continuous treatments or multiple treatments and interactions or for multiple outcomes in this framework.

We demonstrate these points through a medium-n analysis of whether using plurality rules in transitional presidential elections in sub-Saharan Africa in the 1990s increased the severity of opposition harassment in the period leading up to the election. The appendices present the qualitative information from the comparative cases studies that is incorporated into this running example. We find evidence strongly suggestive of this positive effect of plurality rules on opposition harassment, even after accounting for threats to causal inference.

Our method takes advantage of the in-depth investigations of individual observations made feasible by smaller sample sizes (Brady & Collier 2004, Brady, Collier & Seawright 2006), while recognizing that case studies may be time-consuming and potentially ever more detailed. The method delineates how much qualitative information on the outcome and confounders and on which specific sets of cases will be most useful for improving causal inference. Therefore, it implies a statistically principled guideline for how researchers should direct their efforts and resources in conducting case and qualitative comparative studies. With our running example, we show how specific, limited comparisons of a small number of country sets can offer sufficient qualitative information to substantially improve our causal inferences.

This method also differs from existing approaches to "mixed methods" for bolstering quantitative analyses with qualitative case studies. In many instances, case studies are used to illustrate an argument and provide a "plausibility check" (George & Bennett 2005, Fearon & Laitin 2008, Dunning 2012). Lieberman (2005) suggests a nested approach in which an unsatisfactory large-n analysis is followed by a model-building small-n analysis. In these approaches, the case study stands apart from and accompanies the quantitative analysis, rather than being merged into a unified analysis as in our method. Our approach is also more flexible than other formalized procedures for integrating qualitative information, such as Herron & Quinn (2009), which assume binary outcomes or parametric models and often require the elicitation of Bayesian priors.

The paper proceeds as follows. The next section introduces our running example of transitional presidential elections in 1990s sub-Saharan Africa, the formal notation, and randomization inference for pair-matched binary outcome data. Then in each of the following sections, we introduce qualitative information into our analysis to elaborate on our formal mixed method procedure for improving causal inference in medium-n studies. We first incorporate within-pair information on the outcome through the sign statistic and then between-pair information on the outcome through the signed-rank statistic to generate p-values and qualitative confidence intervals. We then discuss how qualitative information on unmeasured confounders reduces the conservativeness of Rosenbaum-style sensitivity analysis. We show how full matching and the Quade statistic can further reduce p-values, and lastly how qualitative information on confounders in matched sets with multiple controls can reduce the sensitivity of the analysis.

2 An Illustrative Example and Notation

2.1 Electoral Rules in Transitional Presidential Elections

To demonstrate these methods, we explore the effect of plurality electoral rules on opposition harassment in multi-party presidential elections that marked transitions away from authoritarian rule in sub-Saharan Africa in the 1990s. These transitional elections were watershed moments in which citizens of these countries, often for the first time in their lives, had the opportunity to replace an authoritarian incumbent at the ballot box. But they were also periods of uncertainty and many potential dangers in which authoritarian incumbents might employ violence against the opposition in order to stay in power.

Twenty-four sub-Saharan countries held these transitional elections in the 1990s, and four of these used plurality rules under which a candidate must obtain more votes than any other candidate in order to be declared the winner.[1] The other countries used some form of runoff rules, which stipulate that should no candidate pass a given vote share threshold (usually 50%) in the first round, weaker candidates are eliminated and the top two finishers compete in a second-round election.[2] This rule and other elements of the election framework were determined by the authoritarian incumbent, with varying degrees of input from opposition representatives and civil society groups through national conferences and constitutional review committees. Foreign constitutional scholars, social scientists, and other experts on democratic institutions were often sponsored by foreign donors' democracy promotion programs and these experts offered advice (Nwajiaku 1994, van Cranenburgh 2011). Since *ex ante* we have reasons to believe that plurality rules increase opposition harassment, on which we elaborate below, our question is whether using plurality rules affected the likelihood and intensity of opposition harassment in these countries' transitional elections.

We start with an incumbent authoritarian regime that has agreed to hold multi-party presidential elections in the face of pressures for political liberalization. The regime wants to hold onto power by having its favored candidate win the election, who may but need not be the incumbent president himself. The incumbent authoritarian regime allocates its finite resources to a combination of opposition co-optation and harassment with the aim of winning this election. We assume that harassment cannot not reliably convert opposition supporters into supporters of the regime's favored candidate, and that harassment can suppress voting by some but not all opposition supporters.[3]

While all are aware of widespread dissatisfaction with the regime, not enough information is available about support for specific challengers to the authoritarian incumbent to ensure Duvergerian coordination in the transitional elections. This means that under plurality rules, a potential challenger who does not have the resources to win a majority but might be able to win a plurality

[1]Although Nigeria's electoral rules did not have a provision to eliminate any candidates, we have not coded this country as a plurality country because only two political parties were permitted to compete in the elections. Including Nigeria as a plurality country in the analysis increases the statistical significance of all results.

[2]These four countries are Cameroon, Kenya, Malawi, and Tanzania. In Kenya, the winning candidate must also receive a minimum of 25% of the valid votes cast in at least 5 of the 8 provinces of the country.

[3]This is because the willingness of a substantial portion of the population to oppose the regime has been demonstrated and is now common knowledge following the riots, strikes, and other costly collective actions that led the regime to accede to multi-party elections. Potential candidates and voters may be willing to endure some harassment to oust the authoritarian incumbent at the ballot box.

might compete in the election and divide opposition support, reducing the vote margin needed to win the election. For the incumbent authoritarian regime, this makes opposition harassment more likely to be decisive for the outcome of the election and a more attractive strategy, particularly if the harassment can be targeted at the supporters of the opposition candidate who is likely to have the most support.

With a runoff provision, the incumbent authoritarian regime could aim to place in the top two rather than try to win a majority of votes cast in the first round. But this strategy is dangerous because the opposition would gain the opportunity to coordinate behind a single candidate for the second round and the regime's favored candidate might place third and be ineligible for the runoff election. Therefore the incumbent regime's strategy will try win an outright majority in the first round by drawing potential challengers and their supporters into its coalition, which in turn encourages weak challengers to contest the election in order to be co-opted by the regime, even if they do not have the resources to muster a majority. Opposition harassment could help the incumbent by reducing turnout and therefore the number of votes needed to comprise a majority, but resources would need to be diverted from co-optation. Moreover, unlike under a plurality rule where harassment can change the threshold for an incumbent win, with a runoff rule, harassment does not change the requirement of a majority. This means that opposition harassment is relatively less effective than co-optation under runoff rules and is less likely to be decisive. Consequently, we expect plurality rules to lead to greater opposition harassment than runoff rules.

Note that empirical study of this proposed plurality effect has several difficulties common to many observational studies. In addition to the small sample size, we are likely to have significant unmeasured confounding because we do not have the information that was available to the key actors who set this electoral rule or how they weighed different considerations. In particular, strong opposition to the incumbent might increase the amount of opposition harassment under either set of electoral rules and might also increase the likelihood of using plurality rules. Moreover, and most basically, the outcome variable of opposition harassment is difficult to measure. We address these concerns throughout the remainder of the paper.

2.2 Notation and First Analysis

Suppose we want to make causal inferences regarding N_1 treated units ($T = 1$) and a comparable subset of $N_0 \geq N_1$ control units ($T = 0$). For illustrative purposes, we follow Rosenbaum (2002) and initially assume that the N_1 treated units have been pair-matched without replacement to N_1 of the control units. In addition, we initially assume that the outcome variable has been coded for pairs $s = 1, ..., N_1$ so that the outcome for the first unit in each pair is denoted Y_{s1} and the outcome for the second unit is denoted Y_{s2}. We denote T_s to be the treatment condition for the first unit in each pair and $1 - T_s$ to be the treatment condition for the second unit in the pair. We assume that causal effects are well defined for each individual as the difference between two potential outcomes or counterfactuals, the outcome if treatment had been received, $Y(1)$, and the outcome if control had been received, $Y(0)$. We also assume that the observed outcome Y is equal to the potential outcome corresponding to treatment T; the other potential outcome is unknown. Therefore, for pair s, $Y_{s1} = T_s \cdot Y_{s1}(1) + (1 - T_s) \cdot Y_{s1}(0)$ and $Y_{s2} = T_s \cdot Y_{s2}(0) + (1 - T_s) \cdot Y_{s2}(1)$.

For the $2 \cdot N_1$ units in the matching study, the causal effects are written as:

$$\tau_{s1} = Y_{s1}(1) - Y_{s1}(0), \text{ and}$$
$$\tau_{s2} = Y_{s2}(1) - Y_{s2}(0), \text{ for } s = 1, ..., N_1$$

Like many observational studies, we begin our analysis with data from a publicly available dataset. The National Elections across Democracy and Autocracy (NELDA) dataset (Hyde & Marinov 2012) covers our population of interest, and we draw on this dataset to code an outcome variable that takes the value 1 if the opposition is harassed in the run-up to the election, and 0 otherwise. Data and sources are described in Appendix A. Because weaker incumbents who face strong opposition and are more worried about obtaining a majority are probably less likely to adopt a runoff provision that demands a majority, we pair-match the four plurality countries ($T = 1$) to the four countries with runoff provisions ($T = 0$) that are the most comparable on predictors of this institutional choice. Appendix B provides the matching details, but we highlight that the plurality countries were exactly matched on the basis of whether the transition follows civil conflict, whether the country had previous experience with military rule, and the level of protest during the transition period. They were also matched on ethnic fractionalization and the log of GDP per capita. These four matched pairs are presented in Table 1, along with their potential outcomes. Note that these countries have been paired in previous comparative studies (Azevedo, 1995, for Cameroon-Gabon; Widner, 1994a, b, c for Kenya–Côte d'Ivoire; Posner, 2004, for Malawi-Zambia; Smith, 2005, for Tanzania–Guinea-Bissau).

Treated (Plurality)	$Y(1)$	$Y(0)$	Controls (Runoff)	$Y(1)$	$Y(0)$
Cameroon	1	?	Gabon	?	0
Kenya	1	?	Côte d'Ivoire	?	1
Malawi	1	?	Zambia	?	0
Tanzania	0	?	Guinea-Bissau	?	0

Table 1: *Potential Outcomes for Matched Pairs*

First, note that as discussed above, the potential outcome under treatment is observed for the plurality countries, while the potential outcome under control is unknown. Analogously, the potential outcome under control is observed for the runoff countries, while the potential outcome under treatment is unknown. Second, note that we inspect the outcome variable only after we match control units to our treated units, and further, that information on the outcome variable for the control units that are not matched does not contribute to our analysis. This significantly reduces the potential coding burden. If the NELDA dataset had not been available and we needed to code the outcomes ourselves for even just the initial analysis, we would only have coded the outcome for these 8 countries in the matched pairs rather than all 24 countries. With the NELDA coding, the difference in outcomes between plurality and runoff countries is positive (2/4), indicating that plurality electoral rules may have caused opposition harassment in these transitional presidential elections.

Because the sample size is small, even if we believe that the matching successfully removed confounding and that in each pair the treated unit and control unit had the same *ex ante* probability of being assigned to treatment, we wonder whether the result could simply be due to chance. A straightforward approach to answering this question is Fisherian randomization inference, which is discussed in detail in Rosenbaum (2002, 2009), and by Bowers and Panagopoulos (2009, 2011), Hansen & Bowers (2008), Ho & Imai (2006), and Keele, McConnaughy & White (2008) in political science. We will consider the assumptions required to use randomization inference in observational studies in later sections. For now, we can think about the following hypothetical question: if we had flipped a coin for each pair to determine which unit would receive treatment and which unit would receive control, then would we find the evidence in the table convincing? This hypothetical question is typically formalized with a test of the sharp null hypothesis of no effect for any unit:

$$H_0 : \tau_{s1} = \tau_{s2} = 0, \text{ for } s = 1, ..., N_1$$

Under this null hypothesis and an assumption of pairwise randomization, we can generate null distributions and p-values by permuting over all possible pairwise randomizations. For our example with four matched pairs, there are $2^4 = 16$ possible pairwise randomizations. Using McNemar's test for binary outcomes, a special case of a randomization test using a sign score statistic, and no additional information on the outcome variable, we obtain a one-sided p-value of $4/16 = 0.25$. In the next section, we explain the logic behind these randomization tests with the sign statistic and the signed-rank statistic for pair-matched data, and show that this approach allows us to incorporate qualitative information on the outcome variable to improve the analysis.

3 Using Qualitative Information on the Outcome

When it is not possible to accurately measure on an interval scale an outcome that is conceptualized correctly as a one-dimensional interval variable, this outcome may be coded as dichotomous or ordinal. This coarse coding may be necessary when creating a multi-use data set, but it may waste available information and lead to the wrong conclusions in a particular analysis. In this section, we present a method for incorporating additional qualitative information on the outcome to improve inferences about whether a particular treatment has an effect and how large this effect may be. Applying this method decreases the p-value for our analysis.

In our example, instances of opposition harassment may differ in the number of people detained, their treatment under detention, whether the regime targeted opposition leaders or supporters or both, whether violence was only threatened, and the extent of the violence if employed. Measurement of this variable may be improved by attending to these components, but we may lack consensus on how much weight should be given to each component when constructing an overall measure of opposition harassment. It may be difficult to obtain the data to construct and place each of these countries on a scale of the severity of opposition harassment, even if there were agreement. However, because we have matched treated units to control units, even small amounts of this information can increase the power of tests of the sharp null hypothesis.

3.1 Incorporating Within-Pair Information

A number of sign score statistics can be used with pair-matched data, but the sign statistic is the easiest to improve through the addition of qualitative information. As indicated by its name, the sign statistic uses the sign of the difference in outcomes for each pair ($sign(Y_{s1} - Y_{s2})$ for $s = 1, ..., N_1$). The statistic is written as:

$$V = \sum_{s=1}^{N_1} [T_s s_{s1} + (1 - T_s) s_{s2}] \; = \sum_{s=1}^{N_1} V_s$$

where $s_{s1} = 1$ if $Y_{s1} > Y_{s2}$ and $= 0$ otherwise, and $s_{s2} = 1$ if $Y_{s2} > Y_{s1}$ and $= 0$ otherwise. This is the statistic implicitly used for McNemar's test in the previous section.

With this statistic V, we return to our hypothetical question: how likely is it that we could have observed the result just by chance, that is, if plurality electoral rules had been assigned by a coin flip within each pair? We calculate this p-value using Table 2 which presents the full permutation distribution for the four pairs in our example under the sharp null hypothesis. Each row in the table represents a different possible set of coin flip outcomes, with the first row corresponding to the observed data with $s = 2$. We assume that the coin is fair, so that $Pr(T_s = 1) = 1/2$ for each pair $s = 1, ..., N_1$, and because there are 4 pairs, the probability of observing any individual row of

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 1	Pair 2	Pair 3	Pair 4	
					$s_{11} = 1$	$s_{21} = 0$	$s_{31} = 1$	$s_{41} = 0$	
					$s_{12} = 0$	$s_{22} = 0$	$s_{32} = 0$	$s_{42} = 0$	
Permutation	T_1	T_2	T_3	T_4	V_1	V_2	V_3	V_4	V
1	1	1	1	1	1	0	1	0	2
2	1	1	1	0	1	0	1	0	2
3	1	1	0	1	1	0	0	0	1
4	1	1	0	0	1	0	0	0	1
5	1	0	1	1	1	0	1	0	2
6	1	0	1	0	1	0	1	0	2
7	1	0	0	1	1	0	0	0	1
8	1	0	0	0	1	0	0	0	1
9	0	1	1	1	0	0	1	0	1
10	0	1	1	0	0	0	1	0	1
11	0	1	0	1	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0
13	0	0	1	1	0	0	1	0	1
14	0	0	1	0	0	0	1	0	1
15	0	0	0	1	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0

Table 2: *Permutation Distribution for McNemar's Test Using the NELDA Data*

Table 5 is 1/16. Because only rows 1, 2, 5, and 6 of the table have $V \geq 2$, the one-sided p-value is 4/16 as noted in the previous section.

The example demonstrates one of the properties of McNemar's test: concordant pairs, which are pairs with the same value of the outcome variable, can be removed from the dataset without affecting the p-value. In our example, the second pair (Kenya–Côte d'Ivoire) and fourth pair (Tanzania–Guinea-Bissau) do not affect the V statistic in Table 2, since $V_2 = 0$ and $V_4 = 0$ for all permutations. This reveals the limited power of using only a binary coding of the outcome variable and a point for potential improvement with the introduction of qualitative information.

For concepts such as opposition harassment, a binary variable coded as zero does not necessarily indicate the complete absence of that phenomenon. There was certainly some opposition harassment in all of the countries coded with $Y = 0$ in Table 1, and not all countries with $Y = 0$ had the same level of opposition harassment. Similarly, two countries coded as $Y = 1$ may not have experienced similar levels of harassment. By examining the cases in each concordant pair, we may be able to provide enough qualitative information to determine a non-zero sign on the pair. This may not be possible for some pairs, in which case, the pair remains coded as a tie. However, reducing the number of concordant pairs through these limited comparative studies will improve the power of the test.

Consider the concordant pair with $Y = 0$ in our example (Tanzania–Guinea-Bissau). Tanzania and Guinea-Bissau were both coded as $Y = 0$ with a binary variable from the NELDA dataset. However, closer investigation shows that both had some opposition harassment at a level that often appears in accounts of transitional elections, though less than other countries that were coded $Y = 1$. In Tanzania, several people were killed in fighting between the ruling party and opposition parties, and two newspaper editors were detained on sedition charges after publishing letters critical of the government (U.S. Department of State 1995). Although the opposition could generally hold large public rallies without harassment on the mainland (Commonwealth Observer Group, 1995, 16), the ruling party "intimidated and harassed the opposition and did not allow opposition rallies until 2 months prior to elections" on the island of Zanzibar (U.S. Department of State 1995). Election observers also noted reports of harassment and the occasional detention of local opposition supporters, but these were generally fairly minor incidents (Commonwealth Observer Group, 1995, 15, AWEPA, 1996, 14, U.S. Department of State, 1995). In Guinea-Bissau, the incumbent initially resisted the formation of opposition parties by "delaying registration procedures and by police violence" (Rudebeck 2002, 116). Human rights reports note that in February 1992, "five members of an opposition party were beaten and then refused hospital treatment." In addition, "police and security forces harassed opposition forces with detentions and physical mistreatment" (U.S. Department of State, 1992, 116). Because of the situation in Zanzibar, we assess Tanzania as having more opposition harassment than Guinea-Bissau, and we code $s_{41} = 1$ for this pair.

Similarly, consider the concordant pair with $Y = 1$ in our example (Kenya–Côte d'Ivoire). Some additional information indicates a difference in the severity of opposition harassment, a difference greater than that between Tanzania and Guinea-Bissau which were coded $Y = 0$. The run-up to the 1992 transitional presidential elections in Kenya were marked by "widespread intimidation, kidnapping, robbing and bribing of opposition candidates" (Tordoff 1992, 58), and widespread problems of voters not appearing on the voters register (*Kenya: The December 29, 1992 Election* 1993, 45). In addition, at least 50,000 people were internally displaced and hundreds killed in violence targeted at ethnic groups that were seen to be supportive of the opposition and making claims to land (Holmquist & Ford 1992, 103). In the run-up to the 1990 elections in Côte d'Ivoire, the ruling *Parti Démocratique de la Côte d'Ivoire* (PDCI) also harassed the opposition organized around Laurent Gbagbo of the *Front Populaire Ivoirien* (FPI), but to a much lesser extent than in Kenya. The PDCI pressured opposition newspapers and journalists, but several opposition newspapers were in circulation (U.S. Department of State 1990, 96). While the opposition were able to hold many peaceful pro-democracy demonstrations and opposition meetings, at times the opposition leader Laurent Gbagbo was prevented from making speeches at rallies (Widner 1991, 39). The police also broke up several political rallies with truncheons and tear gas, resulting in several dozen injuries (*Africa Research Bulletin* (*ARB*), Aug 1990 27:7, 9768; *ARB*, Sept 1990, 27:8, 9799-800 and *ARB*, Sept 1990 27:9, 9826-27).

Notice how these changes in coding affect the test. The V_4 column of Table 2 would now take the value 1 for the odd rows of the table but keep the value 0 for the even rows. The V_2 column of Table 2 would now take the value 1 for rows 1–4 and rows 9–12, but keep the value 0 for the other rows. Our observed data in the first row would now produce the value of $V = 4$ while all other rows will still have $V \leq 3$. Hence the V for the observed data is larger than the V values for all other rows, and we have reduced the one-sided p-value to $1/16 = 0.0625$. These changes are reflected in Table 3. Notice that for a binary outcome variable, we only need to conduct comparative studies for the concordant pairs to improve the power of the test. In this example, we only needed to conduct two comparative case studies and determine which unit had the larger outcome value within each pair.

3.2 Incorporating Between-Pair Information

While including qualitative information about within-pair differences allowed us to improve the power of the sign statistic for binary outcome data, adding between-pair ranking information allows us to use the generally more powerful signed-rank statistic. Between-pair ranking information also

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 1	Pair 2	Pair 3	Pair 4	
					$s_{11}=1$	$s_{21}=1$	$s_{31}=1$	$s_{41}=1$	
					$s_{12}=0$	$s_{22}=0$	$s_{32}=0$	$s_{42}=0$	
Permutation	T_1	T_2	T_3	T_4	V_1	V_2	V_3	V_4	V
1	1	1	1	1	1	1	1	1	4
2	1	1	1	0	1	1	1	0	3
3	1	1	0	1	1	1	0	1	3
4	1	1	0	0	1	1	0	0	2
5	1	0	1	1	1	0	1	1	3
6	1	0	1	0	1	0	1	0	2
7	1	0	0	1	1	0	0	1	2
8	1	0	0	0	1	0	0	0	1
9	0	1	1	1	0	1	1	1	3
10	0	1	1	0	0	1	1	0	2
11	0	1	0	1	0	1	0	1	2
12	0	1	0	0	0	1	0	0	1
13	0	0	1	1	0	0	1	1	2
14	0	0	1	0	0	0	1	0	1
15	0	0	0	1	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0

Table 3: *Permutation Distribution for the Sign Test Using Within-Pair Qualitative Information to Supplement the NELDA Data*

enables us to construct qualitative confidence intervals, as we discuss in the next subsection.

The signed-rank statistic uses the sign of the difference in outcomes for each pair ($sign(Y_{s1}-Y_{s2})$ for $s = 1, ..., N_1$) and the ranks of the absolute values of the within-pair differences in the outcomes ($rank(abs(Y_{s1} - Y_{s2}))$ for $s = 1, ..., N_1$). The pair with the largest absolute difference in outcomes is assigned a rank of N_1, the pair with the smallest absolute difference in outcomes is assigned a rank of 1, and tied pairs are assigned an average of the ranks of those pairs. The statistic is:

$$W = \sum_{s=1}^{N_1} q_s[T_s s_{s1} + (1 - T_s)s_{s2}] \ = \sum_{s=1}^{N_1} W_s$$

where s_{s1} and s_{s2} as defined as before, and q_s is the rank for each pair.

For our running example, we must delve into the details of eight cases, but only as deeply as necessary to rank the absolute differences in outcomes in each pair. It is not necessary to determine exactly how much bigger or smaller the difference is in one pair than in another, only whether the absolute difference is bigger or smaller. Moreover, debates over measurement of complex outcome variables need only be settled to the extent that they produce agreement on the signs and ranks. In other words, scholars may disagree on how much the numbers of deaths and the extent of violence each contribute to the overall assessment of the severity of opposition harassment, as long as they agree enough to produce the same signs and rankings of the absolute differences. Finally, as we discuss below, it is straightforward to perform a sensitivity analysis should there be disagreement over the signs or ranks.

By looking more closely at these cases, we determine that the Kenya–Côte d'Ivoire pair has the largest rank, followed by Cameroon–Gabon, Malawi–Zambia, and finally the Tanzania-Guinea-Bissau pair.[4] Descriptions of these pairs and more details on our rankings are in Appendix C. In each pair, the treated country had more opposition harassment than its paired control country ($Y_{s1} > Y_{s2}$) so that $s_{s1} = 1$ and $s_{s2} = 0$ for these pairs. We present the proposed ranks for our example in Table 4.

Pair s	Treated (Plurality)	Control (Runoff)	q_s	s_{s1}	s_{s2}	observed W_s	alternate W_s
1	Cameroon	Gabon	3	1	0	3	0
2	Kenya	Côte d'Ivoire	4	1	0	4	0
3	Malawi	Zambia	2	1	0	2	0
4	Tanzania	Guinea-Bissau	1	1	0	1	0

Table 4: *Using Qualitative Information to Rank Differences in Outcomes within Matched Pairs.*

We started with binary data constructed from the NELDA dataset for our example, but ordinal data may be available for other analyses from other datasets, which may simplify the task of coding the signs and ranks in two ways. First, when the outcome variable can take on many values, there will likely be fewer concordant matched pairs, and second, the ordinal data will provide partial ranks of the pairs. For example, when two pairs have the same sign, and the control units of the two pairs have the same value of the outcome but the treated units have different values of the outcome, then it is possible to rank the two pairs.

[4]It may initially seem strange that the largest difference is between countries that were both coded as $Y = 1$, but this merely indicates the extreme severity of opposition harassment in Kenya.

Note that most of the leverage comes from these signs and ranks, rather than any additional information beyond these signs and ranks. Moreover, even if interval measures of the outcome were available, we might still use the signed-rank statistic because it provides robust power with non-normal outcomes (Rosenbaum 2002, 2009). Therefore, if ordinal or interval coding of the outcome variable is not available, researchers should focus their efforts on producing the signs and ranks rather than the full ordinal coding.

Table 5 presents the permutation distribution for the signed-rank test for the four pairs under the sharp null hypothesis. Note that again the first row corresponds to the observed data, with $W = 10$. No other value of W within the table is as large as the observed $W = 10$, and hence again the one-sided p-value is $1/16$. The signed-rank statistic is generally more powerful than the sign statistic (Rosenbaum 2002, 2009), but the p-value from the signed-rank statistic is not smaller than the p-value generated by the sign statistic for our particular example. However, as we demonstrate in the next section, this rank information can be used to produce a qualitative confidence interval.

3.3 Qualitative Confidence Intervals

The p-value provides information about the likelihood that the observed evidence could be due to chance. But having rejected the sharp null hypothesis, we would like to have a more descriptive representation of plausible sizes for the effect. The most familiar such description in quantitative analyses is a confidence interval, and if we had a continuous measure of the outcome variable Y, we could form confidence intervals within the framework of randomization inference in the manner suggested in Rosenbaum (2002, 2009). Specifically, we would consider different possible null hypotheses and form a confidence interval on the basis of the null hypotheses that we fail to reject. We describe below the procedure for producing such confidence intervals if Y could be measured as a continuous variable, and then discuss how we form *qualitative* confidence intervals with just the signs and ranks.

If Y can be measured, the first step is to alter the null hypothesis by assuming an effect size for each unit in the study. This can be done in many ways, but one simple approach is to assume

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 1	Pair 2	Pair 3	Pair 4	
					$q_1 = 3$	$q_2 = 4$	$q_3 = 2$	$q_4 = 1$	
					$s_{11} = 1$	$s_{21} = 1$	$s_{31} = 1$	$s_{41} = 1$	
					$s_{12} = 0$	$s_{22} = 0$	$s_{32} = 0$	$s_{42} = 0$	
Permutation	T_1	T_2	T_3	T_4	W_1	W_2	W_3	W_4	W
1	1	1	1	1	3	4	2	1	10
2	1	1	1	0	3	4	2	0	9
3	1	1	0	1	3	4	0	1	8
4	1	1	0	0	3	4	0	0	7
5	1	0	1	1	3	0	2	1	6
6	1	0	1	0	3	0	2	0	5
7	1	0	0	1	3	0	0	1	4
8	1	0	0	0	3	0	0	0	3
9	0	1	1	1	0	4	2	1	7
10	0	1	1	0	0	4	2	0	6
11	0	1	0	1	0	4	0	1	5
12	0	1	0	0	0	4	0	0	4
13	0	0	1	1	0	0	2	1	3
14	0	0	1	0	0	0	2	0	2
15	0	0	0	1	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0

Table 5: *Permutation Distribution for the Signed-Rank Statistic Using Within- and Between-Pair Qualitative Information to Supplement the NELDA Data*

that the effect takes a constant value c for all units:[5]

$$H_0 : \tau_{s1} = \tau_{s2} = c, \text{ for } s = 1, ..., N_1$$

We are interested in positive effects for our example, so we start by considering small positive values of c. We can test the adjusted null hypothesis for a fixed value of c at the 10% level by subtracting c from Y for the treated units, re-calculating the p-value, and rejecting the altered null if the p-value is less than or equal to 10%. For our example, this means adjusting Y for the plurality countries such that, for example $Y^*_{Cameroon} = Y_{Cameroon} - c$. The p-value could be calculated as described in the previous section with Y^* used for the plurality countries and Y used for the runoff countries. This process could then be repeated until we found the smallest value of c such that the p-value

[5] For example, it is sufficient to assume that the effect for the treated unit in the lowest ranked pair is no larger than the effects for the other units in order to test against the following hypothesis $H_0 : \tau_{s1} \leq c, \tau_{s2} \leq c, \text{ for } s = 1, ..., N_1$.

is greater than 10%. This value of c, which we denote $c_{10\%}$, represents the lower bound of the one-sided 90% confidence interval.

Notice that although we do not have a continuous measure for Y, the signs and ranks provide enough information to describe this value of $c_{10\%}$. For our electoral rule example, reversing the sign on any of the pairs generates a p-value greater than 10%. Because the Tanzania–Guinea-Bissau pair has the smallest absolute difference in outcomes, $c_{10\%}$ is the c that reverses the sign of this pair. The lower bound of the 90% one-sided confidence interval, $c_{10\%}$, is therefore the difference in the severity of opposition harassment between Tanzania and Guinea-Bissau, $Y_{Tanzania} - Y_{Guinea-Bissau}$.

We cannot provide a quantitative description of the difference in harassment intensity between Tanzania and Guinea Bissau since quantitative measurements of $Y_{Tanzania}$ and $Y_{Guinea-Bissau}$ are unavailable. However, we have a qualitative description of this difference, provided above and as summarized in Table 6. We see that the major difference in opposition harassment between the two countries was that Tanzania banned opposition rallies on Zanzibar, while the opposition in Guinea-Bissau did not face such restrictions.

Opposition Harassment (Y)		
Tanzania	Guinea-Bissau	90% One-Sided Confidence Interval
detentions	detentions	$[Y_{\text{Tanzania}} - Y_{\text{Guinea-Bissau}}, \infty)$
physical violence	physical violence	
barred rallies in opposition stronghold		

Table 6: *90% One-Sided Qualitative Confidence Interval.* The difference in opposition harassment between Tanzania and Guinea-Bissau represents the lower bound on the 90% one-sided confidence interval for $\tau = \tau_{s1} = \tau_{s2}$ for all $s = 1, ..., N_1$.

Both the p-value for the sharp null of no effect and the qualitative confidence interval are calculated using the signs and ranks presented in Table 4. Our remaining concern is whether our analysis using the signed-rank statistic is sensitive to errors or disagreements about these signs and ranks.

3.4 Sensitivity Analysis on the Signs and Ranks

We can address uncertainty or disagreements about the signs and ranks with a sensitivity analysis that calculates p-values for all possible signs and ranks. First, note that some disagreements over ranks will have no effect on the p-value. For example, perhaps we cannot determine or agree on whether the Cameroon-Gabon pair or the Kenya-Côte d'Ivoire pair has a greater absolute difference in outcomes. Suppose contrary to our ranking in Table 4, the Cameroon-Gabon pair is rank 4 and the Kenya-Côte d'Ivoire pair is rank 3. The p-value we obtain from this adjusted ranking is the same as from our original ranking. More generally, when $T_s s_{s1} + (1 - T_s) s_{s2} = T_{s'} s_{s'1} + (1 - T_{s'}) s_{s'2}$ for $s \neq s'$ and $q_s = q_{s'} + 1$, then exchanging the ranks for pairs s and s' does not affect the p-value.

Furthermore, even errors in the signs are likely to have only small effects on the p-value, since such errors are most likely to occur for the pairs where the magnitude of the within-pair difference and the rank are small. For example, we might be concerned that we have missed some incidents of opposition harassment in Guinea-Bissau that were only reported in Portuguese language media. However, even if upon reassessment we find that Guinea-Bissau had more harassment than Tanzania, it is unlikely that the magnitude of this difference will be large, and therefore the rank will remain the same $q_4 = 1$. Because of the small rank, the sign reversal only generates a small difference in the p-value. Note that in this alternative, $s_{41} = 0$ and $s_{42} = 1$ so that we subtract one from W in the odd rows and add one to W in the even rows of Table 5. Our observed W would now be 9, and the second row would have $W = 10$, but no other row in the table would have a W as large as 9. Therefore, the p-value would only increase from 1/16 to 2/16. Contrast this with a coding change for the sign test. If the outcome in Guinea-Bissau were recoded as larger than that in Tanzania in Table 3, then the p-value would increase from 1/16 to 5/16. This demonstrates the additional power in the signed-rank test.

The qualitative confidence interval is more sensitive to disagreements about ranks in the sense that the pair of countries that describes the bound may change. For example, if we switch the rank of the Malawi-Zambia pair with the Tanzania–Guinea-Bissau pair, the p-value would not change, but the lower bound on the one-sided confidence interval would switch to the difference in severity of opposition harassment between Malawi and Zambia from that between Tanzania and Guinea-Bissau.

4 Using Qualitative Information on Confounders

The analysis in the previous section was predicated on hypothetical coin flips within comparable pairs of units. Even though the thought experiment may produce a small p-value, we may be concerned that the units were somewhat incomparable on a confounder so that the result reflected differences in this variable rather than the effect of the treatment. For this situation, the thought experiment should have employed a weighted coin, since one unit in a pair was *ex ante* more likely to have received treatment than the other.

Formally, two units in a pair will be comparable when they have the same values of X (our matching variables) and U (some unmeasured matching variables) such that learning which of the two units received treatment will not provide information about the potential outcomes for either unit:

$$Y(1), Y(0) \perp\!\!\!\perp T | X, U$$

Suppose that for two units in a matched pair where one unit received treatment and the other received control, we can write the probability that the first unit received treatment as the following:

$$\pi_s = Pr(T_s = 1 | X_{s1} = x_1, X_{s2} = x_2, U_{s1} = u_1, U_{s2} = u_2)$$

If $x_1 = x_2$ and $u_1 = u_2$, then $\pi_s = .5$. However, Rosenbaum (2002) shows that when $x_1 \neq x_2$ or $u_1 \neq u_2$ or both, we can bound p-values by assuming that the first unit in all pairs was at most $\Gamma \geq 1$ times and at least $\frac{1}{\Gamma}$ times as likely to have received the treatment as the second unit,

$$\frac{1}{\Gamma} \leq \frac{\pi_s}{1 - \pi_s} \leq \Gamma \text{ for all } s = 1, ..., N_1.$$

Therefore, we can perform a sensitivity analysis on the p-value by setting Γ to increasingly larger values. Specifically, we can use the value of Γ to determine the upper and lower bounds on π_s. We then use these bounds on π_s to generate the possible probabilities for each of the 16 rows of Table 5 and to generate upper bounds on the p-values. Table 7 presents this sensitivity analysis on the upper bound for the p-value for our electoral rules example, and we see that the result is quite sensitive to unmeasured confounding. If the treated units were twice as likely to receive treatment as the control units due to unmeasured confounding, then we would have obtained a p-value as large as 32%. This is quite a bit larger than the p-value of 6% we obtained by assuming a fair coin flip within each pair.

Γ	p-value upper bound
1	0.063
1.5	0.130
2	0.198
2.5	0.260
3	0.316

Table 7: *Sensitivity Analysis without Qualitative Information*

4.1 Using Qualitative Information on Measured Confounders

Note that the analysis in Table 7 is somewhat conservative when we are only interested in detecting effects in one direction. We are only concerned with positive bias in our running example, since we are only investigating the possibility of positive effects of plurality electoral rule on opposition harassment. This implies that we are only worried about mismatches that will produce positive bias, and hence qualitative information on the direction of bias can reduce the sensitivity of the analysis.

To fix ideas, consider an observational matching study to determine the effects of a drug on health outcomes. Suppose the study shows that the drug is beneficial for health, but that subjects are mismatched on age so we are concerned about bias. However, suppose that all the subjects who received the drug were older than the control subjects matched to them. In this case, we might be willing to assume that mismatches on age could only be negatively biasing the study, and therefore, that we need not perform a sensitivity analysis on the p-value. Intuitively, we can sign the bias if the sign of the effects of the confounder on the treatment and on the outcome are known. If both effects have the same sign, the bias will be positive. If the effects have different signs, the bias will be negative. This can be formalized within a two parameter amplification of the Rosenbaum-style sensitivity analysis (Rosenbaum & Silber 2009), but for our purposes, it will be sufficient to adjust our sensitivity analysis on a pair-by-pair basis.

Within our electoral rules analysis, we might be concerned about the imbalance on former colonial power in two of our matches. In one pair, the plurality country is a former British colony (Kenya), while the runoff country is a former French colony (Côte d'Ivoire). In another pair, the plurality country is a former British colony (Tanzania), while the runoff country is a former Portuguese colony (Guinea-Bissau). However, we worry differently about the bias induced by these mismatches depending on how we think the legacies of rule by different colonial powers affect the likelihood of opposition harassment. More specifically, we might believe that former Portuguese colonies will generally have more severe opposition harassment than former British colonies under the control condition because incumbents in the former are more accustomed to using force for political ends, including winning independence through violent insurgencies against more autocratic powers more recently. Then we may not worry that the mismatch for the Tanzania–Guinea-Bissau pair will add positive bias. In contrast, if we are not as sure about the effects of being a former French colony versus being a former British colony on the outcome, then we might still consider the possibility of positive bias for the Kenya–Côte d'Ivoire pair. Therefore, it may be possible to reduce the sensitivity of our analysis if we use different values of Γ for each pair (i.e., Γ_s for $s = 1, ..., N_1$).

If for simplicity we focus on colonial background, and we believe that the only pair that could produce positive bias is the Kenya–Côte d'Ivoire pair, then we can set $\Gamma_s = 1$ for $s = 1, 3, 4$, while allowing $\Gamma_2 > 1$. In this sensitivity analysis presented in Table 8, notice that the p-value never rises above 10%.

However, there are two problems with this analysis. First, the choice of cutoff for Γ_2 is still arbitrary. One could easily argue that former British colonies were *ex ante* more than three times as likely as former French colonies to use plurality election rules.[6] The second problem is the assumption that we have measured all of the confounding variables. We discuss this second issue in the next subsection.

Γ_2	p-value upper bound
1	0.063
1.5	0.075
2	0.083
2.5	0.089
3	0.094

Table 8: *Sensitivity Analysis with Qualitative Information on Measured Mismatches*

4.2 Using Qualitative Information on Unmeasured Confounders

Even when treated and control units are perfectly matched on the measured confounders such that $x_1 = x_2$, they may be mismatched on unmeasured confounders such that $u_1 \neq u_2$. Although we cannot measure confounders explicitly, case studies may provide some qualitative information about the $u_1 - u_2$ difference. Specifically, if we believe we know the key unmeasured confounder U^*, and if we can learn the sign of $u_1^* - u_2^*$ for some pairs, and if as above we know the sign of the effects of the confounder on the treatment and the outcome, then we can provide sharper bounds on the p-value.

In our electoral rules example, we may be concerned that the matching variables discussed in Appendix B do not fully capture the strength of opposition, the key variable affecting the outcome that we believe makes an incumbent authoritarian regime more likely to adopt plurality rules. If this is true, then the probability of treatment is not equal across the countries in each pair. Further comparative case studies allow us to assess the relative strength of opposition, and hence give a sense of whether treatment was more likely for one case than the other in each pair.

[6] In fact, if we examine all sub-Saharan African transitional presidential elections during this period, we find that 3 of 8 transitional presidential elections in former British colonies used plurality rules, while only 1 in 12 transitional presidential elections in former French colonies used plurality rules. This would imply that $\Gamma_2 = \frac{3/8}{1/12} = 4.5$.

These comparative case studies on the strength of opposition are presented in Appendix D. Although we could not make a clear determination for the Kenya-Côte d'Ivoire pair or the Tanzania–Guinea-Bissau pair, we believe the opposition was marginally stronger in Cameroon than in Gabon and stronger in Zambia than in Malawi. This means the strength of opposition was stronger for the treated unit in the Cameroon–Gabon pair and stronger for the control unit in the Malawi–Zambia pair. If we believed that strength of opposition fully represented the U in the Rosenbaum (2002) model, then we might conduct a sensitivity analysis that utilizes different Γ_s for each pair.

	Γ_4			
$\Gamma_{1,2}$	1	1.25	1.5	$\Gamma_{1,2}$
1	-	-	-	0.063
1.5	0.090	0.100	-	0.108
2	0.111	0.123	0.133	0.148
2.5	0.128	0.142	0.153	0.182
3	0.141	0.156	0.169	0.211

Table 9: *Sensitivity Analysis with Qualitative Information on the Unmeasured Confounder. This analysis assumes that* $\Gamma_3 = 1$, $\Gamma_4 \geq 1$, $\Gamma_1 \geq \Gamma_4 \geq 1$ *and* $\Gamma_2 \geq \Gamma_4 \geq 1$. *To simplify the presentation, we assume that* $\Gamma_1 = \Gamma_2$, *which we write as* $\Gamma_{1,2}$.

Because opposition strength was stronger for the control unit in the Malawi–Zambia pair ($s = 3$), we will hold $\Gamma_3 = 1$. We assume that the Tanzania–Guinea-Bissau pair ($s = 4$) has $\Gamma_4 \geq 1$ in order to reflect our uncertainty about opposition strength. Finally, because opposition strength was stronger for the treated unit in the Cameroon–Gabon pair ($s = 1$), and because we have the British–French mismatch on the Kenya–Côte d'Ivoire pair ($s = 2$) along with our uncertainty about the opposition strength for this pair, we allow $\Gamma_1 \geq \Gamma_4 \geq 1$ and $\Gamma_2 \geq \Gamma_4 \geq 1$. To simplify the presentation, we assume that $\Gamma_1 = \Gamma_2$, which we write as $\Gamma_{1,2}$. This is presented as a two-dimensional sensitivity analysis in Table 9. First, we set $\Gamma_{1,2}$ equal to the values from Table 7. Then we investigate increasing values of Γ_4. The columns represent different values of Γ_4. The far left column sets these ratios equal to 1, the far right column sets these upper bounds at $\Gamma_{1,2}$. The middle columns explore other possibilities in between these two extremes. Note that the use of qualitative information on confounders reduces the sensitivity of the analysis when Table 9 is compared to Table 7 – when $\Gamma_{1,2} = \Gamma_4 = 1.5$ the p-value is only marginally larger than 10%.

5 Using Qualitative Information to Improve Full Matching

Our discussion has so far focused on pair matching, but we may benefit from having a variable number of treated and control units in each matched set through full matching (Hansen 2004). We first show that, as discussed in Hansen (2004), we can reduce mismatches on the measured confounders by allowing more general matched sets. We then demonstrate how full matching allows us to include additional units to reduce sensitivity to unmeasured confounders and to lower p-values.

5.1 Using Qualitative Information on the Outcome in Full Matching

With full matching, our 8 countries are now matched in 3 sets instead of 4 pairs (Table 10). Notice that full matching allows us to match all former French colonies to other former French colonies, so that we no longer have a former French colony as a control for a former British colony. Because we believe the bias would be negative, we continue to allow a former British colony (Tanzania) to be matched to a former Portuguese colony (Guinea-Bissau). Appendix E presents the details of the full matching procedure for our running example.

Set s	Treated (Plurality)	Control (Runoff)	q_s	r_{s1}	r_{s2}	r_{s3}	observed Q_s
1	Cameroon	Gabon, Côte d'Ivoire	2	3	2	1	6
2	Kenya, Malawi	Zambia	3	3	2	1	15
3	Tanzania	Guinea-Bissau	1	2	1	NA	2

Table 10: *Using Qualitative Information for Full Matching.*

While full matching reduces imbalance on the measured confounders (see Appendix E), it also rules out the use of the signed-rank statistic. Fortunately, Quade's statistic (Quade, 1979, Rosenbaum, 2002, 161) provides a straightforward generalization of the signed-rank statistic that can be used with full matching. Quade's statistic uses both within-set and between-set ranks. With pair matching, there were $n_s = 2$ units within each set s, but now we allow arbitrary n_s units in each set. These units are assigned ranks from 1 to n_s according to the size of the outcomes Y_{sj} for $j = 1, ..., n_s$. With $S \leq N_1$ sets, we write these ranks as r_{sj} for $j = 1, ..., n_s$ and $s = 1, ..., S$.

These within-set ranks are presented for our electoral rule example in Table 10. Notice the r_{33} is not defined because there are only two countries in the $s = 3$ set.

As before, the S sets are assigned ranks from 1 to S, which we write as q_s for $s = 1, ..., S$. However, because n_s can now be larger than 2, the ranks q_s are determined by the absolute values of the differences between the maximum and minimum outcomes in the group. ($rank(abs(max_j\{Y_{sj}\} - min_j\{Y_{sj}\}))$ for $s = 1, ..., S$). This means that for our running example the ranks are determined by $abs(Y_{Cameroon} - Y_{Cote\ d'Ivoire})$, $abs(Y_{Kenya} - Y_{Zambia})$, and $abs(Y_{Tanzania} - Y_{Guinea-Bissau})$. Finally, because we allow more than one treated and/or control unit within each group, we define T_{sj} to be a treatment indicator for the jth unit in set s, such that $T_{sj} = 1$ if that unit receives treatment and $T_{sj} = 0$ if not. With these definitions the Quade statistic can be written as the following:

$$Q = \sum_{s=1}^{S} q_s \sum_{j=1}^{n_s} T_{sj} r_{sj} = \sum_{s=1}^{S} Q_s.$$

where $Q_s = \sum_{j=1}^{n_s} T_{sj} r_{sj}$.

If we define m_s to be the number of treated units in set s ($\sum_{j=1}^{n_s} T_{sj} = m_s$), then conditional on $\{q_s, n_s, m_s, r_{sj}\}$ for $s = 1, ..., S$, the permutation distribution for Quade's statistic can be derived in a manner analogous to the permutation distribution for the signed-rank statistic. Table 11 presents this distribution for the three sets in our example under the sharp null hypothesis. Note that again the first row corresponds to the observed data, with $Q = 23$. No other value of Q within the table is as large as the observed $Q = 23$, but now the table has 18 rows ($3 \times 3 \times 2$), so the one-sided p-value is 1/18.

5.2 Using Qualitative Information on Unmeasured Confounders in Full Matching

Full matching can also relieve some of the mismatches on opposition strength, our key unmeasured matching variable. First, the potential for mismatch on opposition strength for Kenya is reduced, because it is now in the same set as Zambia (control), which we judge to have greater opposition strength than both Kenya and Malawi (both treated). Second, we assessed that Cameroon

	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3	
				$q_1 = 2$	$q_2 = 3$	$q_3 = 1$	
				$r_{11} = 3$	$r_{21} = 3$	$r_{31} = 2$	
				$r_{12} = 2$	$r_{22} = 2$	$r_{32} = 1$	
				$r_{13} = 1$	$r_{23} = 1$		
Permutation	T_{11}, T_{12}, T_{13}	T_{21}, T_{22}, T_{23}	T_{31}, T_{32}	Q_1	Q_2	Q_3	Q
1	1,0,0	1,1,0	1,0	6	15	2	23
2	1,0,0	1,1,0	0,1	6	15	1	22
3	1,0,0	1,0,1	1,0	6	12	2	20
4	1,0,0	1,0,1	0,1	6	12	1	19
5	1,0,0	0,1,1	1,0	6	9	2	17
6	1,0,0	0,1,1	0,1	6	9	1	16
7	0,1,0	1,1,0	1,0	4	15	2	21
8	0,1,0	1,1,0	0,1	4	15	1	20
9	0,1,0	1,0,1	1,0	4	12	2	18
10	0,1,0	1,0,1	0,1	4	12	1	17
11	0,1,0	0,1,1	1,0	4	9	2	15
12	0,1,0	0,1,1	0,1	4	9	1	14
13	0,0,1	1,1,0	1,0	2	15	2	19
14	0,0,1	1,1,0	0,1	2	15	1	18
15	0,0,1	1,0,1	1,0	2	12	2	16
16	0,0,1	1,0,1	0,1	2	12	1	15
17	0,0,1	0,1,1	1,0	2	9	2	13
18	0,0,1	0,1,1	0,1	2	9	1	12

Table 11: *Permutation Distribution for Quade's Statistic Using Within- and Between-Set Qualitative Information to Supplement the NELDA Data*

(treated unit) had stronger opposition than Gabon (control), and further that Gabon had stronger opposition than Côte d'Ivoire (Appendix E). Full matching allows us to partially relieve this mismatch on opposition strength within the set of former French colonies by including an additional control unit in the analysis, and consequently lower the p-value.

To formalize the sensitivity analysis with full matching, it is most straightforward to define the parameter π_{sj} for each unit in each set as the *ex ante* probability of each unit receiving treatment, conditional on the number of treated and control units in the set. If we define $\mathbf{X_s}$ and $\mathbf{U_s}$ to be the matrices of measured and unmeasured confounders for all units within set s, then this *ex ante* probability can be written as the following:

$$\pi_{sj} = Pr(T_{sj} = 1 | \sum_{j=1}^{n_s} T_{sj} = m_s, \mathbf{X_s}, \mathbf{U_s})$$

Because the observed data provides the maximum value of Q=23 (row 1 of Table 11), our sensitivity analysis need only focus on that row of the table. For the former French colonies set ($s = 1$), the concern is that because Cameroon was judged to have greater opposition strength than Gabon or Côte d'Ivoire, we believe that $1 \geq \pi_{11} \geq \pi_{12} \geq \pi_{13} \geq 0$, or in other words, that Cameroon was more likely to have received treatment than Gabon or Côte d'Ivoire. Therefore $\pi_{11} \geq 1/3$, which implies that the p-value will be greater than if the treatment were effectively randomly assigned within this set.

However, it is possible to limit the size of π_{11} by including an additional control unit in this set. If we can find another potential former French colony as a control unit that matches well on the measured variables and has opposition strength greater than Cameroon, then this unit would increase the size of the set ($n_1 = 4$), and it would change the definition of π_{1j} for all j. In particular, if opposition strength dictates the values of π_{1j}, then we would believe that $1 \geq \pi_{14} \geq \pi_{11} \geq \pi_{12} \geq \pi_{13} \geq 0$ and therefore $\pi_{11} \leq 1/2$. Note this important change on the restriction for π_{11}. With the previous set of three countries, π_{11} must be at least 1/3 and could be as large as 1. With the new set of four countries, π_{11} can at most be 1/2 and could be as small as zero. Note also that with the new set, π_{11} can equal its randomization probability (i.e., $m_s/n_s = 1/4$) for a variety of different values of π_{11}, π_{12}, and π_{14} and in particular for values of $\pi_{13} \leq \pi_{12} < 1/4$. In the case where the unmeasured confounder is some amalgamation of pre-treatment values of the outcome, this result provides some of the benefits of synthetic matching (Abadie, Diamond & Hainmueller 2010).

For our electoral rules example, the inclusion of an additional former French colony as a control unit requires the relaxation of our treatment of military rule, because all other potential Francophone controls had some experience with military rule. Among these potential control countries, we select Madagascar because it only spent 3% of its years since independence under military rule and because we assess that it has stronger opposition than does Cameroon. We are unconcerned by the mismatch on past experience under military rule for two reasons. First, 3% is a very limited experience, and second, a military with experience in politics and repression of opposition should result in negative bias.[7]

Despite its strength of opposition, we find less opposition harassment in Madagascar than in all other countries in the former French colonies set (see Appendix E for details). We also find that the ranking of this set ($s = 1$) stays the same ($q_1 = 2$). Therefore when we incorporate this country into the permutation distribution (Table 12), the observed data represented in the first row still has the largest value of Quade's statistic, but there are now 24 rows in the table so the randomization p-value is 1/24. Note that because we still assess the difference in opposition harassment to be smaller for the Tanzania–Guinea-Bissau then for any other treatment and control pairing (see Appendix E for details). If the ranking of this pair were to flip, the p-value would become 2/24, and therefore the Tanzania–Guinea-Bissau difference now defines the lower bound of a one-sided 95% qualitative confidence interval.

Again, we perform a sensitivity analysis, but due to the previous discussion, we need only consider the parameters π_{11} (corresponding to Cameroon) and π_{31} (corresponding to Tanzania). This is presented in Table 13 with values of π_{11} between 1/4 and 1/2 and values of π_{31} set to correspond to the Γ values considered for the Tanzania–Guinea-Bissau pair in Table 9. In order to simplify the comparison across tables, we will write these probabilities as $\frac{\Gamma_s}{(n_s-1)+\Gamma_s}$ without reducing fractions. Recall as well that we have already discounted the effects of the mismatch in British-Portuguese colonial background for this pair, so assessment of the likely values of π_{31} should not consider this difference.[8]

[7]Moreover, all other potential Francophone control units had longer experience with military rule, and even by relaxing the cutoff up to 30%, we would only reduce the p-values presented in the paper.

[8]Within the two parameter amplification of the sensitivity analysis (Rosenbaum & Silber 2009), this can be formalized for the Tanzania–Guinea-Bissau pair in two steps. First, we can combine in the parameter λ the positive effects of the mismatch in British-Portuguese colonial background on Tanzania receiving the treatment with the potentially positive effects of an opposition strength mismatch on Tanzania receiving the treatment. Second, we can combine in the parameter δ the negative effects of the mismatch in British-Portuguese colonial background on the outcome difference under the control condition and the potentially positive effects of an opposition strength mismatch on the outcome difference under the control condition. Then we can write $\pi_{31} = \frac{exp(\lambda+\delta)+1}{(1+exp(\delta))(1+exp(\lambda))}$. Intuitively, we can make λ relatively large and δ relatively small to incorporate our qualitative knowledge about this mismatch on colonial background.

	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3	
				$q_1 = 2$	$q_2 = 3$	$q_3 = 1$	
				$r_{11} = 4$	$r_{21} = 3$	$r_{31} = 2$	
				$r_{12} = 3$	$r_{22} = 2$	$r_{32} = 1$	
				$r_{13} = 2$	$r_{23} = 1$		
				$r_{14} = 1$			
Permutation	$T_{11}, T_{12}, T_{13}, T_{14}$	T_{21}, T_{22}, T_{23}	T_{31}, T_{32}	Q_1	Q_2	Q_3	Q
1	1,0,0,0	1,1,0	1,0	8	15	2	25
2	1,0,0,0	1,1,0	0,1	8	15	1	24
3	1,0,0,0	1,0,1	1,0	8	12	2	22
4	1,0,0,0	1,0,1	0,1	8	12	1	21
5	1,0,0,0	0,1,1	1,0	8	9	2	19
6	1,0,0,0	0,1,1	0,1	8	9	1	18
7	0,1,0,0	1,1,0	1,0	6	15	2	23
8	0,1,0,0	1,1,0	0,1	6	15	1	22
9	0,1,0,0	1,0,1	1,0	6	12	2	20
10	0,1,0,0	1,0,1	0,1	6	12	1	19
11	0,1,0,0	0,1,1	1,0	6	9	2	17
12	0,1,0,0	0,1,1	0,1	6	9	1	16
13	0,0,1,0	1,1,0	1,0	4	15	2	21
14	0,0,1,0	1,1,0	0,1	4	15	1	10
15	0,0,1,0	1,0,1	1,0	4	12	2	18
16	0,0,1,0	1,0,1	0,1	4	12	1	17
17	0,0,1,0	0,1,1	1,0	4	9	2	15
18	0,0,1,0	0,1,1	0,1	4	9	1	14
19	0,0,0,1	1,1,0	1,0	2	15	2	19
20	0,0,0,1	1,1,0	0,1	2	15	1	18
21	0,0,0,1	1,0,1	1,0	2	12	2	16
22	0,0,0,1	1,0,1	0,1	2	12	1	15
23	0,0,0,1	0,1,1	1,0	2	9	2	13
24	0,0,0,1	0,1,1	0,1	2	9	1	12

Table 12: *Permutation Distribution for Quade's Statistic Using Within- and Between-Set Qualitative Information to Supplement the NELDA Data, as well as an additional former French colony as a control unit.*

This analysis is presented in Table 13. Notice that if $\pi_{11} = 1/4$, and $\pi_{31} \leq 1.5/2.5$, then the p-value is at most 5%. Furthermore, the maximum p-value based on the upper bound for π_{11} and a value of $\pi_{31} = 3/4$ still provides a p-value of 0.125.

	π_{31}					
π_{11}	1/2	1.25/2.25	1.5/2.5	2/3	2.5/3.5	3/4
1/4	0.042	0.046	0.050	0.056	0.60	0.063
1.25/4.25	0.049	0.054	0.059	0.065	0.070	0.074
1.5/4.5	0.056	0.062	0.067	0.074	0.079	0.083
2/5	0.067	0.074	0.080	0.089	0.095	0.100
2.5/5.5	0.076	0.084	0.091	0.101	0.108	0.114
3/6	0.083	0.093	0.100	0.111	0.119	0.125

Table 13: *Sensitivity Analysis with Qualitative Information included on the unmeasured confounder within full matching. This analysis assumes that* $(1 - \pi_{23}) \leq 1/3$*.*

6 Conclusion

For many questions in political science, researchers face the challenges of poorly measured outcomes, imbalance on measured and unmeasured confounders, and small sample size after removing incomparable units from the study. Analyses of the effects of country-level institutions on large-scale social or political outcomes are particularly vulnerable to these problems, since these institutions are generally chosen endogenously through complex political processes and the population of units is limited. But, as this paper has demonstrated, the small sample size makes feasible the use of qualitative information to improve causal inferences in an observational study.

In our analysis of the effect of presidential electoral rules on opposition harassment in African countries undergoing regime transition in the 1990s, comparative case studies allowed us to rank within and between set differences in the severity of opposition harassment and rank the direction of within-set differences in unmeasured strength of opposition. The techniques described in the previous sections provided a principled way in which to use the qualitative information we learned from these case studies in order to improve our analysis. We showed that by incorporating case knowledge within the Rosenbaum (2002, 2009) approach, we can improve power and potentially reduce p-values, provide qualitative confidence intervals, and reduce sensitivity to unmeasured confounders.

By showing how and how much additional qualitative information can improve causal inference, we have offered statistically-grounded guidelines for how mixed methods researchers should direct their efforts in data collection for small- and medium-n studies. The first step should be to understand the treatment assignment process and to focus on measurement of the more important matching variables rather than improving the outcome variable. Then after matching with these variables, researchers should focus on signing and ranking differences in outcomes within concordant pairs or sets in order to, but do no more than, establish within-set and between-set rankings. Existing datasets can be very helpful starting points for both of these steps, but it is only necessary to work on the outcome variable for the cases that appear in the matched sets, not the entire sample. Our methods also point out which differences will be candidates to define the bounds of a qualitative confidence interval for some specified level, so that the researcher can focus on characterizing more precisely the difference in outcomes among those candidates. Finally, deep knowledge beyond the information encoded in quantitative datasets should be used to assess the relative probabilities of treatment assignment within these sets in order to strengthen and clarify the credibility of a study.

More generally, we have shown that even when we have a small sample size, randomization inference allows qualitative information to be incorporated in a nonparametric statistical framework. The methods we have described allow the formal synthesis of quantitative and qualitative information without the use of parametric assumptions or the elicitation of Bayesian priors. Unlike other mixed-methods approaches, our method formally integrates qualitative information with quantitative analysis. This allowed us to provide evidence suggesting that plurality rules may increase the severity of opposition harassment and to characterize the lower bound for the size of that effect. That this was possible even with only four countries with plurality rules suggests the potential benefits from an expanded study.

References

Abadie, Alberto, Alexis Diamond & Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* .

Africa Research Bulletin. N.d. Various issues.

AWEPA. 1996. *Tanzania's Multi-party Elections.* Amsterdam: African-European Institute.

Azevedo, Mario. 1995. Ethnicity and Democratization: Cameroon and Gabon. In *Ethnic Conflict and Democratization in Africa*, ed. Harvey Glickman. Atlanta, GA: African Studies Association Press pp. 255–88.

Bayalama, Sylvain. 1991. "Pluralism and Political Change in Central Africa." *Africa Today* 38(3):66–71.

Bowers, Jake & Costas Panagopoulos. 2009. ""Probability of What?": A Randomization-based Method for Hypothesis Tests and Confidence Intervals about Treatment Effects." Unpublished manuscript.

Bowers, Jake & Costas Panagopoulos. 2011. "Fisher's randomization mode of statistical inference, then and now." Unpublished manuscript.

Brady, Henry E. & David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* New York: Rowman & Littlefield.

Brady, Henry E., David Collier & Jason Seawright. 2006. "Toward a Pluralistic Vision of Methodology." *Political Analysis* 14:353–368.

Bratton, Michael & Nicolas van de Walle. 1997*a*. *Democratic Experiments in Africa: Regime Transitions in Comparative Perspective.* Cambridge: Cambridge University Press.

Bratton, Michael & Nicolas van de Walle. 1997*b*. "Political Regimes and Regime Transitions in Africa, 1910-1994 (ICPSR Dataset 6996).".

Cardoso, Carlos. 1994. "A Transição Democrática Na Guiné-Bissau: Um Parto Difícil." *Soronda: Revista Estudos Guineeses* 17:5–30.

Commonwealth Observer Group. 1995. The Union Presidential and Parliamentary Elections in Tanzania, 29 October 1995. Technical report.

Dunning, Thad. 2012. *Natural experiments in the social sciences.* New York: Cambridge University Press.

Erdmann, Gero & Neo Simutanyi. 2003. "Transition in Zambia: The Hybridisation of the Third Republic.". Konrad Adenauer Foundation, Occasional Papers.

Fearon, James D. 2003. "Ethnic Structure and Cultural Diversity Around the World: A Cross-National Data Set on Ethnic Groups." *Journal of Economic Growth* 8(2):191–218.

Fearon, James D. & David D. Laitin. 2008. Integrating Qualitative and Quantitative Methods. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady & David Collier. New York: Oxford University Press pp. 756–76.

Forrest, Joshua B. 2005. Democratization in a Divided Urban Political Culture: Guinea-Bissau. In *The Fate of Africa's Democratic Experiments*. Bloomington: Indiana University Press pp. 246–66.

Gardinier, David E. 1997. Gabon: Limited Reform and Regime Survival. In *Political Reform in Francophone Africa*, ed. John F. Clark & David E. Gardinier. Westview Press chapter 9, pp. 145–61.

George, A.L. & A. Bennett. 2005. *Case studies and theory development in the social sciences*. MIT Press.

Hansen, Ben B. 2004. "Full matching in an observational study of coaching for the SAT." *Journal of the American Statistical Association* 99(467):609–618.

Hansen, Ben B. & Jake Bowers. 2008. "Attributing Effects to A Clustered Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association* 104(487):873–85.

Hansen, Ben B. & Stephanie Olsen Klopfer. 2006. "Optimal full matching and related designs via network flows." *Journal of Computational and Graphical Statistics* 15(3):609–627.

Herron, Michael C. & Kevin M. Quinn. 2009. "A Careful Look at Modern Qualitative Case Selection Methods." Presented at the 67th Annual Meetings of the Midwest Political Science Association.

Ho, Daniel E. & Kosuke Imai. 2006. "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 Election." *Journal of the American Statistical Association* 101:888–900.

Ho, Daniel E., Kosuke Imai, Gary King & Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Holmquist, Frank & Michael Ford. 1992. "Kenya: Slouching toward Democracy." *Africa Today* 39(3):97–111.

Hyde, Susan D. & Nikolay Marinov. 2012. "Which Elections Can Be Lost?" *Political Analysis* 20(2):191–210.

Ihonvbere, Julius O. 1997. "From despotism to democracy: The rise of multiparty politics in Malawi." *Third World Quarterly* 18(2):225–247.

Keele, Luke, Corrine McConnaughy & Ismail White. 2008. "Statistical Inference for Experiments." Unpublished manuscript, Ohio State University.

Kelsall, Tim. 2003. "Governance, Democracy and Recent Political Struggles in Mainland Tanzania." *Commonwealth & Comparative Politics* 41(2):55–82.

Kenya: The December 29, 1992 Election. 1993. IRI.

Krieger, Milton. 1994. "Cameroon's Democratic Crossroads, 1990-4." *Journal of Modern African Studies* 32(4):605–628.

Kruskal, J.B. 1964. "Nonmetric Multidimensional Scaling: A Numerical Method." *Psychometrika* 29(2):115–129.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-method Strategy for Comparative Research." *American Political Science Review* 99(3):435–52.

Lijphart, Arend E. 1975. "II. The Comparable-Cases Strategy in Comparative Research." *Comparative Political Studies* 8(2):158–77.

Lyons, Terrence. 2004. "Post-conflict Elections and the Process of Demilitarizing Politics: The Role of Electoral Administration." *Democratization* 11(3):36–62.

Malawi. 2002. In *Compendium of elections in Southern Africa*, ed. Tom Lodge, Denis Kadima & David Pottie. Johannesburg: Electoral Institute of Southern Africa.

Marcus, Richard R. 2004. "Political Change in Madagascar: Populist Democracy or Neopatrimonialism by Another Name?" Institute for Security Studies (South Africa) Occasional Paper 89, http://www.iss.co.za/pubs/papers/89/Paper89.htm, accessed July 6, 2012.

Mentan, Tatah. 1998. Cameroon: A Flawed Transition to Democracy. In *Democratization in Late Twentieth-Century Africa: Coping with Uncertainty*, ed. Jean-Germain Gros. Greenwood Press pp. 41–58.

Messone, Nelson N. & Jean-Germain Gros. 1998. The Irony of Wealth: Democratization in Gabon. In *Democratization in Late Twentieth-Century Africa: Coping with Uncertainty*, ed. Jean-Germain Gros. Greenwood Press pp. 41–58.

Miguel, Edward. 2004. "Tribe or Nation? Nation-Building and Public Goods in Kenya versus Tanzania." *World Politics* 56(3):327–362.

Mwase, Ngila & Mary Raphael. 1997. The Presidential Elections. In *Multiparty Democracy in Transition: Tanzania's 1995 General Elections*, ed. Samuel S. Mushi & Rwekaza S. Kumandala. Department of Political Science and Public Administration, University of Dar es Salaam pp. 149–84.

NDI & Carter Center. 1992. *The October 31 1991 National Elections in Zambia.* National Democratic Institute for International Affairs.

Newell, J. 1995. "A Moment of Truth, the Church and Political Change in Malawi, 1992." *Journal of Modern African Studies* 33(2):243–62.

Nwajiaku, Kathryn. 1994. "The National Conferences in Benin and Togo Revisited." *Journal of Modern African Studies* 32(3):429–47.

Posner, Daniel N. 2004*a*. "Measuring Ethnic Fractionalization in Africa." *American Journal of Political Science* 48(4):849–63.

Posner, Daniel N. 2004*b*. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98(4):529–546.

Quade, Dana. 1979. "Using weighted rankings in the analysis of complete blocks with additive block effects." *Journal of the American Statistical Association* 74(367):680–683.

Rosenbaum, Paul R. 2002. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Statistical Science* 17:286–304.

Rosenbaum, Paul R. & Jeffrey H. Silber. 2009. "Amplification of Sensitivity Analysis in Matched Observational Studies." *Journal of the American Statistical Association* 104(488):1398–1405.

Rudebeck, Lars. 2002. Multi-party Elections in Guinea-Bissau. In *Multi-party Elections in Africa.* Oxford: James Currey pp. 104–27.

Schraeder, Peter J. 1994. "Elites as Facilitators or Impediments to Political Development? Some Lessons from the "Third Wave" of Democratization in Africa." *Journal of Developing Areas* 29(1):69–90.

Smith, Benjamin. 2005. "Life of the Party: The Origins of Regime Breakdown and Persistence under Single-Party Rule." *World Politics* 57(3):421–51.

Takougang, Joseph. 1997. Cameroon: Biya and Incremental Reform. In *Political Reform in Francophone Africa*. Boulder, CO: Westview Press pp. 162–81.

Teorell, Jan, Marcus Samanni, Sören Holmberg & Bo Rothstein. 2011. "The Quality of Government Dataset, version 6 Apr 11." University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se.

Throup, David. 1993. "Elections and Political Legitimacy in Kenya." *Africa* 63(3):371–396.

Tordoff, Lord. 1992. "Multiparty elections in Kenya December 1992." *Representation* 31(115):57–59.

U.S. Department of State. 1990. *Country Reports on Human Rights Practices for 1990*.

U.S. Department of State. 1991. *Country Reports on Human Rights Practices for 1991*.

U.S. Department of State. 1992. *Country Reports on Human Rights Practices for 1992*.

U.S. Department of State. 1993. *1993 Human Rights Report: Central African Republic*.

U.S. Department of State. 1994. *Gabon Human Rights Practices, 1993*.

U.S. Department of State. 1995. *1995 Human Rights Report: Tanzania*.

van Cranenburgh, Oda. 1996. "Tanzania's 1995 Multi-Party Elections: The Emerging Party System." *Party Politics* 2(4):535–47.

van Cranenburgh, Oda. 2011. "Democracy promotion in Africa: the institutional context." *Democratization* 18(2):443–61.

van Donge, Jan Kees. 1995. "Kamuzu's Legacy – the democratization of Malawi." *African Affairs* 94(375):227–57.

VonDoepp, Peter. 1996. "Political transition and civil society: The cases of Kenya and Zambia." *Studies in Comparative International Development* 31(1):24–47.

Widner, Jennifer. 1991. "The 1990 Elections in Côte d'Ivoire." *Issue* 20(1):31–40.

Widner, Jennifer. 1994*a*. "Two Leadership Styles and Patterns of Political Liberalization." *African Studies Review* 37(1):151–174.

Widner, Jennifer A. 1994*b*. Political Reform in Anglophone and Francophone Africa. In *Economic Change and Political Liberalization in Sub-Saharan Africa*, ed. Jennifer A. Widner. Baltimore: Johns Hopkins University Press pp. 49–79.

Widner, Jennifer A. 1994*c*. "Single Party States and Agricultural Policies: The Cases of Ivory Coast and Kenya." *Comparative Politics* 26(2):127–147.

Appendix A: Data

We exclude countries with population less than a half million in 1989 (Cape Verde, Comoros, Djibouti, Equatorial Guinea, Seychelles, São Tomé and Príncipe). We also exclude Djibouti, Sudan, and Mauritania, which the United Nations and many scholars consider to be part of the Middle East/North Africa region rather than sub-Saharan Africa. Botswana, Lesotho, Mauritius, and South Africa are not presidential systems; Eritrea, Rwanda, Sudan, Somalia, Swaziland, Uganda, and Zaire/DRC did not hold transitional multi-party presidential elections in the 1990s. We also exclude Namibia, which held its first post-independence election just before the turn of the decade.

Plurality rule for transitional presidential election (T) = 1 if transitional presidential election used plurality rules, = 0 if it had a run-off provision to eliminate weaker candidates and hold a second round election if no candidate met a vote threshold in the first round. Coded from constitutions available through the African Legislatures Project (www.africanlegislatures project.org), Africa Elections Database (africanelections.tripod.com), Consortium for Elections and Political Process Strengthening (CEPPS) (www.electionguide.org), and Electoral Institute for the Sustainability of Democracy in Africa website (www.eisa.org.za).

Opposition harassment (Y) = 1 if there is any reported incident of opposition harassment before the election, = 0 otherwise. From National Elections across Democracy and Autocracy (NELDA) dataset v2 (Hyde & Marinov 2012).

Civil conflict = 1 if there was a civil conflict leading into the transition, = 0 otherwise.

Frequency of political protests during the transition period, 1988–92 = 0 if none, = 1 if some, = 2 frequent. To be counted, protests had to include explicit demands for political rights or changes in political rulers. Coded by Bratton & van de Walle (1997*b*).

Experience with military rule Share of years from independence to the transition under military rule, calculated using data from Bratton & van de Walle (1997*b*).

Ethnic Fractionalization Probability that two randomly selected people from a given country will belong to different ethnic groups, restricted to groups with at least 1 percent of country population (1990). The fractionalization index is $1 - \sum_{i=1}^{N} p_i^2$, where p_i is the proportion of people in each ethnic group and N is the total number of groups and ranges from 0 (perfectly homogeneous) to 1 (highly fragmented). From Fearon (2003).

Log GDP per capita Log GDP per capita, PPP adjusted. From World Development Indicators (data.worldbank.org/data-catalog), available through Quality of Government Dataset (Teorell et al. 2011).

Former French Colony = 1 if formerly a French colony, = 0 otherwise.

Appendix B: Matching

Authoritarian incumbents agree to hold multi-party elections because their right to rule has been challenged. We are concerned that weaker incumbents – those without a strong party organization or face a unified opposition – might be more likely to resort to opposition harassment and also be more likely to adopt electoral rules without a runoff provision for these transitional presidential elections. To address this concern, we match a country that had a runoff provision ($T = 0$) to each country that used plurality rules for the transitional presidential elections ($T = 1$). We pair match without replacement, exactly on three variables and on Mahalonobis distance with two variables, using the `optmatch` package version 0.8-1 (Hansen & Klopfer 2006) in `R` version 2.15.2.

The first matching variable is whether there was a civil conflict prior to the transition, since former combatants can be readily mobilized to harass opponents, and electoral rules under these circumstances are part of negotiated peace settlements designed to draw the armed factions into electoral politics (Lyons 2004). The second matching variable is the frequency of political protests in the transition period (coded as none, some, or frequent), as they indicate the public's dissatisfaction with the incumbent authoritarian regime. The third matching variable is experience with military rule, measured as the share of years since independence that the country was under military rule. We include this variable since countries with longer experience with civilian rather than military rule have a greater opportunity to develop political parties, and incumbents with political parties have stronger ties and better information about citizens and have the organization in place for voter mobilization (Bratton & van de Walle 1997*a*). We match exactly on these three variables. We match on Mahalonobis distance with two other variables. The first variable is log GDP per capita, which affects the regime's capacity to co-opt or repress the opposition (Bratton & van de Walle 1997*a*). The second variable is ethnic diversity from Fearon (2003), operationalized as ethnic fractionalization, which affects the complexity and difficulty of forming coalitions behind specific presidential candidates. This matching procedure produces four pairs: Cameroon–Gabon, Kenya–Côte d'Ivoire, Malawi–Zambia, and Tanzania–Guinea-Bissau. The data are presented in Table 14.

Pair (s)	Country	Treated (T)	Ethnic Frac.	Log GDP per capita	Civil Conflict	Protest into Transition	Military Rule Exp.
1	Cameroon	1	0.887	7.509	0	Frequent	0
1	Gabon	0	0.858	9.585	0	Frequent	0
2	Kenya	1	0.852	7.199	0	Frequent	0
2	Côte d'Ivoire	0	0.784	7.548	0	Frequent	0
3	Malawi	1	0.829	6.364	0	Frequent	0
3	Zambia	0	0.726	7.092	0	Frequent	0
4	Tanzania	1	0.953	6.712	0	Some	0
4	Guinea-Bissau	0	0.818	6.457	0	Some	0

Table 14: Data with Pair Matching

Log GDP per capita is larger for the treated unit in the fourth pair (Tanzania) than its matched control country (Guinea-Bissau), but it is smaller for the treated unit in all other pairs. The difference is greatest for the first pair, which includes Gabon, with its high oil revenues and a small population. This imbalance, with a one-sided p-value of 2/16 using the signed rank statistic, is not too troublesome for our analysis, because it will be improved by the subsequent full matching procedure.

There is also some imbalance on ethnic fractionalization (exact p-value of 1/16), which is lower for the matched control country than the treated country in every pair. In particular, Tanzania has a much greater ethnic fractionalization score than any of the twenty-four countries in our sample, and the within-pair difference on this variable is largest for the Tanzania–Guinea-Bissau pair. There are also a variety of concerns with this and other similar fractionalization measures, including debates about the what collection of people should count as an ethnic group, given the fluid nature of ethnicity, inter-ethnic marriages, and the possibility of membership in multiple groups; differing depths of divisions or social and cultural distance between groups; differing levels of salience of ethnicity in politics; whether all possibly enumerated groups are relevant for a particular analysis; and how the population shares of these groups can be best summarized. Tanzanian politics is frequently noted for its lesser emphasis on ethnicity compared to other African countries, sometimes attributed precisely to its large number of very small ethnic groups and to President Nyerere's sustained agenda of socialism and national integration (Kelsall 2003, Miguel 2004). Posner (2004*a*) reviews these concerns and has proposed an alternative fractionalization index of *politically relevant* ethnic groups, but it is not clear how appropriate such a measure is in a highly fluid moment of political transition and mobilization.

Nevertheless, it is important to match on some measure of ethnic diversity, since we believe the coalition possibilities of the ethnic groups in a country affects opposition strength and hence treatment assignment. Because we do not have strong theory for the relationship between ethnic diversity and treatment assignment, we adopt this measure widely used in cross-national analyses for the matching. We address the imbalance through additional qualitative information on opposition strength affected by ethnic diversity.

Appendix C: Ranking the Pairs

Using annual U.S. State Department Reports on Human Rights, monthly *Africa Research Bulletin* reports (*ARB*), news reports, election observer reports, and secondary sources, we rank the within-pair absolute differences in outcomes.

As we described in the main text, Kenya-Côte d'Ivoire is the pair with the largest absolute difference in outcomes.

Cameroon–Gabon is the pair with next largest absolute difference in outcomes. The October 1992 presidential elections in Cameroon were preceded by heavy intimidation of the opposition, including arrests and torture without explanation (Mentan 1998, 44–46). About 300 people were killed in just 1991 (Takougang 1997, 169) and a total of at least 400 democracy activists were killed in the two years leading up to the elections (Schraeder 1994, 81). In Gabon, transitional presidential elections in 1993 followed an extended period of demonstrations and strikes in urban areas and barricaded roads in rural areas (Gardinier 1997, Messone & Gros 1998). Throughout 1993, government security services intimidated opposition media with electronic jamming and the destruction or confiscation of radio transmission equipment (U.S. Department of State 1994). The National Assembly approved presidential decrees that severely curtailed press freedoms and most opposition newspapers were banned (Gardinier 1997). The U.S. State Department reports that "[Police were] absent from some opposition gatherings which were disrupted by violence attributed to street gangs paid by rival parties," but "[d]uring the 2 weeks prior to the elections, police acted quickly and effectively to assure that demonstrations and confrontations between the opposition and the [ruling] PDG remained peaceful" (U.S. Department of State 1994).

The difference in opposition harassment was still smaller between Malawi and Zambia. In Malawi, there were repeated instances of mass arrests of hundreds of opposition members and the detention and prosecutions of opposition leaders (*ARB*, Apr 1992 29:4, 10548–50; *ARB*, June 1992 29:6, 10618–9; *ARB*, Jul 1992 29:7, 10659; *ARB*, Nov 1992 29:11, 10793–4). Lodge et al. (2002) note that, "Both the UN Joint International Observer Group (JIOG) and the Malawi Electoral Commission (MEC) reported campaign violence and widespread intimidation, bribery and misuse of official positions" (130). The ruling Malawi Congress Party's paramilitary arm, the Young Pioneers, rounded up opposition supporters and detained them illegally on such a scale that "there was not enough room, and scores of detainees [had] to be held under guard in tents set up near Blantyre jail" (Ihonvbere, 1997, 238, citing *Africa Research Bulletin* 1992). There was also serious harassment of the opposition in Zambia [NDI and Carter Center, 1992, 44-45; *ARB*, Feb 1991 28:2, 10166; *ARB*, Sep 1991 28:9, 10284-5], but to a lesser extent than in Malawi. After political prisoners were released in July 1990, the government did not detail any additional opposition supporters. Still, several deaths resulted from ruling party supporters attacking opposition members, but in many instances the police arrested the attackers and generally allowed opposition rallies to be held (U.S. Department of State 1991, 453).

As noted in the main text, Tanzania and Guinea-Bissau has the smallest difference in outcomes of all the pairs. We find evidence of harassment over a more extended period in Tanzania than in Guinea-Bissau, although that in Tanzania is less than that in Malawi.

Appendix D: Strength of Opposition

We may be concerned that the matching described in Appendix B did not sufficiently model treatment assignment. More specifically, the matching variables may not fully capture the strength of opposition which affects the choice of electoral rules, so that the probability of receiving treatment is not equal across the countries in each pair. Further comparative case studies allow us to assess the relative strength of opposition, and hence whether treatment was more likely for one case than the other in each pair.

Kenya-Côte d'Ivoire: We cannot determine whether Kenya or Côte d'Ivoire had stronger opposition before the transition. The opposition in Kenya prior to the determination of electoral rules was fairly strong, but disunited. Kenya had a fairly active civil society and religious leaders who denounced obviously rigged single-party elections. Moreover, the opposition was strong enough that it could organize rallies, and riots followed the detention of two prominent politicians (Throup 1993, 389). In Côte d'Ivoire, President Houphouët-Boigny similarly announced a return to multi-party competition after decades of single-party rule following a general strike and several large demonstrations (U.S. Department of State 1990, 98). In contrast to Kenya, the opposition in Côte d'Ivoire had the advantage of having one main leader, Laurent Gbagbo, but he was in exile for most of the 1980s. Disaffected segments of society like students clashed with the government and organized demonstrations, but these events were not directed by the main opposition party or established religious leaders (*ARB*, Mar 1990 27:2, 9592).

Cameroon–Gabon: We believe that the opposition was somewhat stronger in Cameroon than in Gabon, but that the difference is not great. In both countries, the transition began with riots and strikes; the difference is that the Cameroonian opposition found a central opposition leader, while the Gabonese opposition were more divided between a long-standing opposition group in exile and other in-country leaders.

In Cameroon, the transition began with a security crackdown in 1990 on a large opposition rally in Bamenda, a city in the minority Anglophone region of the primarily Francophone country. This led to the deaths of six protestors and spurred further action. John Fru Ndi, an obscure former member of the governing party who had called for the rally, became a major opposition leader. The opposition parties under the umbrella National Coordination of Opposition Parties and Associations (NCOPA) and the unions began a general strike and implemented their "ghost town" strategy, demanding a national conference (Mentan 1998, 44-46). President Biya eventually conceded to this demand and the ruling CPDM started to fragment, but the opposition also fragmented (Krieger 1994, 608–12).

The transition in Gabon also began with strikes and riots, first among students but then spread to workers in both the public and private sectors that brought the country to a standstill. President Bongo called for a national conference with advisory status to the president, but the opposition rejected this arrangement and demanded multi-party competition. The Gabonese opposition was divided into at least two groups. One group was the *Parti Gabonais du Progrès*, which had support from the coastal regions and other groups, including workers and professors persecuted by the Bongo regime in the 1970s (Gardinier 1997, 149–52). It was the secretary-general of this party whose assassination in early 1990 provoked major rioting (Bayalama 1991, 68). Another group was *Mouvement de Redressement National* (MORENA), a group of exiles formed in France in the 1980s and originally led by the Catholic priest Paul Mba-Abessolé. It had support in the northern part of the country and from among Catholics and Protestants, who were concerned with Bongo's Islam and Masonry. Mba eventually broke away and formed the *Rassemblement National des Bûcherons*, but the opposition sought to unite for the presidential election (Messone & Gros 1998, 139).

Malawi–Zambia: The opposition was weaker in Malawi than in Zambia. In Malawi, political liberalization was spurred by pressure from foreign governments and international financial institutions, as well as domestic social unrest (Newell 1995). After "almost thirty years under the totalitarian control" of the Malawi Congress Party, there was little organized internal opposition to the regime (van Donge 1995, 229). Industrial action and student protests emerged only after the coordinated reading of the Pastoral Letter criticizing the regime in March 1992. Regime opponents in exile had little organization on the ground and formed new pressure groups only upon their return (Newell 1995).

One-party rule in Zambia was much less repressive than in Malawi, with the churches and student organizations more frequently expressing their discontent. The Zambian economy was heavily dependent upon copper mining, and the trade union movement remained powerful under President Kaunda and the United National Independence Party (UNIP). Beginning in 1989, the single-party regime was challenged by the Movement for Multi-party Democracy (MMD), led by Frederick Chiluba, the leader of the Zambian Congress of Trade Unions, with the support of prominent defectors from the ruling party, professionals, business people, and students (VonDoepp 1996, 32–3). Riots led the regime to announce a referendum on the return to multi-party politics, and the postponement of the referendum led to MMD-led mass demonstrations, forcing Kaunda to accept multi-partyism without a referendum and to schedule elections (Erdmann & Simutanyi 2003, 10-11).

Tanzania–Guinea-Bissau: We are unsure about whether the opposition was stronger in Tanzania or in Guinea-Bissau, although as a former Portuguese colony, Guinea-Bissau was probably more likely to have adopted runoff rules, like in Portugal, than Tanzania, which is a former British colony. In Tanzania, the transition was "managed" by the incumbent *Chama cha Mapinduzi* (CCM) party, the successor party to the Tanganyika African National Union (TANU) that brought the country to independence. Several opposition parties with regional bases were formed, but broad opposition coalitions repeatedly collapsed, and the opposition was "fragmented and weak" (van Cranenburgh 1996, 541–2). Mwase & Raphael (1997) also characterize the opposition as "fragmented, with parties of doubtful credibility and leadership" (153). Like the CCM in Tanzania, the ruling *Partido Africano da Independência da Guiné e Cabo Verde* (PAICG) in Guinea-Bissau was the party that led the fight for the country's independence. The demand for political liberalization in Guinea-Bissau initially came from younger ruling party members with high positions in the state (Rudebeck 2002, 115). This led to the legalization of the formation of many political parties, and as in Tanzania, the stronger opposition parties were those formed by ruling party defectors with a regional or ethnic base of support after this liberalization (Rudebeck 2002, 115, Forrest, 2005, 252–3, Cardoso, 1994, 26). Neither country had a group or politician that would likely become a strong opposition leader upon political liberalization.

Appendix E: Full Matching

With full matching, including exact matching on whether the country is a former French colony, our data are grouped into three sets. The first set is former French colonies with "frequent" protests in the transition period – Cameroon, Côte d'Ivoire, and Gabon, and later on also Madagascar. The second set is former British colonies with "frequent" protests in the transition period – Kenya, Malawi, and Zambia. The third set of countries are those with "some" protests in the transition period – Tanzania and Guinea-Bissau. These comprised the fourth pair in the original analysis. Table 15 presents the data.

Set (s)	Country	Treated (T)	Ethnic Frac.	Log GDP per capita	Former French	Civil Conflict	Protest	Military Rule
1	Cameroon	1	0.887	7.509	1	0	Frequent	0
1	Gabon	0	0.858	9.585	1	0	Frequent	0
1	Côte d'Ivoire	0	0.784	7.548	1	0	Frequent	0
1	Madagascar	0	0.861	6.821	1	0	Frequent	0.033
2	Kenya	1	0.852	7.199	0	0	Frequent	0
2	Malawi	1	0.829	6.364	0	0	Frequent	0
2	Zambia	0	0.726	7.092	0	0	Frequent	0
3	Tanzania	1	0.953	6.712	0	0	Some	0
3	Guinea-Bissau	0	0.818	6.457	0	0	Some	0

Table 15: Data with Full Matching

With full matching, we assess balance on the measured matching variables using Quade's statistic (Section 5.1). In the first set, adding Côte d'Ivoire to the original Cameroon–Gabon pair reduces the difference in log GDP per capita between treated and control units but increases the difference in ethnic fractionalization in that set. By adding Madagascar and accepting a small imbalance on previous experience with military rule, balance on both of these variables is further improved in this set. While Zambia has lower ethnic fractionalization than both Kenya and Malawi, its log GDP per capita falls in between those two treated units. For log GDP per capita, the one-sided

p-value with Quade's statistic is 5/18 without Madagascar and 11/24 with Madagascar. For ethnic fractionalization, the p-values are 1/18 and 1/24, respectively. As before, we address the imbalance on ethnic fractionalization through the overall qualitative assessment of opposition strength.

We assess opposition harassment to be greatest in Cameroon, second greatest in Gabon, then Côte d'Ivoire, and least in Madagascar. We consider $Y_{Gabon} > Y_{Cote\ d'Ivoire}$ because opposition rallies were disrupted more frequently and opposition media was repressed more violently in Gabon than in Côte d'Ivoire. Furthermore, there was greater opposition harassment in both of these countries than in Madagascar. Once the Malagasy opposition successfully pressured incumbent President Ratsiraka into holding a constitutional convention and multi-party elections, the election itself proceeded fairly smoothly (Marcus 2004). The one incident of violence reported involves the army, which was loyal to the transitional government, killing several Ratsiraka supporters while the latter demonstrated in favor of secession by several regions of the country (*ARB*, Oct 1992 29:10, 10759–60).

Opposition harassment (Y) was greater in Kenya, where hundreds were killed and tens of thousands displaced, than in Malawi, which had mass arrests and detention of opposition members. As we discussed earlier, Malawi had more opposition harassment than in Zambia (Appendix C), so $Y_{Kenya} > Y_{Malawi} > Y_{Zambia}$. Moreover, because of the great extent of opposition harassment in Kenya, we believe $abs(Y_{Kenya} - Y_{Zambia}) > abs(Y_{Cameroon} - Y_{Madagascar}) > abs(Y_{Cameroon} - Y_{Cote\ d'Ivoire})$. We also assess $abs(Y_{Cameroon} - Y_{Cote\ d'Ivoire}) > abs(Y_{Tanzania} - Y_{Guinea-Bissau})$. These relationships allow us to provide the ranks in Table 12 as well as to determine that the Tanzania–Guinea-Bissau pair defines the lower bound of the one-sided 95% confidence interval.

Within the set of former French colonies, opposition strength was greatest in Madagascar, followed by Cameroon, and then Côte d'Ivoire and Gabon. The opposition in Madagascar pushed for liberalization extremely forcefully behind Zafy Albert, who led a strike of nearly 80,000 civil servants and later a march of 400,0000 people on the city center which was effectively a general strike. Zafy Albert was able to claim a parallel government, proclaim himself prime minister, and push for a transitional government (Marcus 2004). The opposition in Cameroon was also capable of organizing extended, large general strikes ("ghost town" strategy), but was less unified than the Malagasy opposition. Appendix D described the Gabonese opposition as weaker than the Cameroonian opposition. It is difficult to compare the strength of opposition in Côte d'Ivoire and Gabon. The Ivoirien opposition was more unified under Laurent Gbagbo than the Gabonese opposition which had several leaders, but it is not clear that the pressure they put on their respective authoritarian incumbents was so different as to affect their relative probability of adopting plurality rule. Our best guess is that plurality rule for transitional presidential elections was most likely to have been adopted in Madagascar, followed by Cameroon, then Côte d'Ivoire and Gabon.

Within the set of former British colonies, opposition strength was greater in Zambia than in Kenya, which in turn had a stronger opposition than Malawi. Zambia had a unified opposition led by the head of the Zambian Congress of Trade Unions and broad support from society. Kenya had an active civil society and an internal opposition that had the capacity to organize large rallies, while the opposition in Malawi was much more divided and was the least threatening to the incumbent regime in this group.