Portfolio

Jacob B. Schumaker

Table of Contents

Time Series Project

Disease and Conservatism

May 2014

Jacob Schumaker

Abstract

This study examines whether fear of disease causes political conservatism. A large body of psychological research finds that fear is a major factor in conservatism. A smaller body of studies theorizes that fear of disease specifically causes conservatism as well. However, most of these studies use cross-sectional data. The present study uses panel data to employ a stronger identification strategy. The results of the study do not support the fear of disease-conservatism hypothesis. Instead, it finds that increases in deaths from the influenza virus are not associated with a person becoming more liberal.

Introduction

Since the beginning of the 20th century, the developed world has undergone a massive improvement in public health. People are living longer, infant mortality has dropped dramatically, and disease pandemics occur less often. At the same time, society has become more liberal and tolerant. Attitudes toward ethnic minorities, homosexuals, women, and other traditionally stigmatized social groups have become significantly more progressive and accepting. The research presented here investigates whether these two trends are linked.

Recent psychological research has provided initial evidence that the trends are related. One strand of this research has found that political conservatism and prejudice are strongly linked to fear of disease. For example, Green et al. (2010) finds that people who are averse to germs are opposed to immigration. Importantly, they find that this is not because they believe the probability of a disease pandemic is higher than individuals who are not averse to germs believe it is. Instead, they find that germ aversion leads to anti-immigrant attitudes by making social attitudes more conservative. This suggests that people are opposed to immigration not for direct, possibly rational reasons but because fear of disease changes how they view the world.

Other research reaches similar conclusions. Van Leeuwen, Park, Koenig, & Graham, (2012) finds that people in countries where disease is more prevalent are more likely to have concerns about society that have been linked to conservatism and prejudice, including a preference for their in-group, a preference for authority, and a concern about moral and bodily purity. Another study, Huang, Sedlovskaya, Ackerman, & Bargh (2011), finds that 1) people who are vaccinated against the flu have more positive feelings toward outgroups, 2) framing vaccines as injections of the disease rather than protection from the disease causes more negative feelings toward outgroups, and 3) that making people feel more protected against illness by giving them a hand wipe makes people feel more positively toward outgroups. These studies suggest that reducing concerns about illness through public health improvements could reduce prejudice.

This literature suggests that a fear of pathogens could make a person prejudiced because of evolutionary pressure (Huang et al. 2011). Early humans may have evolved a fear of out-groups because people who were unfamiliar to them were more likely to carry diseases they lacked immunity for. This fear of out-groups results in not just instincts to physically avoid them but in a mental distaste—or prejudice—toward them (Huang et al. 2011).

There is also a large and important body of recent psychological research by Haidt and others that suggests that a major cause of conservatism is fear and a need for protection (Haidt 2008; Janoff-Bulman 2009; Jost 2006; Jost et al. 2003; van Leeuwen and Park 2009; Park and Isherwood 2011). For example, Jost (2006) finds that conservatives are more afraid of death, less open to new experiences, and have a greater fear of threats and losses. Along the same lines, the van Leeuwen et al. (2012) study cited above that theorizes that higher disease prevalence in a country represents a threat to a person's health and that this threat causes a psychological desire for protection in individuals.

If disease fears lead to conservative social attitudes as existing research shows, fear of disease should also lead to conservative political attitudes. The present study examines this possibility by looking at the link between disease deaths and political conservatism. Specifically, it examines whether the number of people who die from the influenza virus in a person's state is associated with an increase in their self-reported political ideology using panel data from American National Election Studies and the Centers for Disease Control. Estimating the effect of flu deaths using respondent fixed effects, the results do not support the study's theory. Instead, the results show that more flu deaths do not make survey respondents more liberal.

Data, Estimation, and Results

The goal of this study is to examine whether fear of disease causes people to become more conservative. To operationalize that test, the study uses panel survey data to look at whether respondents tend to become more conservative when more people die of the influenza virus in their state. The survey data used comes from the American National Election Studies 2008-09 and 2010 panel study. Data on influenza deaths comes from the Centers for Disease Control. To estimate the effect of changes in the influenza death rate, the study takes advantage of the panel data to use survey respondent fixed effects. It finds that increases in flu deaths are not associated with respondents identifying themselves as more liberal, counter to the study's theory. In the following sections, I describe the data used in the study, the method used to analyze the data, and the results of that analysis.

Data

The data used in this study comes from two sources: the American National Election Studies (ANES) and the Centers for Disease Control (CDC). The ANES data comes from the ANES Panel Study 2008-09 ("American National Election Studies (ANES) Panel Study, 2008-2009" 2011) and the ANES Panel Recontact Study 2010 ("American National Election Studies (ANES) Panel Recontact Study, 2010" 2011). The ANES Panel Study 2008-09 consisted of a series of internet surveys that re-interviewed the same panel of respondents over the course of the two years of the study.[1] The ANES Panel Study used a representative panel of Americans, although the results presented below do not use respondents with missing data which could make it less representative. The ANES Panel Recontact Study 2010 re-interviewed participants from the 2008-09 ANES Panel Study in June and July of that year.

The outcome of interest—political conservatism—comes from the ANES. It was assessed by asking survey respondents about their ideological leanings. Using a series of

[1] The first cohort of respondents were interviewed in January 2008 and a second cohort was added to the study in September 2008.

questions about how liberal or conservative they feel, respondents chose one of seven categories ranging from “Very Liberal”—a 1 on the scale—to “Neither”—4 on the scale—to “Very Conservative”—7 on the scale. It was recorded in seven waves of the survey. Table 1 shows that average ideology over the course of the panel was 4.31, or leaning slightly conservative.

TABLE 1 SUMMARY STATISTICS

Variable	Data	Mean	SD	Min	Max	Obs
Ideology	overall	4.31	1.88	1	7	N = 11425
	between		1.76	1	7	n = 2960
	within		0.69	-0.02	8.31	T = 3.8598
Non-disease	overall	11.72	24.39	0.15	237.03	N = 108180
Death Rate	between		24.34	0.24	207.56	n = 3005
	within		1.65	-13.99	41.19	T = 36
Flu Death Rate	overall	1.46	0.49	0.47	5.06	N = 104543
	between		0.24	0.87	2.62	n = 2936
						T-bar =
	within		0.43	0.27	4.90	35.6073
Respondent Income	overall	12.31	4.08	1	19	N = 4485
	between		4.03	1	19	n = 2967
	within		1.02	3.31	21.31	T = 1.51163
State Income	overall	5274.24	4186.90	279.00	15389.80	N = 108180
	between		4183.16	288.19	14606.94	n = 3005
	within		192.35	4788.30	6057.10	T = 36

Note: This table displays summary statistics for the panel data used in the study. ‘overall’ is data for the dataset as a whole, ‘between’ is between-observation data, and ‘within’ is within-observation data. N is the total number of observations for each variable, n is the number of observations per time period, and T is the average number of time periods per observation.

Data on the income of respondents also comes from the ANES. Income is used in the estimation presented later as a control variable. It was also derived using a series of questions. Income is on a 19-point scale, ranging from 1 to 19. Higher values indicate higher income, with the lowest possible income being less than $5,000 and the highest possible income being more than $175,000. The modal response was 13 on this variable, indicating a modal income of between $50,000-$60,000. Respondent income was derived for two waves of the study, the first wave in 2008 and the recontact wave in 2010 and thus results that use income contain far fewer observations than results that do not use it.

The primary independent variable of interest, the state monthly death rate due to the influenza virus, comes from the Centers for Disease Control's (CDC) WONDER (Centers for Disease Control and Prevention 2012), an online database of disease information. All types of the influenza virus were included in the analysis, including common strains of the flu as well as more rare varieties like avian and swine flu. Avian and swine flu not analyzed separately because they kill so few people that the CDC will not publish monthly state death totals.

The death rate is calculated as the total number of flu deaths in a month per 100,000 state residents. State populations also come from the CDC and are measured at the year level, because monthly population estimates were not available. Death rates are not adjusted to account for differences in age demographics, which is one drawback of using monthly data from WONDER rather than yearly data. Age-adjusted data would take into account that some age groups are more likely to die of diseases than others and adjusts the data to show how high the death rate is relative to what would be expected given the population's age demographics. Future iterations of this study will attempt to age-adjust the data manually to guard against the possibility that changes in the death rate are proxying for the population aging.

Estimation and Results

This study's primary empirical strategy is to look within respondents at how their ideology changed in response to change in their state's influenza death rate. By looking at within-respondent variation using respondent fixed effects, this strategy controls for any unobserved characteristics of respondents that are correlated with the flu death rate of their state and their ideology. For example, it controls for the possibility that people who are more conservative tend to live in more conservative states and that conservative states are less healthy. This study also employs deaths from non-disease causes as a second independent variable to ensure that flu deaths affect ideology through fear or disease and not through some alternative mechanism.

This estimation is described by the following equation:

$$Ideology_{it} = \beta_0 + \beta_1 Respondent_i + \beta_2 Flu\ Death\ Rate_{st} + \beta_3 Income_{it} + \varepsilon_{it}$$

where i indexes respondents, t indexes months, and s indexes states. Ideology is the self-placed ideology of respondent i at month t, Respondent is a respondent fixed effect for each respondent i, Disease Death Rate is the death rate due to disease in the respondent's state s, Income is the respondent i's income at time t, and ε is an error term for respondent i at time t.

Table 2 Dependent Variable: Ideological Self-placement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
State Disease Death	-0.072	-0.072	-0.158	-0.123	-0.033	-0.032	-0.081	-0.083
	(0.027)	(0.027)	(0.083)	(0.087)	(0.043)	(0.043)	(0.090)	(0.090)
	{1}	{1}	{.2}	{.56}	{1}	{.99}	{.85}	{.93}
State Non-disease Death		-0.005**	0.010*	0.025*		0.001	0.001	0.002
		(0.006)	(0.042)	(0.044)		(0.001)	(0.001)	(0.002)
		{0}	{0}	{0}		{.11}	{.30}	{.25}
Income			-0.022	-0.020			0.005	0.006*
			(0.019)	(0.019)			(0.009)	(0.009)
			{.38}	{.69}			{.20}	{.04}
State Income				-0.000***				-0.000***
				(0.000)				(0.000)
				{0}				{0}
Constant	4.405	4.459***	4.718	5.208	4.353	4.337	4.357	4.404
	(0.036)	(0.073)	(0.535)	(0.662)	(0.060)	(0.061)	(0.181)	(0.184)
	{1}	{0}	{1}	{1}	{1}	{.94}	{.88}	{.85}
Fixed Effects	x	x	x	x				
Observations	10985	10985	2966	2966	10985	10985	2966	2966
R^2	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
Adjusted R^2	-0.35	-0.35	-2.76	-2.75	-0.00	0.00	-0.00	-0.00
AIC	22898.559	22899.563	4001.072	3997.097	44949.684	44949.005	12180.828	12180.927
BIC	22913.168	22921.476	4025.052	4027.071	44964.293	44970.918	12204.807	12210.902

Standard errors in parentheses. Permutation test p-values in brackets. Fixed effects are at the respondent level.

Note: The dependent variable, ideological self-placement, is on a liberal-to-conservative scale. Higher values indicate more conservative ideologies.

Note 2: Models that include respondent income use fewer observations because respondent income was asked about in only two of the seven waves of the survey.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Columns 1 through 4 of Table 2 contain the main results of interest. These columns display results obtained when respondent fixed effects are used. Column 1 display the results from the model that includes only the flu death rate, Column 2 adds the non-disease death rate, and Columns 3 and 4 add income and state income as control variables. Both results shows that the monthly rate of flu deaths is not associated with a respondent's political ideology. When fixed effects are not used the results remain the same (Columns 5 through 8).

This study also employs a permutation test to deal with non-independent error terms between respondents. Respondent error terms will not be independent of each other because they are affected by disease rates at the state level. This inflates the apparent precision of standard error calculations that do not account for the dependence. To correct this problem, the study uses p-values calculated from permuting state flu death rates within time periods. This will tend to produce significance tests that are more conservative and closer to being correct.

The results of Column 4 show that an increase of 1 death per 100,000 state residents in a month is associated with a 0.12 decrease in respondent's ideological self-placement score, which is the equivalent of a person's ideology becoming slightly more liberal. A 0.12 effect of a 1 in 100,000 increase in the flu death rate means that for a respondent to become a full point more liberal, it would take about 8 more deaths due to disease per 100,000 residents. Moving one point lower on the 7-point self-placement scale from a 5 to a 4 corresponds to a survey respondent's ideology moving from closer to conservatives than to liberals to being closer to ideologically neutral.

To guard against the possibility that the flu deaths were related to political ideology because death has an effect on ideology, not because flu deaths increase fear of disease, this study also includes the death rate due to causes besides disease as an independent variable. Data on non-disease or "external" causes of death also comes from the CDC WONDER database and includes deaths from car accidents and homicides. If flu deaths have the same effect on ideology as non-disease deaths that would indicate that death has the same effect on a person's political leanings regardless of whether it is disease-related. As in the regression using only flu deaths, the main models of interest include fixed effects.

This estimation is described with the following equation:

$$Ideology_{it} = \beta_0 + \beta_1 Respondent_i + \beta_2 Flu\ Death\ Rate_{st} + \beta_3 Non - disease\ Death\ Rate_{st} + \beta_4 Income_{it} + \varepsilon_{it}$$

The results that include non-disease deaths are displayed in Columns 2, 3, and 4. They again show no significant relationship between flu deaths and ideology. The effect of flu deaths does come closer to reaching statistical significance, with the p-value from a permutation test dropping from 1 in the model without non-disease deaths (Column 1) to 0.17 in the full model with non-disease deaths and control variables (Column 4).

Conclusion

This study finds that deaths from disease are not associated with political conservatism. Specifically, it finds that when the death rate due to disease increases people do not seem to identify themselves as more conservative. Alternatively, there appears to be no relationship between disease and ideology.

This finding is in contrast to other findings on the relationship between disease and conservatism. Green et al. (2010), Van Leeuwen, Park, Koenig, & Graham, (2012), and Huang et al. (2011) all find that fear of disease is associated with more conservative or prejudiced attitudes. These studies may have reached found significant effects because they use cross-sectional data and therefore may not have well-identified results. The present study uses panel data and could have more accurate results. However, the hand wipe and framing studies in Huang et al. (2011) are experimental, an even stronger identification tool than the present study's fixed effects. While the experiment in Huang et al. (2011) is not conclusive, it suggests that there may be a link between fear of disease and social attitudes, although the nature of that link is not yet clear.

This research adds to the literature on how public policy affects politics. Most political science research studies how politics affects public policy (what policies get chosen and why). The small number of studies that examine how policy affects politics focus on policy feedback, i.e. how citizens' perceptions of a policy affect their support for that policy.

The present study goes beyond this literature by looking at whether policies can affect core political values.

When taken together with previous research on fear of disease and conservatism, this study sheds light on whether public health revolution had a role in making societies more tolerant. As the world has gotten safer, cleaner, and more comfortable society has become more liberal and tolerant. Other studies suggest that one reason this association exists is because people become more open and accepting when they feel less threatened. However, this study does not find evidence of that fact. Nevertheless, there remains the possibility that fear of disease and conservatism or prejudice are linked. This study will be followed up by experiments that investigate that link.

References

"American National Election Studies (ANES) Panel Recontact Study, 2010." 2011. *Research, Inter-university Consortium for Political and Social.* http://doi.org/10.3886/ICPSR30721.v1.

"American National Election Studies (ANES) Panel Study, 2008-2009." 2011. *Research, Inter-university Consortium for Political and Social.* http://doi.org/10.3886/ICPSR29182.v1.

Centers for Disease Control and Prevention, National Center for Health Statistics. 2012. "Underlying Cause of Death." *CDC WONDER Online Database*: 1999–2010. http://wonder.cdc.gov/ucd-icd10.html .

Green, Eva G.T . et al. 2010. "Keeping the Vermin Out : Perceived Disease Threat and Ideological Orientations as Predictors of Exclusionary Immigration Attitudes." 316(April): 299–316.

Haidt, J. 2008. "Morality." *Perspectives on Psychological Science* 3(1): 65–72. http://pps.sagepub.com/lookup/doi/10.1111/j.1745-6916.2008.00063.x.

Huang, Julie Y., Alexandra Sedlovskaya, Joshua M. Ackerman, and and John A. Bargh. 2011. "Immunizing Against Prejudice Effects of Disease Protection on Attitudes Toward Out-Groups." *Psychological Science* (November). http://pss.sagepub.com/content/22/12/1550.short (October 6, 2013).

Janoff-Bulman, Ronnie. 2009. "To Provide or Protect: Motivational Bases of Political Liberalism and Conservatism." *Psychological Inquiry* 20(2-3): 120–28. http://www.tandfonline.com/doi/abs/10.1080/10478400903028581 (November 22, 2013).

Jost, John T. 2006. "The End of the End of Ideology." *The American psychologist* 61(7): 651–70. http://www.ncbi.nlm.nih.gov/pubmed/17032067 (November 7, 2013).

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129(3): 339–75. http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.129.3.339 (November 7, 2013).

Van Leeuwen, Florian, and Justin H. Park. 2009. “Perceptions of Social Dangers, Moral Foundations, and Political Orientation.” *Personality and Individual Differences* 47(3): 169–73. http://linkinghub.elsevier.com/retrieve/pii/S0191886909000919 (November 9, 2013).

Van Leeuwen, Florian, Justin H. Park, Bryan L. Koenig, and Jesse Graham. 2012. “Regional Variation in Pathogen Prevalence Predicts Endorsement of Group-Focused Moral Concerns.” *Evolution and Human Behavior* 33(5): 429–37. http://linkinghub.elsevier.com/retrieve/pii/S1090513811001413 (November 25, 2013).

Park, Justin H., and Edward Isherwood. 2011. “Effects of Concerns About Pathogens on Conservatism and Anti-Fat Prejudice: Are They Mediated by Moral Intuitions?” *The Journal of Social Psychology* 151(4): 391–94. http://www.tandfonline.com/doi/abs/10.1080/00224545.2010.481692 (November 25, 2013).

Examining the Relationship Between Perceived Risk and Availability of Drugs

Stephen Mike McLaughlin and Jacob Schumaker

May 4, 2014

1 Introduction

We are interested in substance abuse, specifically the abuse of prescription drugs. As part of our study we want to understand how people perceive the risks and availability of drugs. To study these issues we will use the National Survey of Drug Use and Health, an annual survey of Americans aged 12 and older.

2 Data and Methodology

2.1 Data

The data for this project comes from the National Survey of Drug Use and Health 2010 (NSDUH 2010). The NSDUH is a series of nationally representative, cross-sectional surveys intended to track trends in drug use in the United States. Each year of the NSDUH contains information from roughly 60,000 individuals.

We use four variables to measure how survey respondents perceive the health risks of various drugs. On the survey, questions about risk are phrased as follows: ”How much do people risk harming themselves physically and in other ways when they use [DRUG] (frequency)?”. [DRUG] is one of: a pack of cigarettes, marijuana, heroin, or cocaine. (frequency) is ”every day” for cigarettes and once a month for all other drugs. The risk variables are all on a 4 point scale, with 1 representing ”No Risk” and 4 representing ”Great Risk.”

There are three variables on the survey that measure the availiability of drugs. Questions in this section are phrased as follows: ”How difficult or easy would it be for you to get some [DRUG], if you wanted some?” where [DRUG] is marijuana, cocaine, or heroin. These variables are one a 1-5 scale, with 1 meaning ”Probably Impossible” and 5 meaning ”Very Easy”.

For this project, we ignore the sampling design that the NSDUH uses. On the whole, this means that the analysis will reflect a population that is slightly younger and ’harder to reach’ than the United States population. This makes it likely that we will underestimate the perceived risk and overestimate the perceived availability of drugs relative to the United States as a whole. There is also missing data, which may be more problematic. We ignore the missing data problem for now. In a more formal analysis we would consider using multiple imputations or similar procedures to address the issue.

2.2 Methods

We use three methods in this project: k-means clustering, principle components analysis, and factor analysis.

3 K-Means Clustering

We begin the analysis by using k-means clustering to classify respondents into general catogories. The intent is to see if there is any sort of ’grouping’ that arises in the data. We estimate the k-means algorithm using R and standardize the data to avoid problems due to our variables being measured using different scales. We use Euclidean distance when conducting this analysis.

The resulting elbow plot is displayed in Figure **??**. We do not include a ’null’ plot because the necessary permutations would take too long to run:

[Figure 1 about here.]

Based on visual inspection of the elbow plot, we conclude that there are three clusters in the data. The centroid of these data are displayed in Table **??**:

[Table 1 about here.]

The centroid means support the conclusion that there are three distinct clusters. Looking at cluster 1, we see a group that seems to have average beliefs about the risk of cigarettes, believes marijuana is slightly less risky than average but heroin and cocaine are more risky, and thinks that marijuana, cocaine, and heroin are much more available than average. Group 2 perceives much less risk in drug use, particularly heroin and cocaine, than average but surprisingly they are below average in when considering the availability of drugs. Group 3 perceives the most risk, but is like Group 2 in their perception of drug availability. The distribution across clusters is 45.23/8.91/45.85 respectively.

Given this information, we conclude that Groups 1 and 3 represent individuals that have little experience with drug use. If anything, a representative individual from Group 1 might smoke marijuana if it's around and might have tried cocaine in the past but is far from a regular uses. Group 3 could be described as tee-totalers: people with little or no drug experience and who think drugs are difficult to get. Neither groups seems likely to know much about the avaiability of drugs.

Group 2 may be the most interesting. A representative individual from Group 2 might actually have some hard drug experience and may even use some drugs regularly (marijuana, cocaine). They probably smoke cigarettes, might have less education and income than either of Group 1 and 2, and might live in an area where drugs are more widely available. However, they are either unwilling to admit that they know more about drug availability or they are are actually telling the truth and in reality most people do not know where to get drugs. Our prior knowledge of public health says that it is a mixture of both factors, but more heavily weighted towards the possibility that they do not know where to find drugs.

In summation, we conclude there are three distinct types of individuals. Going forward, we will describe them as 'Average Joes', 'Teetotalers', and 'The Real Deal' to represent groups 1, 3, and 2 respectively.

4 Principal Components

Using k-means clustering, we concluded that survey respondents can be clustered into three groups, which we describe as 'Average Joes', 'Teetotalers', and 'The Real Deal'. In this section, we use Principal Components Analysis to see if we can find underlying dimensions among which these groups may differ. That is, can we find a way to reduce the risk and availability perceptions to dimensions that might allow us to separate the three groups?

To begin, we standardize data and look at the correlations matrix, which is displayed in Table **??**:

[Table 2 about here.]

It appears that the perceptions of risk are positively correlated, as are perceptions of availability. This is not surprising, as we would expect individuals who think marijuana or cocaine are easily available would also think heroin is easily available. Similarly, we would expect individuals who think there is risk in consuming cocaine will also think there is risk in consuming heroin. The correlations between the risk and availability domains are less clear. It appears that the perception of the risks of marijuana and availability of marijuana are fairly negatively correlated but beyond this we find that the correlations are fairly weak and do not provide much insight.

To check whether PCA is appropriate for the data we calculate the proportion of correlations with absolute value greater than 0.2. Approximately 43% of the correlations meet this criteria, which suggests PCA will perform moderately well. The scree plot resulting from the PCA is displayed in Figure **??**:

[Figure 2 about here.]

The scree plot suggests there are 3 or 4 components that explain most of the variance. We confirm this result using a parallel plots analysis. Because the parallel plots code available at *reuningscherer.net* only accepts data with $n < 1000$, we run the PCA again using only the first 999 observations from the data. The resulting parallel plot is displayed in Figure **??**.

[Figure 3 about here.]

This parallel plot confirms our conclusions from the scree plot: there are 3 components that explain almost all of the variance. We report the first three loadings in Table **??**.

[Table 3 about here.]

It is unclear what these loadings mean. It seems that there is variation along a dimension of perceived availability and the perceived risk of marijuana use, and the weights move against one another (Loading 1). Loading 2 appears to point towards a risk perception dimension, and the meaning of Loading 3 is unclear. In short, it is not clear what sub-space might fully describe the dimensions our three groups vary on.

5 Factor Analysis

Using k-means clustering, we identified three distinct groups of individuals in the data. Using PCA, we also attempted to find a small number of variables that describe the three groups. For example, one dimension appears be 'drug experience' and another appears to be 'drug perceptions'. PCA was not helpful in finding such dimensions[1], so next we try factor analysis.

We first consider whether the data is appropriate for exploratory factor analysis. We have seven indicator variables and two natural ways to classifying survey respondents: by their risk Perception and availability perception. Seven indicators is typically enough to describe two latent factors, so we proceed with factor analysis. We also note that in our PCA analysis, we estimated the correlation matrix between variables. The matrix is displayed in Table **??**. In it, we see moderately high correlation among the risk variables and high correlation among the perceived availability variables. This suggests there might be two factors to separate the groups. Finally, we calculate the KMO measure of accuracy and obtain a value of 0.62785, indicating that the data is acceptable for exploratory factor analysis.

We perform factor analysis with two extraction methods and assume there are two latent factors. The extraction methods are iterative principle components and principle axis factoring. We compare the two results based on the root mean square residual and the proportion of residual correlations that are greater than 0.05 in magnitude. The results suggest that principle axis factoring is more appropriate. The root mean square residual is slightly larger for principal components (0.0722 compared to 0.0682), but the proportion of residual correlations greater than 0.05 in absolute value is much lower for principal components (0.1905 compared to 0.381).

We use both quartimax and varimax rotations. The loadings plot for the different rotations appears in Figure **??**.

[Figure 4 about here.]

The loading plot indicates two pieces of information. First, there is little difference between a varimax and quartimax roation. Second, two factors are adequate to separate the risk and availability perceptions, however there may be a third factor relating to perceptions of the risk of marijuana. The loadings from the varimax rotation are in Table **??**:

[1] This is discussed in the class notes on Factor Analysis. PCA is not suited for identifying unique factors, but we wanted to try it first because it is typically easier to estimate and interpret

[Table 4 about here.]

These loadings suggest there may be a third latent factor. This third factor may be something like experience with marijuana or living in a liberal state. This may explain why the perceived risk of marijuana is different than the perceived risks of drugs in general. In other words, it could be that whether a person perceives marijuana as dangerous depends on whether they live in a culture that has decided that marijuana is not an illicit drug. People in these cultures may be similar to other cultures in how risky they perceive hard drugs, but they have different views on marijuana because they think about it as a soft, non-illicit drug.

6 Conclusion

In this project, we examined the relationship between the perceived health risks and availability of drugs using data from the 2010 National Survey of Drug Use and Health. Using k-means clustering, we find evidence that there are three groups of drug users, which we described as 'Average Joes', 'Teetotalers', and 'The Real Deal'. 'Average Joes' appear to perceive most drugs as risky, but do not perceive marijuana as risky. They also think that drugs are relatively available in their neighborhoods. 'Teetotalers' view all drugs as carrying health risks, including marijuana, but do not perceive drugs to be widely available. 'The Real Deal' are a small fraction of individuals who do not think that drugs are dangerous and believe that drugs are not easily available where they live.

Using principle components analysis and factor analysis, we attempted to find underlying dimensions along which we may be able to explain these three groups. Principle components analysis does not yield insights that are easily interpretable, however factor analysis seems to indicate that risk perception and perception of availability can be separated. However, the perceived risk of marijuana seems to behave differently than risk perceptions of other drugs. We believe there may be a third factor that can explain why this is the case, however we need more indicators to look into that possibility.

In sum, we find k-means clustering to be an effective tool for examining the relationship between risk perception and the perceived availability of drugs. Factor analysis shows promise as a tool for separating the latent variables that may drive the results of clustering, however more investigation is necessary to pin down exactly what those factors may be.

List of Figures

Figure 1: K-Means Within SS

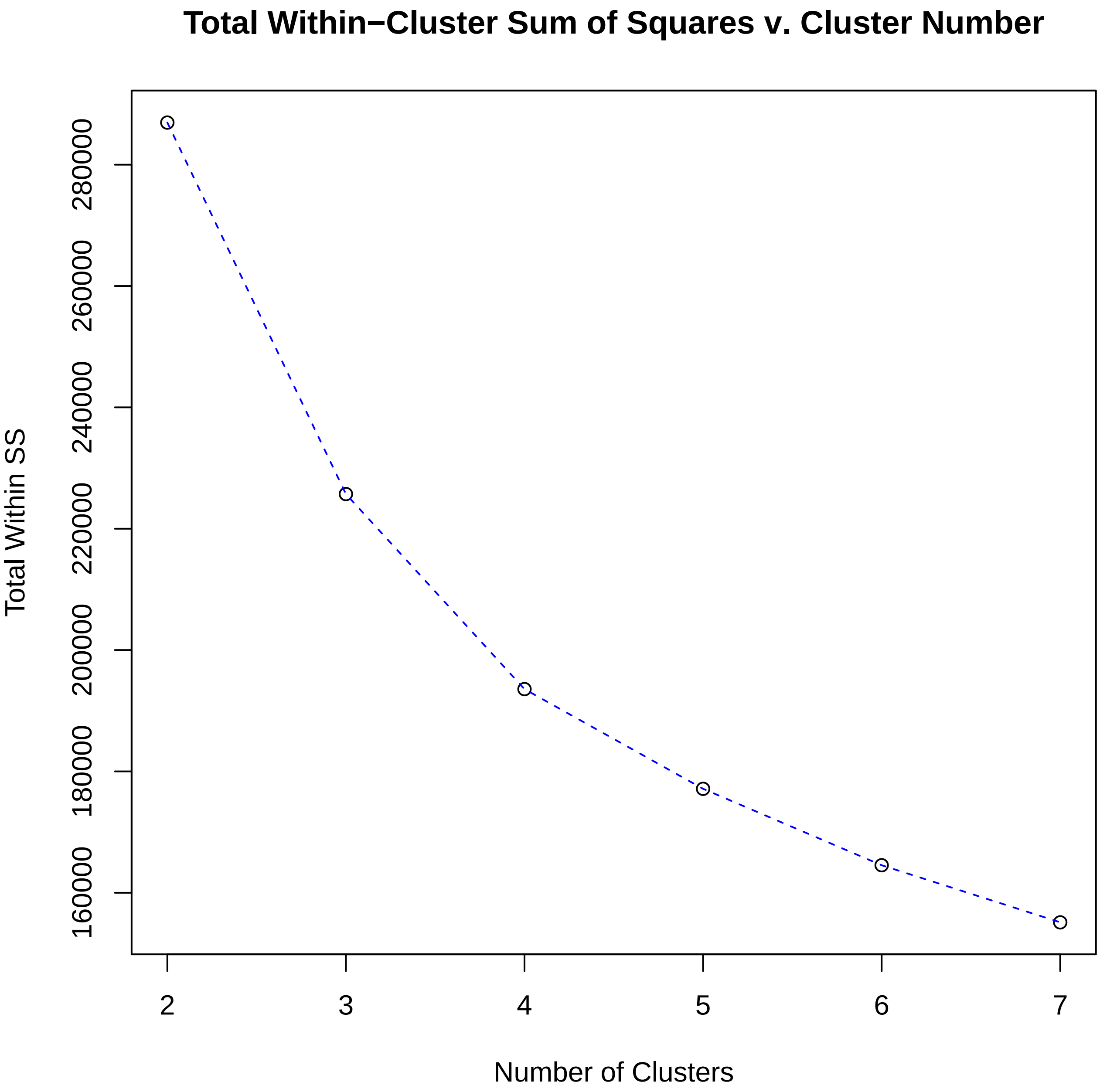


Figure 2: Screeplot From PCA

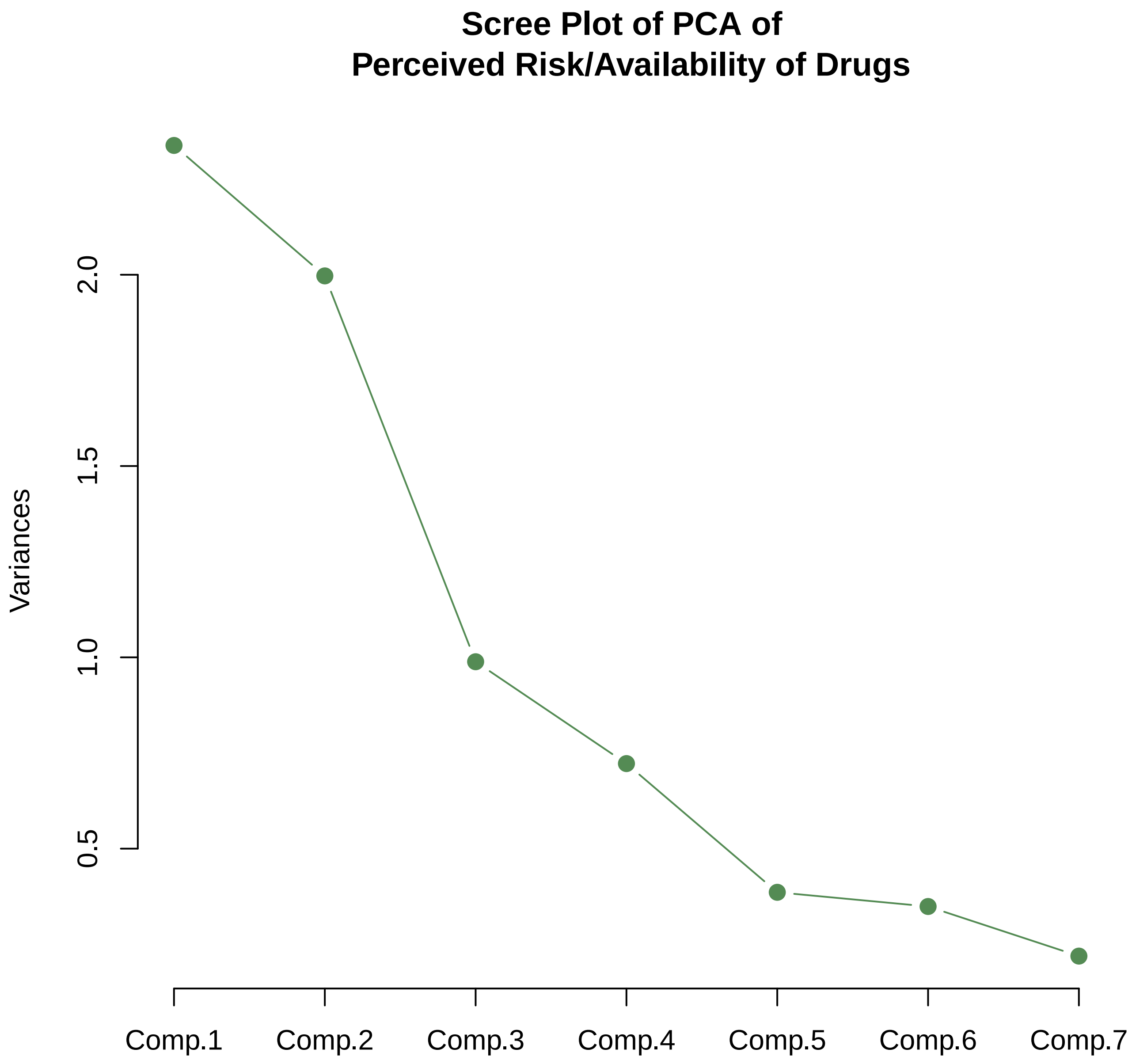


Figure 3: Parallel Plots Analysis

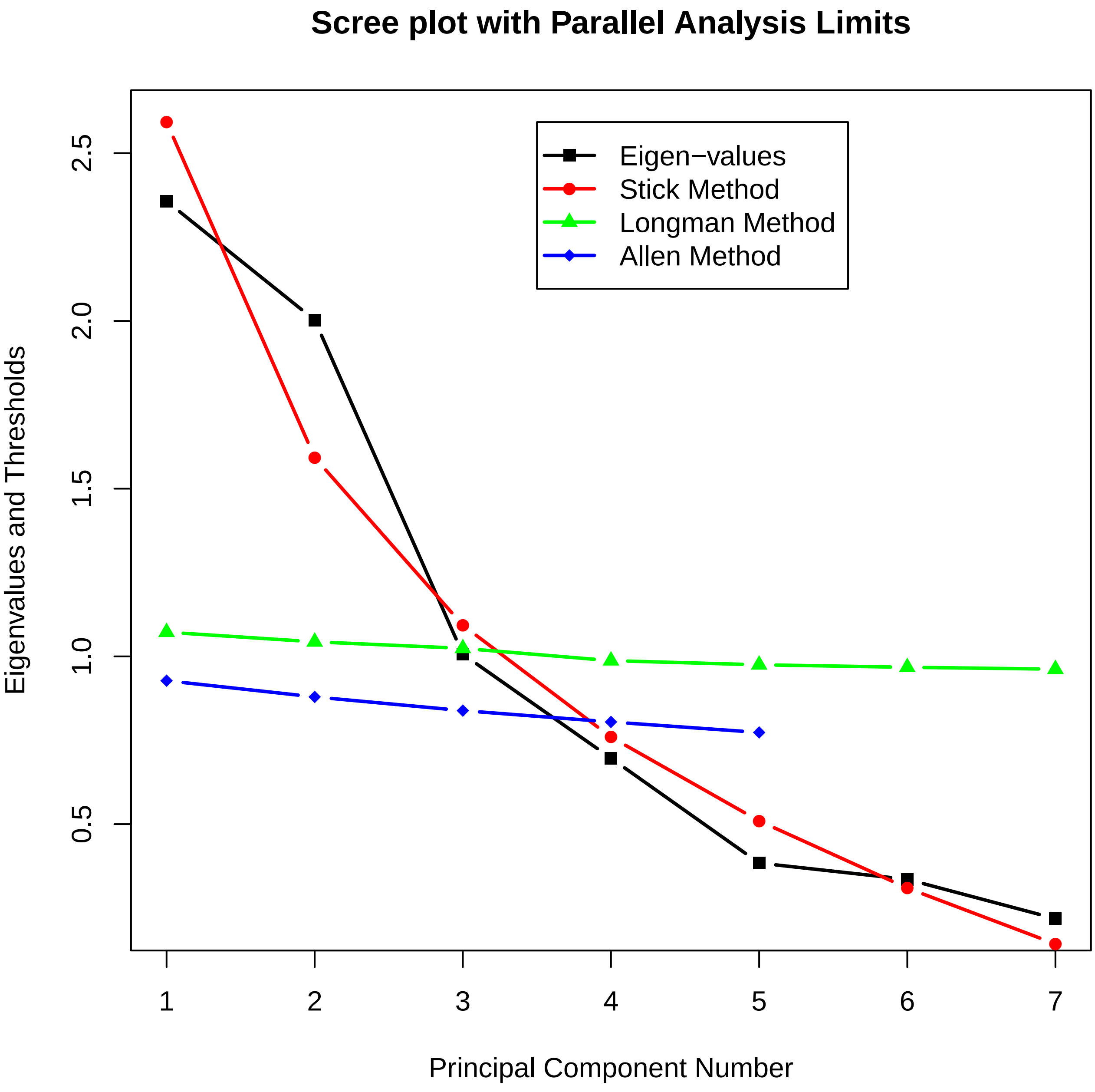
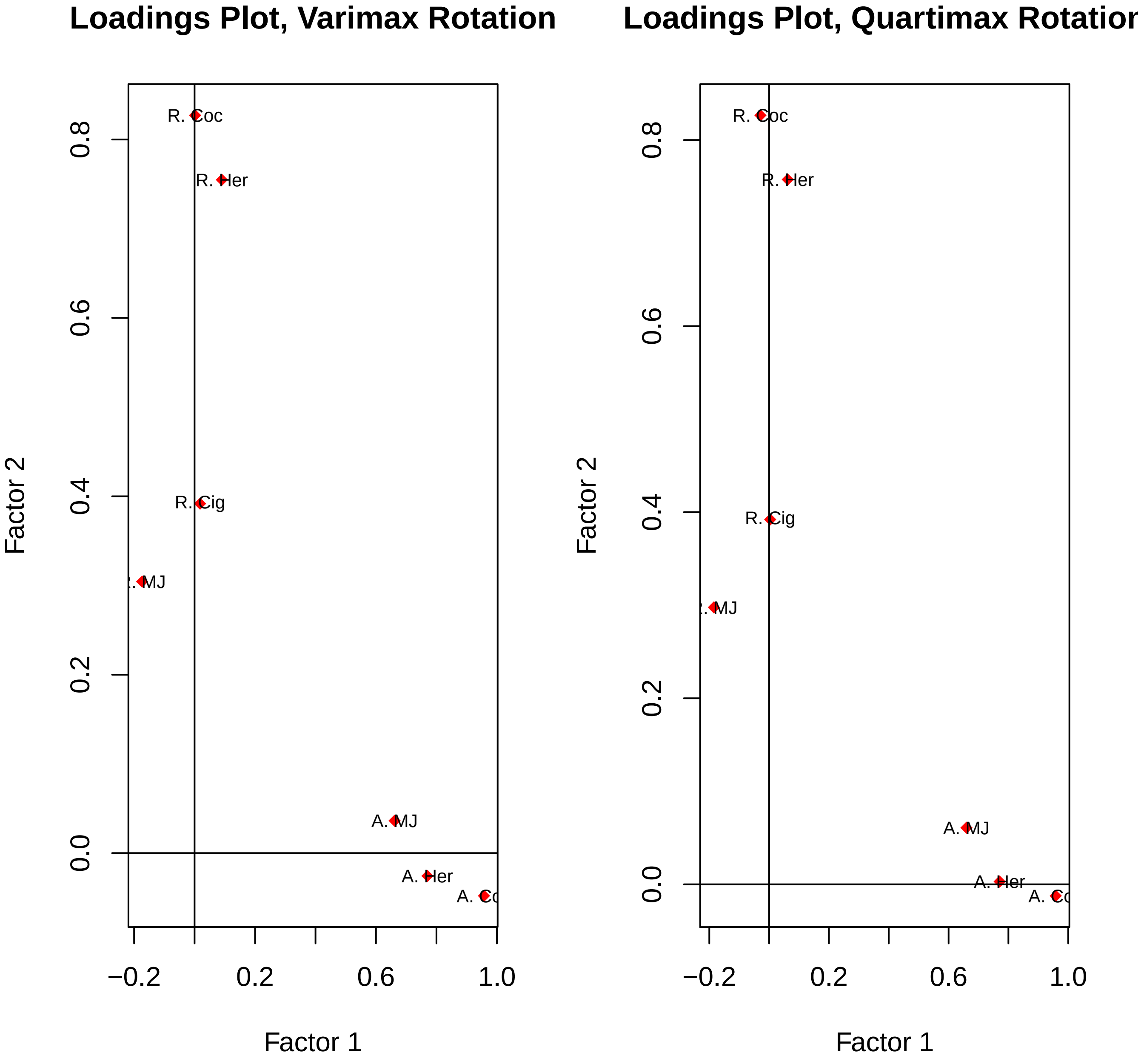


Figure 4: Loadings Plot



List of Tables

Table 1: Centroids

Cluster	Risk Perceptions				Availability		
	Cigarettes	Marijuana	Heroin	Cocaine	Marijuana	Cocaine	Heroin
1	0.05	-0.22	0.25	0.17	0.69	0.83	0.70
2	-0.96	-0.58	-2.46	-2.38	-0.32	-0.11	-0.09
3	0.14	0.33	0.24	0.30	-0.62	-0.80	-0.68

Table 2: Variable Correlations

	R. Cig	R. MJ	R. Her	R. Coc	A. MJ	A. Coc	A. Her
R. Cig	1.00	0.28	0.30	0.30	-0.01	0.00	0.01
R. MJ	0.28	1.00	0.15	0.28	-0.35	-0.17	-0.05
R. Her	0.30	0.15	1.00	0.64	0.13	0.05	0.01
R. Coc	0.30	0.28	0.64	1.00	0.04	-0.04	-0.01
A. MJ	-0.01	-0.35	0.13	0.04	1.00	0.63	0.49
A. Coc	0.00	-0.17	0.05	-0.04	0.63	1.00	0.75
A. Her	0.01	-0.05	0.01	-0.01	0.49	0.75	1.00

Table 3: PCA Loadings

	Load 1	Load 2	Load 3
R. Cig	0.07	-0.44	0.31
R. MJ	0.27	-0.32	0.66
R. Her	0.00	-0.57	-0.40
R. Coc	0.08	-0.59	-0.26
A. MJ	-0.54	-0.09	-0.25
A. Coc	-0.59	-0.09	0.20
A. Her	-0.53	-0.11	0.37

Table 4: Loadings From Factor Analysis

Factor 1	Factor 2
	0.392
-0.174	0.392
	0.755
	0.827
0.662	
0.957	
0.770	

Network Matrix Transformation Visual Basic Code (academic)

```
' This is the Visual Basic code for an Excel macro in this workbook:
' https://www.dropbox.com/s/3h4ymcff4alm4kw/Automated%20Cross-Sum.xlsm?dl=0
' You must run the code inside the workbook for it to function correctly.
' It transforms two-mode matrices into one-mode matrices. For example, if a dataset showed how many
' organizations (column names) individuals (row names) belonged to, it would transform the dataset into
' one that showed how many organizations each pair of individuals both belong to (now individuals are
' both the row and column names). For instance, if the individuals in the third row and fifth row in the
' original dataset both participated in organization in the fourth and sixth columns, then the entry in the
' third row, fifth column of the new matrix would have a 2 in it, because the third listed individual and
' fifth listed individual share two organizations in common.

Attribute VB_Name = "Module1"

Sub CrossSum()
Attribute CrossSum.VB_ProcData.VB_Invoke_Func = "q\n14"


' Get the name of the file containing the Macro
' Useful in case the name of the workbook was changed
Dim filename As String
filename = ThisWorkbook.Name


' Step: Get the Dimensions of the 'i by j' Matrix

' Variables that will hold the number of Rows/Columns
Dim rows As Integer
Dim columns As Integer

' Select the i by j matrix data
Workbooks(filename).Sheets("Input Data -- i by j").Activate
ActiveSheet.Range("C5").Select

' Count the number of Rows/Columns
' Selects the entire matrix and counts the rows/columns
rows = Sheets("Input Data -- i by j").Range("C5").Cells.CurrentRegion.rows.Count - 2
columns = Sheets("Input Data -- i by j").Range("C5").Cells.CurrentRegion.columns.Count - 2


' Step: Create the 'j by i' Matrix (the tranpose)

Worksheets.Add
```

```
ActiveSheet.Name = "j by i"

' Tranpose & Paste the 'i by j' Matrix
Sheets("Input Data -- i by j").Select
Sheets("Input Data -- i by j").Range("C5").Cells.CurrentRegion.Select
    Selection.Copy
    Sheets("j by i").Select
    Range("A3").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:= _
        False, Transpose:=True


' Step: Label the 'i by i' Matrix
' i.e. copy/paste the rows labels in i

Worksheets.Add
ActiveSheet.Name = "i by i"

Sheets("Input Data -- i by j").Select
Sheets("Input Data -- i by j").Range(Cells(5, 2), Cells(5 + rows - 1, 2)).Select
    Selection.Copy
    Sheets("i by i").Select
    Range("B5").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:= _
        False, Transpose:=False
    Range("C4").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:= _
        False, Transpose:=True



' Step: Create the 'i by i' Matrix

' First cell being filled: C5 on Sheet "Final Product Transformed i by i"
' =IF($C5*I$5<>0,$C5+I$5)+IF($D5*I$6<>0,$D5+I$6)+IF($E5*I$7<>0,$E5+I$7)

   i = 0
   Do While i <= (rows - 1)
        j = 0
        Do While j <= (rows - 1)
            k = 0
            Cells(5 + i, 3 + j) = 0
            Do While k <= (columns - 1)
    If (Sheets("Input Data -- i by j").Cells(5 + i, 3 + k) > 0 And Sheets("j by i").Cells(5 + k, 3 + j) > 0) Then
     Cells(5 + i, 3 + j) = Cells(5 + i, 3 + j) + Sheets("Input Data -- i by j").Cells(5 + i, 3 + k) + Sheets("j by i").Cells(5 + k, 3 + j)
    End If
    k = k + 1
   Loop
   j = j + 1
  Loop
  i = i + 1
 Loop



' Create a new Excel file (workbook) for the Cross-summed 'i by i' matrix

Sheets("Input Data -- i by j").Activate
ActiveSheet.Range("A1").Select

Sheets("i by i").Range("C5").Cells.CurrentRegion.Copy


Set NewBook = Workbooks.Add

ActiveSheet.Name = "Cross-Summed i by i matrix"

Range("B4").Select
 Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:= _
  False, Transpose:=False

ActiveSheet.columns(2).Font.Bold = True
ActiveSheet.rows(4).Font.Bold = True

Range("A1") = "Cross-Summed i by i Matrix"
Range("D3") = "i"
Range("A6") = "i"


Application.DisplayAlerts = False
 Workbooks(filename).Sheets("j by i").Delete
 Workbooks(filename).Sheets("i by i").Delete
 Application.DisplayAlerts = True
```

```
ActiveSheet.Range("A1").Select

End Sub
```

Game Theory Project (academic)

Extremist Candidates in a Model with Timing and Uncertainty

Jacob B. Schumaker

Department of Political Science

Yale University

Introduction

In existing formal models of elections, candidates take positions close to the median voter unless they are given both policy preferences and uncertainty (Calvert 1985), candidates are different from each other (Groseclose 2001) or they are deterring entry by opponents (Osborne 2000). I show that candidates can be identical, care only about winning, and face no entry threats by third candidates—yet can still diverge from the median voter.

I demonstrate that candidates do not have to choose the same position by giving them a tradeoff between receiving more information about the median voter and building up a structural advantage on their opponent. This captures the dilemma that politicians face when deciding *when* to announce their candidacy. If they announce early, they have more opportunities to win over voters, to fundraise, to secure endorsements, and to build campaign infrastructure (Stokes 1963; Ansolabehere and Snyder 2000; Groseclose 2001; Kartik and McAfee 2007).[1] If they wait to enter, they can learn more about where voters stand.

I show that if signals are sufficiently more accurate than common knowledge, candidates delay their entry into races and take the position their signal tells them to.[2] Specifically, the following simple condition for this equilibrium emerges:

$$\mathrm{x} - \mathrm{z} > \mathrm{v}$$

This condition says that the difference between probability that a signal of where the median voter is located is accurate (x) and the probability that prior belief about the median voter is accurate (z) must be larger than the advantage of early entry (v). When candidates receive different signals, they adopt different platforms. These positions do not have to be at the median voter.

I. The Model

There are two candidates in a two-stage model. Candidates maximize their probability of winning a primary election. They are uncertain about which of two locations the median voter lies at. In the first stage, candidates can lock themselves into a position or they can wait until the second stage. In the second stage, candidates who have not moved choose a position. If

[1] While it is possible for candidates to lay the groundwork for their campaign before they officially declare their candidacy, this article models the advantage to officially declaring candidacy.

[2] Signals function in the same way in Feddersen and Pesendorfer 1998.

candidates enter in the early stage, they increase their probability of winning. If they enter late, they receive a signal about the location of the median voter.

Candidates can choose to run on either a Traditional platform (T) or an Extremist platform (E). The two choices of platforms represent the options that candidates have in a political primary. Candidates may either appeal to voters who typically control the party or they can appeal to the faction of voters who sometimes take control of the party.

Voters are either Moderate (M) or Radical (R). Voters have symmetric single-peaked preferences and so the winner of the election is the candidate preferred by the median voter (Hotelling 1929). If the median voter is Moderate (M), they prefer Traditional candidates (T). If the median voter is Radical (R), they prefer Extremist candidates (E). If candidates are the same distance from the median voter and neither receives an early entry bonus, they tie and win the election with probably $\frac{1}{2}$. In this two-location model, candidates are equidistant from the median voter when they choose the same position.

Candidates are uncertain of the location of the median voter, but share a common belief about the probability the median voter is moderate when the game begins. They believe the median voter is Moderate (M) with probability z and Radical with probability 1 - z. I consider cases in which a moderate platform is likely to be popular on election day—i.e., when $z \in \left(\frac{1}{2}, 1\right)$.

In both of the two stages of the game, entry is costless. Candidates must enter the race by the end of the game, so if they do not enter in the first stage they must enter in the second. In the first stage of the game, both candidates simultaneously decide whether to choose one of the two platforms or whether to wait until the second stage.

Once a candidate chooses a position, they cannot change that position. A drastic shift in platform would be devastating to a campaign, because it tells voters that the candidate lacks character and does not believe the campaign's message (Kartik and McAfee 2007). This assumption "matches the observed reality" of elections (Bruns 2011).

Stage 1

In the first stage, candidates can either choose a Traditional (T) platform, an Extremist (E) platform, or to wait until the second stage. If they choose either T or E in the first stage, they receive a structural advantage—v—called a valence in the literature (Stokes 1963; Ansolabehere and Snyder 2000; Groseclose 2001; Kartik and McAfee 2007).

Candidates who enter early can spend more time campaigning. They can build up their campaign funds, increase their name recognition, and increase their affinity with voters. Candidates who choose a platform in the first stage do not move in the second stage. They are locked into the position they choose.

If candidates receive v, they increase their probability of winning the election. Specifically:

$$\text{Probability of Winning with Valence} = \frac{\text{Probability of Winning without Valence} + \text{v}}{1\text{+v}}$$

The model will be analyzed for cases when the valence to early entry is positive—i.e. when v>0. Otherwise, candidates would not face a tradeoff between entering early and entering late. Instead, entering late would always be optimal because it would avoid the first stage's zero or negative valence *and* candidates would receive more information about the median voter.

Stage 2

In the second stage, candidates who did not move in the first stage receive a signal and choose a platform according to that signal. They forgo the valence bonus v. Instead, they each receive a private signal s_i about the location of the median voter. Candidates can receive different signals. They are the candidate's personal assessment of the political landscape. Before they receive the signal, their belief about public opinion is common knowledge. By waiting until later in campaign season, they give themselves more time to discover the location of the median voter.

The signal is either that the median is Moderate or Radical. A signal that the median is M is s_i=0. A signal that the median is R is s_i=1. Candidates locate according to their signals– this strategy is referred to as S. Candidates who wait always use the strategy S. If a candidate receives s_i= 0, they choose the platform T in order to please the median voter at M. If a candidate receives s_i= 1, they choose the platform E in order to please the median voter at R.

Signals are accurate with probability x. I assume that signals are not misleading—i.e. that $x \in \left(\frac{1}{2}, 1\right)$. If the median is Moderate candidates receive s_i= 0 with probability x and s_i= 1 with probability 1-x. Conversely, if the median is Radical candidates receive s_i= 0 with probability 1-x and s_i= 1 with probability x. After candidates receive their signals, they use Bayes Rule to update their beliefs about the location of the median.

After the second stage of the game, the primary election occurs. Payoffs are equal to the probabilities of winning after valences have been applied. Candidates choose the strategy (either entering in the first round with T, entering in the first round with E, or waiting and locating at their signal with S) that is a best response to their opponent's strategy. I solve this as a Bayesian Nash Equilibrium game.

II. Equilibria

There is always one equilibrium in this model, except in knife-edge situations[3]. The equilibrium is always symmetric: either both candidates enter early and choose a Traditional platform (T,T) or both candidates enter late and locate at their signal (S,S). If signals are sufficiently more accurate than prior knowledge, candidates wait until the second stage to enter.

In the equilibrium where both candidates wait—(S,S)—candidates are not required to converge *or* to locate at the median. Instead, the only requirement is that they locate where there signal tells them to. If they receive different signals, one candidate will run on an extremist platform while the other runs a traditional platform. Even if they receive the same signal and choose the same platform, if their signals are incorrect they will both choose positions that the median voter disagrees with. In the sections that follow, I prove conditions.

Proposition 1 No candidate ever plays E.[4]

No candidate ever plays E—entering in the first stage with an Extremist platform—because it is strictly dominated by T for all possible parameter values. Candidates will get a higher payoff by choosing the platform that is a priori more likely to win. Even if a candidate faces an opponent who is playing S and therefore may begin an Extremist campaign in the second stage, their expected utility from playing T in response is still larger. Because E is strictly dominated, it does not have to be considered when searching for equilibria.

[3] When T and S yield equal payoffs they are both best-responses to each other and so any combination of the two is an equilibrium. This situation (when x-z = v) is described in Proposition 4.

[4] Proposition 1 is proved in Appendix A.

Proposition 2 When x-z $>$ v, there is one equilibrium—in which candidates wait to enter.[5]

PROOF:

The only equilibrium that exists when x-z $>$ v is both candidates waiting until the second stage to receive their signals and then declare their candidacy (S,S). Substantively, the private signals that candidates receive must be significantly more accurate than their prior belief about the location of the median voter in order for them to wait for a signal. First, I show that a weaker version of this condition (x-z $\geq$ v) must be true for (S,S) to be *an* equilibrium. Then, I show that when the stricter version (x-z $>$ v) is true, it is the *only* equilibrium.

Considering best responses to S shows that changes in parameters have no effect on the payoff to S, but rather only affect the payoff to T. To start, consider the payoff to a candidate who chooses S in response to their opponent choosing S. Candidates are identical and have chosen the same action, therefore they must have the same payoffs. Since payoffs are the chance of winning, they both have an expected utility of $\frac{1}{2}$.

When a candidate plays T in response to S, they win outright whenever their opponent receives a signal that the median is Radical (and runs an Extremist campaign), but the median is in fact Moderate. They tie whenever their opponent receives a signal that the median is Moderate (and runs a Traditional campaign), because they will always be the same distance from the median voter. Considering these scenarios, playing T against S yields a payoff of $\frac{\text{1+2v-x+z}}{\text{2 + 2v}}$.

If Eu_i(S,S)=$\frac{1}{2}$ and Eu_i(T,S)=$\frac{\text{1+2v-(x-z)}}{\text{2 + 2v}}$, then for S to be a best response to itself the following must hold: $\frac{1}{2} \geq \frac{\text{1+2v-(x-z)}}{\text{2 + 2v}}$. Note that only the utility from playing (T,S) depends on the values of parameters. The utility from playing (S,S) remains constant. Simplifying the inequality yields the condition for (S,S) to be an equilibrium: x-z$\geq$ v.

Finding the condition that yields (S,S) as the unique equilibrium requires ruling out all other equilibria. The simplest step is to rule out (S,T), and (T,S)—which are identical because players are interchangeable. Changing the condition from x-z $\geq$ v to x-z $>$ v requires S to be a strictly better response to S than T. This eliminates (T,S) and (S,T), because they involve a candidate playing a T in response to S, which is now strictly optimal.

It turns out that the same condition (x - z $>$ v) that rules out (T,S) and (S,T) also rules out (T,T). Ruling out the equilibrium (T,T) requires finding the condition that makes T a

[5] Payoffs are derived in Appendix B.

suboptimal response to itself. This is true when S is a strictly better response to T than T is to itself. $\mathrm{Eu_i(T,T)}=\frac{1}{2}$ and $\mathrm{Eu_i(S,T)}=\frac{1+x\text{-}z}{2+2v}$. For S to be a strictly better response, $\frac{1}{2}<\frac{1+x\text{-}z}{2+2v}$. This simplifies to x-z> v. Therefore, the condition required for (S,S) to be the only equilibria is x-z > v. This condition makes substantive sense. It says that signals must accurate enough that they outweigh the structural advantage of moving early.

Proposition 3 When x-z < v, there is one equilibrium—in which candidates enter early[6]

PROOF:

The only equilibrium that exists when x-z < v is both candidates entering early with a traditional platform (T,T). Substantively, the advantage to early entry must be larger than the difference in accuracy between signals and prior knowledge. First, I show that a weaker version of this condition (x-z < v) must be true for (T,T) to be *an* equilibrium. Then, I show when the stricter version (x-z < v) is true, it is the *only* equilibrium.

As in the previous section, $\mathrm{Eu_i(T,T)}=\frac{1}{2}$ and $\mathrm{Eu_i(S,T)}=\frac{1+x\text{-}z}{2+2v}$. For T to be in the set of best responses to itself, $\frac{1}{2} \geq \frac{1+x\text{-}z}{2+2v}$. This condition is true when v ≥ x-z. When this condition is true (T,T) is *an* equilibrium.

The other requirement is that no other pair of actions is an equilibrium. As with the equilibrium in the previous section—(S,S)—one requirement for (T,T) to be the only equilibrium is that the inequality for T to be optimal against itself must be strengthened –in this case to v > x-z. Strengthening the condition rules out any pair of actions that includes a candidate entering late in response to an early Traditional candidate—removing both (S,T) and (T,S).

The only remaining equilibrium to rule out is (S,S). For this to be out of equilibrium, $\mathrm{Eu_i(S,S)}< \mathrm{Eu_i(T,S)}$.

This inequality holds when $\mathrm{Eu_i(S,S)}=\frac{1}{2}< \mathrm{Eu_i(T,S)}=\frac{1+2v\text{-}(x\text{-}z)}{2\,+\,2v}$.

Simplifying yields the same condition required to rule out (T,S) and (S,T): v>x-z.

This condition is the converse of the condition required for (S,S) to be the only equilibrium: x-z> v.

[6] Payoffs are derived in Appendix B.

Proposition 4 When x-z= v, any combination of T and S are equilibria.[7]

PROOF:

When the difference in accuracy between signals and prior knowledge is exactly equal to the valence advantage, any combination of the strategies is an equilibrium. Any combination of strategies is an equilibrium when all strategies are best responses to each other. For T and S to both be best responses to T, $\mathrm{Eu_i(T,T)=Eu_i(S,T)}$.

In terms of parameters, this is when $\frac{1}{2}=\frac{\mathrm{1+x\text{-}z}}{\mathrm{2+2v}}$, which is true when x-z= v.

For T and S to both be best responses to S, $\mathrm{Eu_i(S,S)=Eu_i(T,S)}$.

In terms of parameters, this is when $\frac{1}{2}=\frac{\mathrm{1+2v\text{-}(x\text{-}z)}}{\mathrm{2 + 2v}}$, which is true when x-z = v. This implies that (T,T), (S,S), (T,S) and (S,T) are all equilibria when early entry and waiting all yield the same payoffs regardless of what an opponent does.

Proposition 5 There is always at least one equilibrium for all parameter values.

PROOF:

The conditions x-z > v, x-z < v, and x-z = v cover the entire parameter space. One of the three conditions must always be true. If the first is true, there is a unique equilibrium at (T,T). If the second is true, there is a unique equilibrium at (S,S). If the third is true, any combination of undominated strategies is an equilibrium.

Concluding Discussion

Giving candidates the option to receive a signal about the location of the median voter makes it possible for them to be led away from the moderate policy platforms even if they are purely office-motivated. Instead, candidates may appeal to ideologically radical voters. This can occur even when the median voter is not radical. When candidates are uncertain about the location of the median voter, they can make mistakes about what position to take in campaigns. If they are given the option to gain an advantage through early entry, they face a tradeoff between less uncertainty and a structural advantage over their opponent.

[7] Payoffs are derived in Appendix B.

This reflects the decisions that politicians face when declaring their candidacy. They have limited leeway to change their positions and so when they announce a stance on an issue they must keep in mind where public opinion will be as the election progresses. However, politicians often begin their campaigns well before they are legally required to. This makes sense if there is also an advantage to early entry into the race (Stokes 1963; Ansolabehere and Snyder 2000; Groseclose 2001; Kartik and McAfee 2007).

When candidates face this tradeoff, their decision comes down to a simple rule: if the difference in accuracy between signals of the median's location and their prior knowledge of the median's location is larger than the early entry valence, they postpone entry in order to learn more about public opinion. The more confident candidates are in their prior knowledge about the median voter, the more accurate their signals will have to be to make delaying entry optimal. The larger the structural advantage to early entry, the larger the difference in accuracy will have to be to outweigh it. This result is important because it shows that candidates do not need to care about policy, third parties, or to be asymmetric for there to be non-convergence at the median voter. Instead, only a realistic form of uncertainty needs to be introduced for politicians to choose different platforms.

Appendix A: Proofs of Propositions

ASSUMPTIONS:

A1 $z \in \left(\frac{1}{2}, 1\right)$

A2 $x \in \left(\frac{1}{2}, 1\right)$

A3 $v > 0$

PROPOSITION 1: No candidate ever plays E.

PROOF:

Candidates do not enter with an Extremist (E) platform in any equilibrium, because it is strictly dominated by a Traditional (T) platform for all parameter values. Consider the payoffs to E and T against every possible opponent strategy:

Case 1: Opponent plays T

$Eu_i(T,T) = \frac{1}{2}$.

When candidates use identical strategies, their expected utilities must always each be $\frac{1}{2}$. One of the two candidates always wins and the payoff to winning is the probability of winning, which must always sum to 1. In the case of strategies where candidates do not receive signals, they will always be the same distance from the median voter and because ties are determined by a fair coin flip, candidates' expected utilities are $\frac{1}{2}$.

$$Eu_i(E,T) = \frac{(1-z)+v}{1+v+v}$$

When a candidate runs an Extremist campaign against a Moderate campaign, they win whenever the median voter is Radical (before valences are applied). The median voter is Radical with probability (1-z). The valence for entering early is v, but their probability of winning does not increase by v because the other candidate also receives v for entering early and v is also added a second time to the denominator whenever a valence is applied in order to bound the probability of winning at less than 1.

For $Eu_i(T,T)$ to be strictly better than $Eu_i(E,T)$, its expected utility must be strictly larger:

$$\frac{1}{2} > \frac{1-z+v}{1+2v}$$

Simplifying implies $z>\frac{1}{2}$, which is always true by Assumption 1. This makes substantive sense. A Moderate median voter is a priori more likely than a Radical median voter and so locating at the Moderate median voter is the better early strategy.

Case 2: Opponent plays E

$Eu_i(E,E) = \frac{1}{2}$

As with (T,T), candidates who play (E,E) are always the same distance from the median and so they each win with the same probability, $\frac{1}{2}$.

$$Eu_i(T,E) = \frac{z+v}{1+v+v}$$

A candidate running a Traditional platform against an Extremist platform wins whenever the median voter is Moderate, which occurs with probability z. Both candidates receive the valence bonus.

$Eu_i(T,E)$ is strictly larger than $Eu_i(E,E)$ when:

$$\frac{z+v}{1+2v} > \frac{1}{2}$$

Which is true when: $z > \frac{1}{2}$, which is always true by Assumption 1.

Case 3: Opponent plays S

When opponents play S, the probability that they receive each signal must be taken into consideration.

When a candidate runs a Traditional platform against an opponent who waits, they win outright whenever their opponent receives a signal that the median is Radical when it is in fact Moderate and they split the vote whenever the opponent receives a Moderate signal.

$$\text{Eu}_\text{i}(\text{T,S}) = \frac{\frac{\text{Pr(s2=0,w=M)}}{2}+\frac{\text{Pr(s2=0,w=R)}}{2}+\text{Pr(s2=1,w=M)}+\text{ v}}{1+\text{v}}$$

In terms of parameters this is $\frac{\frac{\text{xz}}{2}+\frac{\text{(1-x)(1-z)}}{2}+\text{(1-x)z+ v}}{1+\text{v}}$,

which simplifies to $\frac{1+\text{v-(x-z)}}{2}$.

When a candidate runs an Extremist platform against an opponent who waits, they win outright when the opponent receives a Moderate signal but the median is Radical and they split the vote when the opponent receives a Radical signal:

$$\text{Eu}_\text{i}(\text{E,S}) = \frac{\text{Pr(s2=0,w=R)}+\frac{\text{Pr(s2=1,w=M)}}{2}+\frac{\text{Pr(s2=1,w=R)}}{2}+\text{ v}}{1+\text{v}}$$

In terms of parameters, this is: $\frac{\{\text{(1-x)(1-z)}\}+\frac{\{\text{(1-x)z}\}}{2}+\frac{\{\text{x(1-z)}\}}{2}+\text{ v}}{1+\text{v}}$

which simplifies to: $\frac{\text{2+2v-x-z}}{\text{2+2v}}$.

A Traditional platform is a strictly better response to S than an Extremist platform when $\text{Eu}_\text{i}(\text{T,S}) > \text{Eu}_\text{i}(\text{E,S})$:

$\frac{1+2v-x+z}{2+2v} > \frac{2+2v-x-z}{2+2v}$

As before, this inequality simplifies to z > $\frac{1}{2}$.

- In summary, T yields strictly higher expected utilities than E against all possible opponent strategies and so T strictly dominates E.

Appendix B: Derivation of Payoffs for Propositions 2-4

PROPOSITION 2 When x-z > v, there is one equilibrium—in which candidates wait to enter.

The proof of Proposition 2 is shown in the body of the article. Here I show how the payoffs for the proofs were derived.

$Eu_i(S,S)=\frac{1}{2}$

The expected utility from playing S against itself must be $\frac{1}{2}$. Although candidates will not always be the same distance from the median (if they receive different signals they will choose different platforms), they must still have the same expected utility because they are playing the same action against itself. Payoffs must sum to 1 and the only pair of payoffs that are equal and sum to 1 is $\frac{1}{2}$,$\frac{1}{2}$.

$Eu_i(T,S)=\frac{1+2v-(x-z)}{2 + 2v}$

The derivation of $Eu_i(T,S)$ is shown in Case 3 of the proof of Proposition 1.

$Eu_i(T,T)=\frac{1}{2}$

The derivation of $Eu_i(T,T)$ is shown in Case 1 of the proof of Proposition 1.

$\mathrm{Eu_i(S,T)} = \frac{1+x-z}{2+2v}$

When a candidate waits to enter against a Traditional opponent, they win outright when they receive a signal that the median is Radical and the median is Radical, while they split the vote when they receive a signal that the median is Moderate.

$$\mathrm{Eu_i(S,T)} = \frac{\frac{\Pr(s1=0)}{2} + \Pr(w=R|s1=1)\Pr(s1=1)}{1+v}$$

In terms of parameters, $\frac{\left[\frac{1}{2} * \{xz+(1-x)(1-z)\} + \left\{\frac{x(1-z)}{(1-x)z+x(1-z)}\right\} * \{(1-x)z+x(1-z)\}\right]}{1+v}$,

which simplifies to $\frac{1+x-z}{2+2v}$.

PROPOSITION 3 When x-z < v, there is one equilibrium—in which candidates enter early.

As with Proposition 2, the proof of Proposition 3 is shown in the body of the article. References to the derivations of payoffs are given here.

$\mathrm{Eu_i(T,T)}$

The derivation of $\mathrm{Eu_i(T,T)}$ is shown in Case 1 of the proof of Proposition 1.

$\mathrm{Eu_i(S,T)}$

The derivation of $\mathrm{Eu_i(S,T)}$ is shown in the derivations for Proposition 2.

$\mathrm{Eu_i(S,S)}$

The proof for $\mathrm{Eu_i(S,S)}$ is shown in the derivations for Proposition 2.

$\mathrm{Eu_i(T,S)}$

The derivation of $\mathrm{Eu_i(T,S)}$ is shown in Case 3 of the proof of Proposition 1.

PROPOSITION 4 When x-z = v, any combination of T and S are equilibria.

The proof of Proposition 4 is shown in the body of the article. References to the derivations of payoffs are given here.

Eu_i(T,T)

The derivation of Eu_i(T,T) is shown in Case 1 of the proof of Proposition 1.

Eu_i(S,T)

The derivation of Eu_i(S,T) is shown in the derivations for Proposition 2.

Eu_i(S,S)

The proof for Eu_i(S,S) is shown in the derivations for Proposition 2.

Eu_i(T,S)

The derivation of Eu_i(T,S) is shown in Case 3 of the proof of Proposition 1.

Bibliography

Ansolabehere, Stephen, and J.M. Snyder. 2000. “Valence Politics and Equilibrium in Spatial Election Models.” *Public Choice* 103(3): 327–336. http://www.springerlink.com/index/J3627821WH0R3G21.pdf (Accessed December 14, 2011).

Bruns, Richard. 2011. “Fringe Candidates in Electoral Competition.” Working Paper.

Calvert, RL. 1985. “Robustness of the Multidimensional Voting Model: Candidate Motivations, Uncertainty, and Convergence.” *American Journal of Political Science* 29(1): 69-95. http://www.jstor.org/stable/10.2307/2111212 (Accessed December 14, 2011).

Feddersen, Timothy, and Wolfgang Pesendorfer. 1998. “Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting.” *The American Political Science Review* 92(1): 23. http://www.jstor.org/stable/2585926?origin=crossref.

Groseclose, Tim. 2001. “A Model of Candidate Location When One Candidate Has a Valence Advantage.” *American Journal of Political Science* 45(4): 862–886. http://www.jstor.org/stable/10.2307/2669329 (Accessed December 14, 2011).

Hotelling, Harold. 1929. “Stability in Competition.” *The Economic Journal* 39(153): 41–57. http://www.jstor.org/stable/10.2307/2224214 (Accessed December 15, 2011).

Kartik, Navin, and R. Preston McAfee. 2007. “Signaling Character in Electoral Competition.” *American Economic Review* 97(3): 852-870. http://pubs.aeaweb.org/doi/abs/10.1257/aer.97.3.852.

Osborne, M. 2000. “Entry-deterring policy differentiation by electoral candidates.” *Mathematical Social Sciences* 40(1): 41-62. http://linkinghub.elsevier.com/retrieve/pii/S0165489699000402.

Stokes, D.E. 1963. “Spatial Models of Party Competition.” *The American Political Science Review* 57(2): 368–377. http://www.jstor.org/stable/10.2307/1952828 (Accessed December 14, 2011).

Basketball Analytics Projects

Project 1: Tanking

There is widespread concern that tanking is a problem in the NBA. Teams are accused of attempting to lose games in order to increase their chances of landing a top pick in the draft. However, this is not necessarily the case. Teams may actually stand a better chance of receiving the #1 pick if they improve.

Take the following example: Last year the Celtics finished with the 5th worst record in the NBA. Whether it's two years from now or twenty years from now, what pick are the Celtics more likely to land first, the #1 pick or the #5 pick?
Assume that the Celtics are likely to improve and every year the team will either:
a. Improve by 3 positions (e.g. they go from 5th worst to 8th worst) with a probability of 0.6
b. Or fall 2 positions (e.g. they go from 5th worst to 3rd worst) with a probability of 0.4.
Further assume that the lottery odds and structure remain the same going forward.

In this setup, the Celtics are more likely to land the #1 pick first than the #5 pick. By my calculations, they have a 60.7% chance of landing the #1 pick first. I arrived at this calculation by simulating the draft pick assignment process in Mathematica. Before I started the simulation, my intuition was that the Celtics would be more likely to receive the #1 pick first, because it is only possible to receive the #5 pick from a few positions in the lottery (the 2nd through 5th positions) and in this scenario the Celtics are likely to move farther and farther away from these positions as they tend to improve. I chose to use a simulation despite my intuition because random walks such as the one in this problem can sometimes exhibit counterintuitive behavior.

The annotated Mathematica code for the simulation is below. The code repeatedly executes the draft process described in the question prompt. The probabilities of receiving a pick from each draft position used in the simulation are rounded to three decimal places because I was not able to find probabilities to the full four decimal places.[8] However, I am confident that this does not affect the qualitative outcome that the Celtics are more likely to receive the first pick.

Annotated Mathematica Code[9]:

[8] http://en.wikipedia.org/wiki/NBA_draft_lottery#Process

[9] To run the simulation, use the un-annotated code found below the annotated code.

```
# Initialize a counter for the number of times the team gets the 1st pick

firstpickcount=0;

# Begin a do loop that repeats 100,000 times

Do[

# Initialize the variables to their starting values for each run of the simulation.

# ‘reached’ indicates whether the team has received the 1st or 5th pick

reached=0;

# ‘position’ is the team’s position in the lottery
# The Celtics start every run in 5th position

position=5;

# ‘change’ is how much the team’s draft position changes from the previous year

change=0;

# Begin a while loop that repeats while ‘reached’ equals zero

While[reached==0,

# Inside the loop, determine how much the team’s record changes by from the previous
# year.

# The ‘RandomChoice’ function randomly chooses between the numbers 5 and 0 with the
# probabilities 0.6 and 0.4 respectively, then subtracts 2 from the choice so the change is
# either +3 or -2.

change=RandomChoice[{.6,.4}->{5,0}]-2;

# The team’s draft position changes by the amount determined
```

```
position = position + change;
position = If[position > 30, 30, position];
position = If[position < 1, 1, position];

# The team's draft pick is determined with the probabilities that correspond to their
# draft position through a series of nested if statements

pick = If[position == 1,
   RandomChoice[{.25, .215, .178, .357} -> {1, 2, 3, 4}],
   If[position == 2,
    RandomChoice[{.199, .188, .171, .319, .123} -> {1, 2, 3, 4, 5}],
    If[position == 3,
     RandomChoice[{0.156, 0.1573, 0.1556, 0.2253, 0.2653,
        0.0405} -> {1, 2, 3, 4, 5, 6}],
     If[position == 4,
      RandomChoice[{0.119, 0.126, 0.133, 0.099, 0.351, 0.16,
         0.012} -> {1, 2, 3, 4, 5, 6, 7}],
      If[position == 5,
       RandomChoice[{0.088, 0.097, 0.107, 0.261, 0.36, 0.084,
          0.004} -> {1, 2, 3, 5, 6, 7, 8}],
       If[position == 6,
        RandomChoice[{0.063, 0.071, 0.081, 0.439, 0.305, 0.04,
           0.001} -> {1, 2, 3, 6, 7, 8, 9}],
        If[position == 7,
         RandomChoice[{0.043, 0.049, 0.058, 0.599, 0.232, 0.018,
            0} -> {1, 2, 3, 7, 8, 9, 10}],
         If[position == 8,
          RandomChoice[{0.028, 0.033, 0.039, 0.724, 0.168, 0.008,
             0} -> {1, 2, 3, 8, 9, 10, 11}],
          If[position == 9,
           RandomChoice[{0.017, 0.02, 0.024, 0.813, 0.122, 0.004,
              0} -> {1, 2, 3, 9, 10, 11, 12}],
           If[position == 10,
            RandomChoice[{0.011, 0.013, 0.016, 0.87, 0.089, 0.002,
               0} -> {1, 2, 3, 10, 11, 12, 13}],
            If[position == 11,
             RandomChoice[{0.008, 0.009, 0.012, 0.907, 0.063, 0.001,
                0} -> {1, 2, 3, 11, 12, 13, 14}],
            If[position == 12,
             RandomChoice[{0.007, 0.008, 0.01, 0.935, 0.039,
                0} -> {1, 2, 3, 12, 13, 14}],
             If[position == 13,
              RandomChoice[{0.006, 0.007, 0.009, 0.96, 0.018} -> {1,
                 2, 3, 13, 14}],
              If[position == 14,
               RandomChoice[{0.005, 0.006, 0.007, 0.982} -> {1, 2, 3,
                   14}], pick = 1000]]]]]]]]]]]]]];

# 'reached' is set equal to 1 if the team gets the 1st or 5th pick, or else it remains 0

reached=If[pick==1||pick==5,1,0]];

# End of while loop

# The counter for the number of times the team received the first pick increases by 1 if
# the team received the first pick.

firstpickcount=If[pick==1,firstpickcount+1,firstpickcount],{100000}]

# End of do loop

# Display the number of times the team received the first pick

Print[firstpickcount]
```

End of Mathematica code

Unannotated Mathematica Code (paste and run)

```
firstpickcount = 0;
Do[
 reached = 0;
 position = 5;
 change = 0;
 While[reached == 0,
  change = RandomChoice[{.6, .4} -> {5, 0}] - 2;
```

```
position = position + change;
position = If[position > 30, 30, position];
position = If[position < 1, 1, position];
pick = If[position == 1,
 RandomChoice[{.25, .215, .178, .357} -> {1, 2, 3, 4}],
 If[position == 2,
  RandomChoice[{.199, .188, .171, .319, .123} -> {1, 2, 3, 4, 5}],
  If[position == 3,
   RandomChoice[{0.156, 0.1573, 0.1556, 0.2253, 0.2653,
      0.0405} -> {1, 2, 3, 4, 5, 6}],
   If[position == 4,
    RandomChoice[{0.119, 0.126, 0.133, 0.099, 0.351, 0.16,
       0.012} -> {1, 2, 3, 4, 5, 6, 7}],
    If[position == 5,
     RandomChoice[{0.088, 0.097, 0.107, 0.261, 0.36, 0.084,
        0.004} -> {1, 2, 3, 5, 6, 7, 8}],
     If[position == 6,
      RandomChoice[{0.063, 0.071, 0.081, 0.439, 0.305, 0.04,
         0.001} -> {1, 2, 3, 6, 7, 8, 9}],
      If[position == 7,
       RandomChoice[{0.043, 0.049, 0.058, 0.599, 0.232, 0.018, 0} ->
        {1, 2, 3, 7, 8, 9, 10}],

       If[position == 8,
        RandomChoice[{0.028, 0.033, 0.039, 0.724, 0.168, 0.008,
           0} -> {1, 2, 3, 8, 9, 10, 11}],

        If[position == 9,
         RandomChoice[{0.017, 0.02, 0.024, 0.813, 0.122, 0.004,
            0} -> {1, 2, 3, 9, 10, 11, 12}],

         If[position == 10,
          RandomChoice[{0.011, 0.013, 0.016, 0.87, 0.089, 0.002,
             0} -> {1, 2, 3, 10, 11, 12, 13}],

          If[position == 11,
           RandomChoice[{0.008, 0.009, 0.012, 0.907, 0.063, 0.001,
              0} -> {1, 2, 3, 11, 12, 13, 14}],
             If[position == 12,
              RandomChoice[{0.007, 0.008, 0.01, 0.935, 0.039,
                 0} -> {1, 2, 3, 12, 13, 14}],

              If[position == 13,
               RandomChoice[{0.006, 0.007, 0.009, 0.96, 0.018} -> {1,
                   2, 3, 13, 14}],

               If[position == 14,
                RandomChoice[{0.005, 0.006, 0.007, 0.982} -> {1, 2,
                   3, 14}], pick = 1000]]]]]]]]]]]]]];
  reached = If[pick == 1 || pick == 5, 1, 0]];
 firstpickcount =
  If[pick == 1, firstpickcount + 1, firstpickcount], {100000}]
Print[firstpickcount]
```

End of Unannotated Mathematica Code

Draft Pick Probabilities (rounded to 3 decimal places)[10]:

Seed	Chances	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th	14th
1	250	.250	.215	.178	.357										
2	199	.199	.188	.171	.319	.123									
3	156	.156	.157	.156	.226	.265	.040								
4	119	.119	.126	.133	.099	.351	.160	.012							
5	88	.088	.097	.107		.261	.360	.084	.004						
6	63	.063	.071	.081			.439	.305	.040	.001					
7	43	.043	.049	.058				.599	.232	.018	.000				
8	28	.028	.033	.039					.724	.168	.008	.000			
9	17	.017	.020	.024						.813	.122	.004	.000		
10	11	.011	.013	.016							.870	.089	.002	.000	
11	8	.008	.009	.012								.907	.063	.001	.000

[10] http://en.wikipedia.org/wiki/NBA_draft_lottery#Process

12	7	.007	.008	.010									.935	.039	.000
13	6	.006	.007	.009										.960	.018
14	5	.005	.006	.007											.982

Project 2: Offensive Rebounding

There is a debate in basketball over offensive rebounding. Some experts believe that offensive rebounds are overrated and that to be successful teams should focus on getting back on defense rather than crashing the boards. Others believe that offensive rebounding is important because it creates extra possessions for the team and can be valuable even if it leads to some fastbreaks for the opponent. How many more games should a team expect to win if its offensive rebound rate increases a small amount, for example, from 25% to 27%, all else equal?

The following terms are used in this analysis:

Pos: Possessions Per Game
OReb: Offensive Rebound
$Pos_{\ ORR(0)}$: Possessions Per Game if a team's Offensive Rebounding Rate is zero
$Pos_{\ ORR(0)}\ = Actual\ Possessions\ Per\ Game - Offensive\ Rebounds\ Per\ Game$
$Pos_{\ ORR(\cdot)}$: Possessions Per Game if a team's Offensive Rebounding Rate is $\Pr(OReb|Miss)$
$Pos_{\ ORR(.25)}$: Possessions Per Game if a team's Offensive Rebounding Rate is 0.25
$Pos_{\ ORR(.27)}$: Possessions Per Game if a team's Offensive Rebounding Rate is 0.27
$\Pr(Miss|Pos)$: the probability a missed shot will occur on a possession
$\Pr(OReb|Miss)$: the probability an offensive rebound will occur given a missed shot has occurred

To calculate the expected number of wins a team would gain by increasing their offensive rebounding rate, I first derive the expected number of possessions the team has for any offensive rebounding rate. To do that, I calculate how many possessions a team adds through offensive rebounds. I model this as the sum of an infinite series, in which a team's offensive rebounding rate helps them increase not just the number of consecutive possessions in which they miss a shot but retain the ball through an offensive rebound, but also helps them increase the number of times they have three, four, five, etc. possessions in a row by continuing to get offensive rebounds on missed shots. This may be an upper bound on the number of possessions a team can expect to have, as it does not account for the fact that games have finite length and that eventually the clock will run out even if the team continues to get offensive rebounds.

$Pos_{\ ORR(\cdot)} =$	
$Pos_{ORR(0)} +$	Pos w/ 0 OReb
$Pos_{ORR(0)}\ \Pr(Miss\vert Pos)\ \Pr(OReb\vert Miss) +$	1 OReb
$Pos_{ORR(0)}\ \Pr(Miss\vert Pos)\ \Pr(Oreb\vert Miss)\ \Pr(Miss\vert Poss)\ \Pr(Oreb\vert Miss) +$	2 OReb
$Pos_{ORR(0)}\ \Pr(Miss\vert Pos)\ \Pr(Oreb\vert Miss)\ \Pr(Miss\vert Poss)$ $\Pr(Oreb\vert Miss)\ \Pr(Miss\vert Poss)\ \Pr(Oreb\vert Miss) +$	3 OReb
… +	
$Pos_{ORR(0)}\ [\Pr(Miss\vert Pos)\ \Pr(OReb\vert Miss)]^{\infty}$	∞ OReb

where “x OReb” means “x possessions in a row in which the team retained the ball because of an offensive rebound.”

Every subsequent element in this infinite series multiplies the previous element by $\Pr(Miss|Pos)\ \Pr(OReb|Miss)$, so we can simplify every element to $Pos_{ORR(0)}$ multiplied by $\Pr(Miss|Pos)\ \Pr(OReb|Miss)$ raised to the exponent corresponding to its position in the series:

$Pos_{\ ORR(\cdot)} =$	
$Pos_{ORR(0)} +$	Possessions w/ 0 OReb
$Pos_{ORR(0)}\ \Pr(Miss\vert Pos)\ \Pr(OReb\vert Miss) +$	1 OReb
$Pos_{ORR(0)}[\Pr(Miss\vert Pos)\ \Pr(Oreb\vert Miss)]^2 +$	2 OReb
$Pos_{ORR(0)}[\Pr(Miss\vert Pos)\ \Pr(Oreb\vert Miss)]^3 +$	3 OReb
… +	
$Pos_{ORR(0)}\ [\Pr(Miss\vert Pos)\ \Pr(OReb\vert Miss)]^{\infty}$	∞ OReb

We can factor out $Pos_{ORR(0)}$ from every element in this expression and refer to the term $\Pr(Miss|Pos)\ \Pr(Oreb|Miss)$ with δ, making the expression:

$$Pos_{\ ORR(\cdot)} = Pos_{ORR(0)} * (1 + \delta + \delta^2 + \delta^3 + \ldots + \delta^{\infty})$$

The sum of an infinite series in the form $1 + \delta + \delta^2 + \delta^3 + \cdots + \delta^{\infty}$ simplifies to $\frac{1}{1-\delta}$, so the above express reduces to $Pos_{\ ORR(\cdot)} = Pos_{ORR(0)}\frac{1}{1-\delta}$.[11] Plugging in $\Pr(Miss|Pos)\ \Pr(Oreb|Miss)$ for δ we get $Pos_{\ ORR(\cdot)} = Pos_{ORR(0)}\frac{1}{1-\Pr(Miss|Pos)\ \Pr(OReb|Miss)}$.

[11] As long as $\delta < 1$. Source: http://sites.duke.edu/niou/files/2011/05/Lecture-5-Repeated-Games.pdf, p.5

How do we get from this general formula to how many more games a team will win if their offensive rebounding rate increases from .25 to .27? First, we need to calculate their $Pos_{\ ORR(0)}$, the number of possessions they would have had without any offensive rebounds. This is calculated by taking the actual number of possessions the team used per game and subtracting their offensive rebounds per game:

$$Pos_{\ ORR(0)} \ = Actual\ Possessions\ Per\ Game - Offensive\ Rebounds\ Per\ Game$$

The idea is that if the team's offensive rebounding rate were zero they would have gotten no offensive rebounds and so they would have lost every possession they acquired from an offensive rebound.

This is an approximation of how many possessions they would have, as reducing the number of offensive rebounds to zero could affect the number of possessions in other ways even if the probability of events remained constant. For example, the time that elapsed during the possessions that followed offensive rebounds would not disappear if the offensive rebound did not occur. Instead, that time would be split between possessions for the team and its opponent. For this reason, the number of possessions given an offensive rebounding rate may be more like a lower bound on the number rather than the true average value. It is not immediately clear whether this lower bound effect countervails the upper bound effect discussed earlier.

Next, we need $\Pr(Miss|Pos)$, the probability a missed shot occurs on a possession. This is different than the reverse of a team's field goal percentage. Instead, it is the probability that the event 'missed shot' actually happens on a possession. This is calculated by taking the probability that field goal occurs on a possession and multiplying that by the probability a team misses when they take a field goal attempt:

$$\Pr(Miss|Pos) = \Pr(Field\ Goal\ Attempt|Possession)\Pr(Miss|Field\ Goal\ Attempt)$$

$$= \frac{Field\ Goal\ Attempts}{Possessions} * \frac{Field\ Goal\ Misses}{Field\ Goal\ Attempts}$$

$$= \frac{Field\ Goal\ Misses}{Possessions}$$

With these expressions substituted into the equation we have

$$Pos_{\ ORR(\cdot)} = (Possessions - ORebs)\frac{1}{1 - \frac{Field\ Goal\ Misses}{Possessions}\Pr(OReb|Miss)}$$

For $(Possessions - ORebs)$ and $\frac{Field\ Goal\ Misses}{Possessions}$ the average values for the 2013-2014 NBA season are used. I checked that this is equivalent to carrying out the calculation for every team and then averaging those results to get the average effect. The average $(Possessions - ORebs)$ is 84.7. The average $\frac{Field\ Goal\ Misses}{Possessions}$ is 0.4701145954.

Substituting these values, we get:

$$Pos_{\ ORR(\cdot)} = 84.7\frac{1}{1 - 0.4701145954 * \Pr(OReb|Miss)}$$

To get the team's expected number of possessions, all that remains is to substitute the offensive rebounding rates. For an offensive rebounding rate of 0.25, we have:

$$Pos_{\ ORR(.25)} = 85.42\frac{1}{1 - 0.4701145954 * 0.25} = 96.79\ (rounded\ to\ two\ decimal\ places)$$

For an offensive rebounding rate of 0.27, we have:

$$Pos_{\ ORR(.27)} = 85.42\frac{1}{1 - 0.4701145954 * 0.27} = 97.83\ (rounded\ to\ two\ decimal\ places)$$

The next step is to go from possessions to points. We do that by making the assumption that every possession is worth the same number of points, although this may not be true. Again, we use the average NBA value, which is 1.04 points per possession. Multiplying that by the number of possessions for each rebound rate, we get:

$$Points\ Per\ Game_{\ ORR(.25)} = 1.04 * 96.79 = 100.67$$

$$Points\ Per\ Game_{\ ORR(.27)} = 1.04 * 97.83 = 101.76$$

With points per game, we can calculate the expected winning percentage and the expected number of wins in a season. To calculate the winning percentage, the Pythagorean expected winning percentage method adapted to basketball by John Hollinger (http://espn.go.com/nba/stats/rpi) is used. The formula for winning percentage is:

$$Expected\ Winning\ \% = \frac{Points\ Per\ Game^{16.5}}{Points\ Per\ Game^{16.5} + Points\ Per\ Game\ Allowed^{16.5}}$$

The average points per game allowed in the NBA in the 2013-14 NBA season was 101.01.

Using the points per game for the two offensive rebounding rates, we get:

$$Expected\ Winning\ \%_{.25} = \frac{100.67^{\ 16.5}}{100.67^{\ 16.5} + 101.01^{16.5}} = 0.49$$

$$Expected\ Winning\ \%_{.27} = \frac{100.67^{\ 16.5}}{100.67^{\ 16.5} + 101.01^{16.5}} = 0.53$$

Over and 82 game season, this corresponds to:

$Expected\, Wins_{.25} = 0.49 * 82 = 40.1$

$Expected\, Wins_{.27} = 0.53 * 82 = 43.3$

Subtracting $Expected\, Wins_{.27} - Expected\, Wins_{.25}$, we get our final answer: a team should expect to win 3.2 more games if their offensive rebounding rate increases from 0.25 to 0.27.