After months of vitriolic debate, sweeping legislation changing the very nature of health care in the United States passed through Congress and was signed into law by President Obama in March of 2010. As health care begins to fade from the limelight, the twenty-first Journal of Politics & Society seeks to bring to your attention other prominent policy issues, pushing boundaries, borders, and coastlines. The 2010 Journal continues its mission of being a truly global publication, providing insight into a kaleidoscope of issues while still remaining relevant to those sitting on the steps of Columbia University in the City of New York. As world leaders pass often through Columbia’s gates to speak to undergraduates, we take this opportunity to provide a forum for this attentive audience to let their views be known to the world.

Two prominent thinkers contribute to this relevant conversation, discussing the rising importance of those nations beyond the Atlantic as well as the erratic global economic environment. Dr. Shashi Tharoor speaks of the new “Neighborhood of an Ocean.” Expounding on the potential and power of a new and unique non-Western regional organization united by the Indian Ocean, he writes now as India’s Minister of External Affairs, reflecting as well on his experience as UN Under-Secretary-General. Commenting on the persistent and painful issue of the past few years, Dr. Dominick Salvatore explains the causes and effects of the continuing financial crisis.

But it is our eight excellent undergraduate authors that make this discussion robust. This year’s recipient of the Peter and Katherine Tomassi Prize is José Martinez of Williams College, who examines elite power structures of the Shiite Ulama in Iran preceding the 1979 revolution. Martinez challenges conventional notions of Islamic extremism in Iran, and suggest approaching negotiations with an entirely new perspective.

Connecting the past with the present, Nishi Kumar considers the use of new social media by dissidents in Iran’s recent election and its impact on mobilization and human rights. Idon Natanzon's
paper on Israeli Unilateralism is particularly timely as the United States challenges Israel on its construction of settlements in East Jerusalem. As Kyrgyzstan experiences mass protests and upheaval, Grace Zhou provides a linguistic perspective to our thoughts by writing about the role of language in Central-Asian nation-building. Of course, no dialogue would be complete without reference to China, and Brian Chao provides a detailed analysis on U.S-China naval competition. Andrei Stetsenko takes us back across the ocean to describe the need for a coordinated transatlantic energy security agreement in order to deny Russia its growing monopoly power. Bill DeBaun anchors us home, tackling crucial domestic concerns by examining the distribution of U.S. education aid. Reminding us of our neighbors down South, Sergio Broholm explores NAFTA’s problematic impact on Mexican corn farmers.

The Twenty-First Editorial Board has chosen scholarship that provides readers with a nuanced understanding of those issues that frequent our daily headlines: China, energy, Iran, education, Central Asia, India, the economic recession. The essays’ novel contributions equip us with the complex perspectives necessary to understand world events and the growing number of relationships between countries.

This Journal is the product of the diligence of the Helvidius Group’s excellent staff, and it has been a true privilege to work with them over the past four years. I hope that you enjoy this twenty-first edition and keep the Journal in mind as you engage with the relevant issues of our time.

Anusha Sriram
Editor-in-Chief

New York City
April 2010
paper on Israeli Unilateralism is particularly timely as the United States challenges Israel on its construction of settlements in East Jerusalem. As Kyrgyzstan experiences mass protests and upheaval, Grace Zhou provides a linguistic perspective to our thoughts by writing about the role of language in Central-Asian nation-building. Of course, no dialogue would be complete without reference to China, and Brian Chao provides a detailed analysis on U.S-China naval competition. Andrei Stetsenko takes us back across the ocean to describe the need for a coordinated transatlantic energy security agreement in order to deny Russia its growing monopoly power. Bill DeBaun anchors us home, tackling crucial domestic concerns by examining the distribution of U.S. education aid. Reminding us of our neighbors down South, Sergio Broholm explores NAFTA's problematic impact on Mexican corn farmers.

The Twenty-First Editorial Board has chosen scholarship that provides readers with a nuanced understanding of those issues that frequent our daily headlines: China, energy, Iran, education, Central Asia, India, the economic recession. The essays' novel contributions equip us with the complex perspectives necessary to understand world events and the growing number of relationships between countries.

This Journal is the product of the diligence of the Helvidius Group's excellent staff, and it has been a true privilege to work with them over the past four years. I hope that you enjoy this twenty-first edition and keep the Journal in mind as you engage with the relevant issues of our time.

Anusha Sriram
Editor-in-Chief

New York City
April 2010

Guest Essay

THE NEIGHBORHOOD OF AN OCEAN

Dr. Shashi Tharoor

What international association brings together eighteen countries straddling three continents thousands of miles apart, united solely by their sharing of a common body of water? That’s a quiz question likely to stump the most devoted aficionado of global politics. It’s the Indian Ocean Rim Countries’ Association for Regional Co-operation, blessed with the unwieldy acronym IOR-ARC, which bids fair to be the most extraordinary international grouping you’ve never heard of.

The Association manages to unite Australia and Iran, Singapore and India, Madagascar and the United Arab Emirates, and a dozen other states large and small—unlikely partners brought together by the fact that the Indian Ocean washes their shores. As India’s Minister of State for External Affairs, I have attended their ministerial meeting in Sana’a, Yemen, and despite being used to my eyes glazing over at the alphabet soup of international organizations I’ve encountered during a three-decade UN career, I find myself excited by the potential of this one.

Regional associations have been created on a variety of premises: geographical, as with the African Union; geopolitical, as with the Organization of American States; economic and commercial, as with ASEAN or Mercosur; security-driven, as with NATO. There are multi-continental ones too, like IBSA, which brings together India, Brazil, and South Africa, or the better-known G-20. And even Goldman Sachs can claim to have invented an inter-governmental body, since the “BRIC” concept coined by that Wall Street

Dr. Tharoor, an award-winning novelist and commentator, is the former United Nations Under-Secretary-General. He is currently Minister of State for External Affairs in the Government of India.
firm was recently reified by a meeting of the heads of government of Brazil, Russia, India and China in Yekaterinburg last month. But it’s fair to say there’s nothing quite like IOR-ARC in the annals of global diplomacy.

For one thing, there isn’t another ocean on the planet that takes in Asia, Africa and Oceania (and could embrace Europe too, since the French department of Reunion, in the Indian Ocean, gives Paris observer status in IOR-ARC, and the Quai d’Orsay is considering seeking full membership). For another, every one of Huntington’s famously clashing civilizations finds a representative amongst the members, giving a common roof to the widest possible array of worldviews in their smallest imaginable combination (just eighteen countries). When IOR-ARC meets, new windows are opened between countries separated by distance as well as politics. Malaysians talk to Mauritians, Arabs with Australians, South Africans with Sri Lankans, Iranians with Indonesians. The Indian Ocean serves as both a sea separating them and a bridge linking them together.

And the potential of the organization is huge. There are opportunities to learn from each other, to share experiences and to pool resources on such water-borne issues as blue-water fishing, maritime transport, and piracy (in the Gulf of Aden and the waters off Somalia, as well as in the straits of Malacca). But IOR-ARC doesn’t have to confine itself to the water: it’s the member countries that are members, not just their coastlines. So everything from the development of tourism in the eighteen countries to the transfer of science and technology is on the table. The poorer developing countries have new partners to offer educational scholarships to their young and training courses for their government officers. There’s already talk of new projects in capacity building, agriculture and the promotion of cultural co-operation.

This is not to imply that IOR-ARC has yet fulfilled its potential in the decade that it has been in existence. As often happens with brilliant ideas, the creative spark consumes itself in the act of creation, and IOR-ARC has been treading water, not having done enough to get beyond the declaratory phase that marks most new initiatives. The organization itself is lean to the point of emaciation, with just half-a-dozen staff in its Mauritius secretariat (including the gardener!). I visited the rather forlorn-looking Headquarters in Port Louis and was concerned at the staff’s perception that the Member States had not yet accorded adequate priority to the Association. The formula of pursuing work in an Academic Group, a Business Forum, and a Working Group on Trade and Investment has not yet brought either focus or drive to the parent body.

But such teething troubles are inevitable in any new group, and the seeds of future co-operation have already been sown. Making a success of an association that unites large countries and small ones, island states and continental ones, Islamic republics, monarchies and liberal democracies, and every race known to mankind, represents both a challenge and an opportunity. This very diversity of interests and capabilities can easily impede substantive co-operation, but it can also make such co-operation far more rewarding. In this diversity, we in India see immense possibilities, and in Sana’a we pledged ourselves to energizing and reviving this semi-dormant organization.

The brotherhood of man is a tired cliché; the neighborhood of an ocean is a refreshing new idea. The world as a whole stands to benefit if eighteen littoral states can find common ground in the churning waters of a mighty ocean.
CAUSES AND EFFECTS OF THE GLOBAL FINANCIAL CRISIS

Dominick Salvatore

Advanced countries are currently in the midst of a financial crisis and deep economic recession, and emerging markets are experiencing either a recession or a sharp slowdown in economic growth. In this paper, I will examine the causes, effects, policies, and prospects for the financial crisis and how it has led to such serious consequences.

CAUSES OF THE FINANCIAL CRISIS

The present financial crisis started in the U.S. subprime mortgage market in 2007. After it expanded into the financial and real sectors of the U.S. economy in 2008, it then spread to the rest of the world. The initial causes of the financial crisis are clear: huge and increasing amounts of home mortgages, often based on weak underwriting, were given to individuals and families that clearly could not afford them. These mortgages were made at variable rates when rates were the lowest in fifty years. It was only to be expected that a rise in interest rates would render many homeowners unable to make their mortgage payments and default. The crisis could only have been avoided if housing prices had continued to rise at the unrealistically high rates of 2000–2005.

These subprime home mortgages were then repackaged into mortgage-backed securities (MBS) and sold to credit market investors. Rating agencies, such as Moody's and Standard & Poor, gave some of these financial instruments triple-A ratings. Finally, the Securities and Exchange Commission (SEC), responsible for regulating this market, was not adequately involved in these transactions.

Although the problem of subprime mortgages greatly expanded during the presidency of George W. Bush, the practice began in 1999 during the Clinton Administration when Fannie Mae and Freddie Mac were encouraged to grant home mortgages to people who clearly could not afford them in order to support the American Dream of owning a home. Some economists cite deregulation as the primary cause of the crisis. Indeed, the repeal of the Depression-era Glass-Steagall Act in 1999—pushed by Alan Greenspan and Robert Rubin when Larry Summers was Treasury Secretary under Clinton—ended the separation of commercial banking from other financial activities, such as insurance, underwriting, and investment banking, and made possible some of the financial excesses that led to the present crisis. In 1998, Greenspan, Rubin, and Summers were also the ones who objected to the imposition of any regulation on credit default swaps (CDS), which the famed investor Warren Buffett once called “weapons of financial mass destruction.”

Crises of the present financial crisis were thus deregulation or inadequate regulation of investment banking, the inadequate application of regulations already on the books (rating agencies and the SEC), unfortunate economic policies (granting home mortgages to those unable to afford them), economic greed (financial firms caught in a gigantic profit-seeking schemes with insufficient risk management), and outright fraud (Bernard Madoff’s $65 billion Ponzi scheme).

CONTAGION: THE SPREAD OF THE FINANCIAL CRISIS AROUND THE GLOBE

Then there was contagion, by which the crisis in the United States spread first to other advanced countries through the global financial system and finally to emerging markets when the former fell into recession and sharply reduced their imports from and capital investments in the latter.
Europe, however, was subject to even bigger excesses than what had occurred in the United States, leading to a faster spread of contagion. Bank leverage, which measures the risk a bank faces, was thirty-five for the largest twelve European banks but twelve for the largest twelve U.S. banks. Bank leverage at Lehman Brothers was thirty-one when it failed in March 2008, but at European banks UBS, Deutsche Bank, and Barclays, it was forty-two, fifty-six, and sixty-three, respectively. The housing bubble was also even more inflated in some European countries than in the United States. The peak housing prices were 2.6 times higher than their long-run trend in Ireland, 2.1 in the United Kingdom, and 1.9 in Spain, as compared with 1.7 in the United States. If Europe had not faced even greater excesses in some sectors, contagion would not have occurred as rapidly and as strongly as it did through the financial sector.

The crucial event that triggered the crisis was, of course, the failure of Lehman Brothers in September 2008. Lehman was allowed to fail presumably because its assets were less solid than those of Bear Sterns, which was acquired by J.P. Morgan Chase the previous March to prevent its failure, because there were no buyers after Treasury Secretary Henry Paulson refused to provide $60 billion of loss guarantees to Barclays and Bank of America, the two banks that had shown interest in acquiring Lehman. It is more likely that Paulson wanted to use the failure of Lehman Brothers to avoid the accusation of falling into the moral hazard trap—the situation in which profits are private and losses are public—and to teach a lesson to financial markets. However, he subsequently admitted to having underestimated the size of Lehman and the problem its failure would create. At the time of its failure, Lehman had sold nearly $700 billion in bonds and derivatives, of which about $160 billion were unsecured. Nonetheless, rescuing Lehman, would have only postponed the crisis, not prevented it.

Some economists blame the operation of the international monetary system for the crisis, but the present crisis has for the most part a domestic rather than international origin. A more efficient and effective international monetary system would not have prevented contagion around the world because, as we have seen, greater financial excesses than in the United States had occurred in Europe, Japan and other countries.

**EFFECTS OF THE CRISIS**

The outcome of the current financial crisis is evident to all. The severity of its impact is evidenced by the following:

1. Stock markets crashed all over the world in 2008, with declines ranging from thirty-one percent in the United Kingdom to fifty percent in Italy among advanced countries, and from twenty-four percent in Mexico to sixty-five percent in China and Russia among emerging markets.
2. The capitalization of banks was cut by more than half from more than $8 trillion at the end of 2007 to $4 trillion at the end of 2008. Between March and September 2008, the entire U.S. investment banking sector as we had known it disappeared. Investment banking in the United States will be conducted mostly by commercial banks under more highly regulated and less speculative conditions permitted by the Federal Reserve.
3. All advanced countries fell into the “Great Recession” (the deepest of the postwar period) with real GDP falling by 2.5 percent in the United States, 3.9 percent in the Euro Area, 4.8 percent in the United Kingdom, and 5.3 percent in Japan in 2009.
4. All of the most important and largest emerging market economies, with the exception of China, India, and Indonesia, fell into recession with real GDP falling from 0.4 percent in Brazil to 6.8 percent in Mexico, and 9.0 percent in Russia. From 2007 to 2008, the percentage growth or real GDP declined from 9.0 to 8.5 in China, from 7.3 to 5.4 in India, and from 6.1 to 4.0 in Indonesia. It is important to note that China, India, and Indonesia need very high rates of growth to accommodate their large populations and to absorb their still
significant subsistence sectors into the market economy.

**MONETARY POLICIES, INCREASED LIQUIDITY, AND BANK RESCUES**

The United States and Europe did almost everything possible to avoid the recession, but their efforts only succeeded in preventing a deeper recession or depression. At the beginning of 2008, the United States introduced a $168 billion dollar stimulus package. Although it contributed to a 2.8 percent rise in real GDP in the second quarter of 2009, its effect soon faded away.

The United States lowered its interest rate from 5.25 percent in September 2007, to 1.0 percent in October 2008, and to practically zero in December 2008. Needing more stimulus, the Fed also flooded the market with liquidity, as evidenced by the reserves of commercial banks held at the Fed and the increase in its balance sheet from $900 billion in the summer of 2008 to over $2 trillion today. This could potentially generate an explosive rise in future bank lending and in the money supply, thus leading to a huge inflationary spiral.

To avoid the accusation of moral hazard, the Fed helped J.P. Morgan Chase acquire Bear Sterns in March 2008 at a deeply discounted price with a $29 billion debt guarantee. In May, the Treasury acquired $100 billion of nonvoting Fannie Mae stock, $100 billion of Freddie Mac, and from May to December a total of $185 billion from American Investment Group (AIG). In September, it encouraged and facilitated the acquisition of Merrill Lynch by Bank of America and approved the conversion of Morgan Stanley and Goldman Sachs into commercial banks. In October, it increased insurance on bank deposits from $100,000 to $250,000 and adopted a $700 billion bank rescue plan, with half of the money spent by the end of the year to recapitalize the banking sector and purchase money and commercial paper from firms to compensate for the drying up of this crucial lending activity by commercial banks.

Then in November 2008, the U.S. Treasury injected another $20 billion of new capital to Citigroup, the weakest of the largest U.S. banks, on top of the $25 billion provided in September and, together with the Fed, provided guarantees against excessive losses on $301 billion of toxic assets—mostly subprime personal and commercial loans owned by Citigroup—to prevent its collapse. In January 2009, the Treasury injected another $20 billion of new capital to Bank of America and Merrill Lynch on top of the $25 billion injected in September. In conjunction with the Fed, the Treasury also provided guarantees against excessive losses on $100 billion of toxic assets to prevent Bank of America from withdrawing from the purchase of Merrill Lynch after it discovered that the latter had even more toxic assets than anticipated when Bank of America initially agreed to purchase it. At the end of February, the U.S. government agreed to become the biggest single shareholder of Citigroup by taking a 36 percent stake of the troubled lender to prevent its collapse.

Many European countries adopted similar but less ambitious policies to stimulate their economies. In January 2009, the European Central bank cut the interest rate from 4.25 percent in July 2008 to 1.0 percent in April 2009, but indicated that it would not follow the U.S. and Japanese counterparts down the path of practically zero interest. The Bank of England cut the interest rate even more drastically from 5 percent in October 2008 to 0.5 percent in April 2009, the lowest since its creation in 1694. All of these measures, however, did not prevent an even deeper recession in Europe than in the United States.

**EXIT STRATEGY, THE DANGER OF INFLATION, AND U.S. DEBT**

With the resumption of growth, the potential for inflationary explosion becomes a serious danger. Indeed, the fear in the market that the Fed would not be able to reverse course in unwinding its huge unconventional monetary stimulus, prompted Chairman of the Federal Reserve Ben Bernanke to outline the Fed’s exit strategy in July 2009. This immediately calmed markets and led to a sharp decline in U.S. Treasuries. Bernanke testified that he expected the
U.S. economy to start growing again at the end of the year but, with unemployment likely to reach nearly ten percent, growth would very likely be slow through 2011, so that the economy would face little inflationary pressure.

When necessary, Bernanke indicated that an exit strategy could be established very quickly to mop up the excess liquidity by letting emergency lending programs wind down or expire, raising the short-term interest rates paid on reserve balances (to help set a floor under interest rates), letting short-term credits expire, and selling longer-term assets to the public. He acknowledged that, as always, the difficulty will be deciding the precise timing and set pace of the tightening effort.

At the same time, Bernanke warned Congress and the White House to get budget deficits under control or risk damaging the recovery. The stimulus package and all other expenditures to bail out the banking sector will lead to much higher U.S. government debt and taxes in the years to come. Balancing the out-year budget projected by the Congressional Budget Office (CBO) would require a 44 percent increase in taxes. Without an increase in taxes, the U.S. government debt as a percentage of GDP is expected to increase from 40 in 2008, to 65 in 2010, 70 in 2012, and 103 in 2017. Faced with a drastic decline in their wealth as a result of the deep recession and anticipating much higher taxes in the future to pay for the stimulus package and the other huge government programs\(^1\) to overcome the crisis, business will invest less and individuals and families will tend to save more and spend less, leading to a very low multiplier (barely above one) for every stimulus dollar spent.

Europe and Japan generally have smaller stimulus packages in relation to their GDP than does the United States because of a stronger social welfare net and to avoid very large increases in their already high government debts. Despite this, all advanced countries will face sharply rising national debts in the years to come.

\(^1\) In March 2010, the U.S. Congress pushed through a health care reform plan that may cost $1 trillion according to CBO estimates.
on mortgages and credit cards to consumers hard pressed by the recession. Banks will also be pressured by losses on commercial real estate, which account for thirteen percent of U.S. GDP. Although Goldman Sachs and J.P. Morgan are reporting strong earnings, they are based on reduced competition and funds deployed in risky businesses where spreads are high. Most other large banks (e.g., Citigroup, Wells Fargo, and Morgan Stanley) as well as smaller regional ones reported or face huge losses on bad loans.

Designed to enable banks to dump troubled assets, the Public-Private Investment Program that was introduced by Treasury Secretary Timothy Geithner in March 2009 has been scaled down considerably and now focuses solely on toxic securities, not the bad loans that are the main source of problems for many smaller banks. As these loans deteriorate and securities lose value, banks may be unwilling or unable to make new loans to consumers and business, thus hindering the recovery. As bad as the banking problem is in the United States, it is expected to be worse in Europe, where only a small portion of bad loans have been written off as of the end of 2009 as compared with about half in the United States.

THE UNEMPLOYMENT PROBLEM

Despite the resumption of growth, unemployment remains a serious concern in all advanced countries. In the United States, nearly seven million jobs have been lost since the beginning of the crisis and ten million jobs are likely needed to reach full employment since an additional three million workers have entered the labor force during the past two years. As a lagging indicator, the rate of unemployment, which was ten percent at the end of 2009, may fall only slightly during 2010.

The unemployment problem is much worse than the Obama Administration had projected. The stimulus package simply did not deliver the type of “immediate help” expected and promised by the administration. Simulations, however, indicate that without the stimulus package, the number of unemployed would be approximately one million higher than current levels. One important reason is that most of the billions of dollars “made available” for increased expenditures appear to have gone instead into savings. This was to be expected in view of the $13 billion destruction incurred in household wealth during the past two years. To be fair, the stimulus has been in place for just over a year and more time is necessary to evaluate its full effect.

Unemployment has also increased in Europe since the beginning of the current recession, but somewhat less rapidly than in the United States because of work policies that have reduced the number of layoffs in European firms.

REFORMS TO PREVENT FUTURE CRISSES

Important reforms are clearly needed to avoid future financial crises. Reforms, however, need to be comprehensive and general. Comprehensive change is necessary, because both inadequate regulations on investment banking and inadequate application of regulations that were on the books by the SEC contributed to the current financial crisis. Regulations also need to be broad and general rather than specific and pointed. Money is fungible, and when a specific regulation closes one avenue of creative financial excess, operators soon find other ways to bypass the regulation. Regulations should restrict or prohibit the use of exotic derivatives. Sellers often cannot explain and buyers cannot understand how these derivatives are priced or how they work.

The United States needs to save more and learn to live within its means. Americans save little relative to individuals in other developed countries and the U.S. government is a large net borrower. However, these factors did not quash demand for profitable investment opportunities in the United States. The result was a huge inflow of funds from abroad, which led to an overvalued dollar and unsustainable U.S. trade deficits. Substantially raising the household savings rate in the United States will be difficult with Americans addicted to overspending, but it is a goal the United States

\[2\] This is true with the exception of Spain, where it reached 18 percent by the end of 2009.
could potentially achieve gradually. Surprisingly household savings already seem to have begun rising in the past year.

In the present crisis atmosphere, many nations may over-regulate and impose excessive restrictions on financial activities that would be detrimental to future growth. There is also the danger that the large injection of liquidity in the United States and in other advanced countries to jump-start their economies will lead to hyperinflation in two to three years’ time, which would then require a sharp tightening of monetary policy.

CONCLUSION

Eventually this crisis will end as all crises do. The important question then becomes: will growth in advanced countries, especially in the United States, be rapid or slow? In short, will there be growth or stagnation after recession? Of course, no one can know for sure, but it is likely that some of the growth factors of the last decade will be weaker during this new decade. Specifically, the deregulation process is going into reverse and there will certainly be less financial innovation as drivers of growth during this decade compared to the last decade. The United States may also learning to live within its means by saving more and spending less; by so doing it would remove about $500 billion of U.S. imports from global demand.

On the other hand, the continued rapid application of information technology is likely to continue to provide as important a stimulus to growth as it did during the past decade. Some emerging markets, especially China and India, are likely to continue to grow very rapidly and provide a strong stimulus to growth. These positive growth factors may be sufficient to overcome the negative growth factors, but we will have to wait and see to determine if they will be sufficient to propel the world economy toward high and sustained growth in the future.

The Peter and Katherine Tomassi Essay

THE IRANIAN REVOLUTION: ASSESSING THE POWER, INFLUENCE AND SOCIAL POSITION OF SHIITE ULAMA IN IRAN, 1890–1979

José Ciro Martinez

It was an oft-unrecognized assumption of modernization theory, the dominant social science paradigm of the 1960s and 1970s, that the character and trajectory of historical change was both universal and unilinear. Drawing mainly on the work of Max Weber, scholars such as David Apter, Seymour Martin Lipset, and Middle East expert Daniel Lerner argued that economic growth, capitalism, urbanization, and the impact of Western cultural forms were essential factors for democratic development and would result in the eradication of primitive or traditional forms of societal organization and everyday life. In response to this historical advance, religious piety and conviction (both on a personal and societal level) in places like the Middle East would be radically reduced if not eliminated, all as part of a teleological process of disenchantment. Then, Weber’s “traditional forms of authority” would slowly erode and full-blown rationalization of government and society would follow. It is thus not surprising that the events of 1978–79 and the establishment of the Islamic Republic of Iran stunned many of the scholars who, in the period following Muhammad Reza Shah’s “White Revolution,” predicted a decline in the ulama’s power and the inevitable triumph of secularism and means-end rationality.

Although most modernization theorists draw much of their theoretical arguments from the work of Weber, I argue that their usage of his texts in explaining historical change relies on a fundamental misreading of his work. Although the rationalization of economic forms and daily activities does erode and eventually eliminate many of the time-honored forms of life and authority seen in ‘traditional’ society, this process is not inevitable. Indeed I argue that Weber’s entire book aims to explain how odd this transition is and why it has, until 1905, mostly confined itself to Britain and the United States.

Nikki Keddie, a close and careful scholar of Iranian history, was as swayed by José Martinez is a senior at Williams College majoring in Political Science and History. He will pursue a Thomas J. Watson Fellowship for independent study outside of the United States after graduation.
could potentially achieve gradually. Surprisingly household savings already seem to have begun rising in the past year.

In the present crisis atmosphere, many nations may over-regulate and impose excessive restrictions on financial activities that would be detrimental to future growth. There is also the danger that the large injection of liquidity in the United States and in other advanced countries to jump-start their economies will lead to hyperinflation in two to three years’ time, which would then require a sharp tightening of monetary policy.

**CONCLUSION**

Eventually this crisis will end as all crises do. The important question then becomes: will growth in advanced countries, especially in the United States, be rapid or slow? In short, will there be growth or stagnation after recession? Of course, no one can know for sure, but it is likely that some of the growth factors of the last decade will be weaker during this new decade. Specifically, the deregulation process is going into reverse and there will certainly be less financial innovation as drivers of growth during this decade compared to the last decade. The United States may also learning to live within its means by saving more and spending less; by so doing it would remove about $500 billion of U.S. imports from global demand.

On the other hand, the continued rapid application of information technology is likely to continue to provide as important a stimulus to growth as it did during the past decade. Some emerging markets, especially China and India, are likely to continue to grow very rapidly and provide a strong stimulus to growth. These positive growth factors may be sufficient to overcome the negative growth factors, but we will have to wait and see to determine if they will be sufficient to propel the world economy toward high and sustained growth in the future.

---

1 Although most modernization theorists draw much of their theoretical arguments from the work of Weber, I argue that their usage of his texts in explaining historical change relies on a fundamental misreading of his work. Although the rationalization of economic forms and daily activities does erode and eventually eliminate many of the time-honored forms of life and authority seen in ‘traditional’ society, this process is not inevitable. Indeed I argue that Weber’s entire book aims to explain how odd this transition is and why it has, until 1905, mostly confined itself to Britain and the United States.

2 Nikki Keddie, a close and careful scholar of Iranian history, was as swayed by José Martinez is a senior at Williams College majoring in Political Science and History. He will pursue a Thomas J. Watson Fellowship for independent study outside of the United States after graduation.
Modernization theory proved ill-equipped to explain the reemergence of religious actors and Islamist political forms in the public arena during what these theorists often described as the modern age. Certain schools of thought, most prominently represented by the work of historian Bernard Lewis, saw the Middle East as an exceptional region, due primarily to the nature of Islam, unique in its rejection of modernity and rationality (Lewis 2002). In making sense of the Iranian Revolution, many of these authors decided not to grapple with the historical roots of the actors that helped bring about this seemingly inconceivable event. Instead, they described it as “mystifying,” and especially hard to understand as Iranians had apparently decided to “demand less freedom and fewer material things” (Wright 1988). Other, more perceptive and well-read critics have resorted to describing the innovative political strategies and theological justifications used by Ayatollah Khomeini and his allies as representing some kind of return to a Shiite fundamentalism more reminiscent of the Middle Ages than the late twentieth century.

Following the theoretical framework employed by Ervand Abrahamian, this work aims to elucidate the diverse and distinctive populist, nationalist, and Islamist tendencies and strategies employed by the nominal leader of the Iranian Revolution, Ayatollah Ruhollah Musavi Khomeini (Abrahamian 1982). It does this by exploring the historical role of the clerical class in Iran, a country with a large Shiite majority. By utilizing the theories of political development and state building employed by Barrington Moore and Charles Tilly, we have the tools to better examine and comprehend the influence and power of the Shiite mujtahids (Shiite religious scholars who are empowered to interpret legal issues not explicitly addressed in Quran), a topic that has confounded many analysts and policymakers. Through an examination of the position of Shiite clerics in relation to both the dominant mode of production and the processes of state-making, the ulama are unmasked and revealed as to be a social category or occupational group defined by its distinctive religious and occupational functions that has proven to be ideologically malleable, intellectually flexible, and concretely responsive to the political, economic, and social milieu that surrounds it.

Although our task is limited to explaining the variation in mujtahid influence over the past century, this work will contribute to the literature on Twelver Shiite forms, the role of Iranian ulama in Iran’s politics and foreign policy. It informs U.S. and Western foreign policy-makers who are interested in the power structure of a country that may be pursuing a nuclear weapons capability. My findings directly challenge the popular narrative that portrays Islamic revolutionaries and overtly Shiite political actors as ideologically-driven, religious fanatics.

METHODOLOGY

I employ macro-historical analysis to evaluate the role and social position of clerical elites in relation to political formations, class actors, and state institutions in the prelude to the Iranian Revolution of 1978–79. In doing so, I focus on two major historical events: the Tobacco Revolt of 1891–92 and the Iranian Revolution of 1978–79. I seek to “frame the analytical problem and suggest causal ideas that are evaluated and explored through a within-case analysis.”

4 Some prefer the latter term because it helps capture the notion of craft consciousness. I will employ the former throughout the essay as I feel it is more attuned to relations of production and consumption.

5 My definition of a social category (and its application to the Shiite ulama) relies on the work of Mansoor Moaddel for its empirical grounding and on Nico Poulantzas’s conceptualization of the term, for its theoretical underpinnings. Poulantzas defines social categories as ‘social ensembles with ‘pertinent effects,’ which may become social forces whose distinguishing feature is based on their specific and over-determining relation to the structures other than economic ones.”
Shiism will be theorized as one among a number of constructed identities that need to be equally generated, adopted and used for mobilization by every new generation (Cole 2005, 2–3). When analyzing Shiite religious elites, this account will examine the role played by material, political, and coercive structures and circumstances on the social position of Shiite religious elites.6

This investigation begins by defining the universe of cases to clarify the scope of its empirical categories while refining the concepts—social categories, social position, coercion, capital, and state-making—in usage. I then use the critical juncture technique,7 close knowledge of cases, and within-case analysis to establish scope restrictions on my findings, limiting possible causal mechanisms and links to the case-studies themselves while positing a broader theory for analyzing religious actors in Shiite societies. I employ “contextualized comparison” (Locke and Thelen 1995, 337–367) and qualitative content analysis to collect data and increase measurement validity. Next, I re-employ within-case analysis (mainly to evaluate hypothesis by revealing intervening processes and other observable implications of arguments) and the critical juncture framework to assess historical causation (Munch 2004, 111–119). After iterated refinement to modify my hypotheses, I find that the substantial amount of explanatory variables necessitated a detailed, case-oriented analysis. I argue that chronology and the use of narrative is crucial to contextually-based judgment and for establishing plausible causal links in a comparative macro-historical research.

Through this multi-faceted methodological approach, I explain the emergence of Shiite religious actors in Iran during two particular historical events or political conjunctures. I do this through the use of general concepts and approaches, thus refuting the predominant literature concerning political development in the Middle East that takes the region’s religion and culture as its starting point of analysis. By focusing on a group of the most explicitly religious social actors in a majority-Muslim country, this essay will disprove flawed understandings of the nature of political authority in Shiism and Islam while undermining the common theological or culturalist explanations that are often cited as explaining the lack of democracy, secularism, or ‘modernity’ in the Islamic world (Berman 2003, 257–272). Ultimately, I argue for the failure of Islamic-based exceptionalism or a religious or ethnically based essentialism as an explanatory variable in the Middle East by consciously employing canonical works in comparative politics to the analysis of the region, an area of the world that is often neglected and wholly under-theorized in this literature.

HISTORICAL BACKGROUND: THE ULAMA’S POWER IN THE QAJAR PERIOD (1794–1925)

After the eighteenth century, the ulama that dominated Twelver Shiite approaches to Islamic jurisprudence and participate in historically significant roles in Iraqi and Iranian politics were followers of a particular brand of Shiism. Labeled as Usulism, this school of thought stresses the use of independent legal judgment, or *ijtihad*, by a group of competent Shiite religious scholars, or *muftahids*. Usulis believed that *muftahids* could serve as the surrogates of the hidden Imam during the *ghayba*, or occultation.8

6 In this respect, the work relies on the insights of Talal Akad, who describes religious tradition and theology as a set of arguments and debates that are a product of a dynamic social construction instead of a static and unchanging object of transmission.

7 My use of the framework leans more towards its usage by R. Collier and D. Collier (2002) than Capoccia and Kelemen (2007) who focus on individual agency and selection versus structural antecedents in institutional choice.

8 The Akbari-Usuli debate represents a long-existing schism between the two main legal schools of thought in Twelver Shiism. During the Imamate (660–871), Shiites placed all spiritual authority in the hands of the divine Imam. When the Imamate ended as a worldly institution (this occurred when the twelfth Imam, Muhammad al-Mahdi, goes into hiding), Shiites experienced what numerous analysts have called a crisis of authority. Twelver Shiism left no explicit designation for temporal or spiritual power after the disappearance of the Imam; in response, Shiite theologians put forward two chief responses. The Akbaris, also known as traditionalists or literalists, were willing to see certain crucial Muslim customs (authorization of Friday prayers, collection of taxes, designation of religious leaders, etc.) lapse in the absence of the spiritual figure of the Imam. They rejected any use of *ijtihad* and promoted the Quran and the hadith as the only legitimate sources of law (hence their neglect for rituals not explicitly contained in these two works). They claimed that the Imams were the only true Ayatollahs, rejecting and even cursing the historically novel position of *muftahid*. In contrast, the ‘rationalist’ Usulis believed that a group of Shiite religious scholars could stand as the surrogate or deputy of the Imam during the *ghayba* or occultation. Nevertheless, Usulis differ in terms of their position on the clerics'
promotes *taqlid* or imitation, the worshippers’ practice of following rulings on religious and personal matters made by their mujtahids.

While I disagree with most interpretations that trace the triumph of Usulism to the beginning of the Qajar dynasty, it is undeniable that processes of historical development and state-building were crucial in consolidating the triumph of Usulism towards the end of the eighteenth century.\(^9\) Usulism’s emphasis on a hierarchical, legalistic form of Shiism along with its legitimization of state-related functions (including but not limited to tax collection, juridical functions, and defensive holy war) made it far more amenable to the needs or rising rulers in places as varied as Iran and North India (Cole 1986, 461–480; Nakash 1994, 443–463; Satia 2007, 211–255). Consequently, this theological and philosophical approach allowed the Shiite clergy to ally with local rulers and monarchs on favorable terms, often resulting in expansion of their landholdings and close ties to the royal court and aristocracy.

Mostly of nomadic descent, the Qajar dynasty lacked both the bureaucratic apparatus and the necessary legitimacy to rule the expansive Persian plateau. Much like their European counterparts, the Qajars decided to rely on religious groups (in this case, the Usuli *ulama*) to perform educational, judicial, and legitimatizing functions that were crucial to the statebuilding process (Algar 1980, 33–43; Moaddel 1986, 519–524). It is thus not surprising that as the Qajars began to consolidate their power at the end of the eighteenth century, the Usuli *ulama* began to establish, in conjunction with the state, their own exclusive spheres of religious relationship to authority, although the dominant position since the beginning of the Qajar period has been to legitimize non-sharia based states while attempting to maintain certain levels of clerical autonomy.

\(^9\) In dating the triumph of Usuli approaches to Islamic jurisprudence in both Iran and Iraq, I mostly agree with Juan Cole (2005, 58–77), who traces this historical development to the end of the Zand era in Iran (1760–1780). Interestingly, the triumph of Usulism in the shrine cities of what is known today as Iraq was intimately tied to the decentralized autonomy gained during the late Ottoman Empire. In contrast, Usulism in Iran was intimately tied to the state-building process under both the Zand and Qajar dynasties. This is not surprising given the historical role played by Usuli scholars in the Persian plateau, especially in legitimizing the temporal and religious position of Iran’s first Shiite dynasty, the Safavids (1501–1722).

As a result of this complex historical process, the *ulama* of Iran cemented links to a variety of institutions and social classes. Negotiations between the *ulama* and the Qajars were made inevitable due to the Qajar’s well-known inability to centralize power (as in Weber’s modern state\(^11\)) and the *ulama*’s relatively high level of capital accumulation compared to other Iranian social classes. This makes the institutionalization of certain forms of *ulama* power hardly a shocking development.\(^12\) The need for financial and juridical autonomy and patronage resulted in the creation of a functional tie between the Shiite religious elites and the newly created Qajar state. Structurally, the *ulama*’s societal position as large landholders and owners (especially in the case of the high ranking mujtahids)\(^13\) created a host of shared interests between them and ‘feudal’ or landholding elites.\(^14\) Finally, a substantial portion of domination,\(^10\) making their joint emergence a “mutually reinforcing historical process” (Moaddel 1986, 522).

10 Throughout most of the Qajar period, Usul *ulama* held firm control of the educational and juridical apparatus.

11 Defined by Weber as a human community that (successfully) claims the monopoly of the legitimate use of physical force within a given territory (Weber 1997).

12 Indeed, Tilly would find the *ulama*’s power during the Qajar period to be a classic example of why, despite obvious interests to the contrary, rulers would accept the establishment of certain institutions representing or claiming to represent important social classes within their jurisdiction. The Qajars’ lackluster tax collection, lack of substantial coercive means, and constant battles with the Ottomans and Russians made their dependence on the *ulama* all the more acute.

13 High-ranking mujtahids were often privy to ties with the bazaar (usually through family connections) and highly dependent on the taxes and donations of merchants, the petty bourgeoisie and the Qajar dynasty. In contrast, lower ranking *ulama* were usually bereft of access to wealth or sources of patronage with little landowning privileges. They were generally more attuned to the needs and wants of the lower classes and were more willing than their senior counterparts to incite rebellions and local uprisings. As Moaddel points out, “the intense struggle between landlord and peasant that occurred in the mid-nineteenth century had an integrating effect on the *ulama*-state relationship while it created a somewhat horizontal division between the high- and low-ranking *ulama*. A considerable portion of the low ranking *ulama* supported the cause of the peasants and the urban poor” (Moaddel 1986, 524).

14 The increase in *ulama* landownership is, to a certain extent, a legacy of the Safavid Empire, during which the composition of the landowning class saw a great increase in the amount held by religious elites. Pre-*ulama* policies of the Qajar rulers, especially after the suppression of the Babi movement in the mid-nineteenth century, saw an even greater
the ulama’s revenues came from religious taxes (khums and zakat), donations and payments for juridical rulings, and intermediations made largely by the merchants of the bazaar and the traditional petty bourgeoisie. As a result, the ulama had an instrumental relationship to these two classes. Akin to developments in Europe during the state-building process, religious elites’ ties to a diverse group of interests and social classes were the chief determinants of their political (and, to a certain extent, theological) positions and provided the main reason for agreements and disagreements among this heterogeneous group.  

As Barrington Moore points out, the English bourgeoisie had an economic base largely independent of the Tudor monarchy. The gentry and the nobility increasingly adopted capitalist tendencies and, after the end of the enclosure movement, did not have the same need for political power or influence as their European counterparts (Moore 1993, 3–39). Therefore, the bourgeoisie were able to gradually gain control of the state’s institutions and began to direct government policies towards protecting domestic markets or to promote free trade in accordance with their shifting socio-economic interests. In contrast, the Qajar dynasty is a classic example of a weak state (with little coercive power or means of fiscal extraction) that came to rely on foreign interests and a heavy dose of patronage to perpetuate its rule (Eliash 1979; Keddie 1966; Hairi 1988; Moaddel 1986, 523–534). The Shah’s court bore something of a resemblance to that of Louis XVI, where landholders and budding capitalists could only gain favor and economic privileges with the assistance of the sovereign. As a result, both the merchants and the petty bourgeoisie became increasingly marginalized, both politically and economically. Their common cause against the state increase in the amount of land owned by a significant group of ulama (Lambton 1991, 129–50; Floor 2001, 1–36; Floor 1998, 522–24).  

15 This is crucial to explaining the different political positions taken by ulama throughout Iranian history and seen clearly in events such as the 1906 Constitutional Revolution, the 1953 overthrow of Prime Minister Muhammad Mossadegh and during debates over the constitution of the Islamic Republic in 1979. The ulama have never been a homogenous group in terms of their political, economic or social outlook. Their varied interaction with a variety of class actors and coercive apparatuses ensures the heterogeneity of this social group.

16 The tobacco crop was valuable not only for sale in domestic markets but because of its export to numerous foreign markets. Persian tobacco was highly valued in much of Western Europe. Peasants and members of the proletariat also mobilized at certain moments of the Protest due to their large presence in the tobacco industry (The industry employed around 200,000 people at the time of the Protest). (Mottahedeh 1985, 215–18; Keddie 1966, 65–6).
united over this issue. Although the Islamic Republic has gone to great pains to alter historical accounts of this crucial event, a closer examination of Shiite religious elites will prove their responsiveness to political and economic calculations.

Ayatollah Behbahani, one of the prominent jurists of Tehran, was a leading clerical opponent of the tobacco protests. Behbahani had well-known links to the Shah’s prime minister, Amin-us-Sultan, and is widely thought to have been bribed by the British (Moaddel 1986, 530). In addition, his reliance on court patronage, a product of his lack of a significant independent economic base and his residence in the imperial capital, provided him with little reason to participate in the movement. The clergy of Mashad had a history of intimate ties with the Qajar court (especially since its role in the suppression of the Babism in the 1850s) and was supported mostly by export merchants who largely depended on British markets and access to trade with India. This caused ayatollahs who had ties to Mashad, such as Shaykh Muhammad Rahim and Habibullah Hosseini, to become intolerant toward the “ruffians and rogues” once societal chaos threatened their economic position (Moaddel 1986, 528–532). The faction of the ulama that supported the concession were almost always those who had close ties with either the Qajar court, the British government, or both.

Although the ulama played a disproportionate role in the leadership of the Tobacco Protest, it is increasingly clear that those who participated were often financially dependent on the classes that comprised the backbone of the movement: merchants, the petty bourgeoisie and the peasantry (Amanat 1988: 114–122). This is evidently the case with the author of the initial fatwa that banned the usage of tobacco, Grand Ayatollah Mirza Shirazi. Pressured by numerous clerical allies (most famously by Islamic modernist and philosopher Sayyid Jamal al-Din al-Afghani) who were vulnerable to the impact of the concession, Shirazi also had crucial links to classes that would be affected by the concession such as the vibrant merchants and dealers of tobacco in Shiraz, the petty bourgeoisie of Samarra, and the financiers of Fars and Isfahan (Gilbar 1977, 282–288). Though initially uninvolved, there was eventually a considerable amount of pressure by merchants, peasants, and sympathetic ulama to increase the grand ayatollah’s involvement. In due course, the lobbying worked. Its traces can be seen in Shirazi’s letters to Nasir al-Din Shah, which reflected a clever combination of religious and merchant concerns. Bereft of real political power, the merchants and the petty bourgeoisie cleverly resorted to exerting pressure on the ulama class—an avenue that, if pressured enough, could lead to a confrontation with the Shah. It is hard to conceive of a successful Tobacco Protest without the pressure exerted by the ulama and their bazaar-linked supporters. In this case, a combination of societal pressure (exerted mostly through religious channels) and Nasir al-Din Shah’s perception of British vacillation led to the cancellation of the concession in January of 1892.  

17 Moaddel posits that, “the presence of diverse factions among the ulama over the tobacco concession is a matter of historical fact” (Moaddel 1986, 530).
18 The Behbahani family has a long history of supporting British goals in Iran. The Ayatollah Behbahani mentioned in this essay had a son who followed in his footsteps as a cleric and played a crucial role in opposing the oil nationalization objectives of Prime Minister Mossadegh. He is remembered fondly by royalists for his role in organizing hooligans and looting National Front headquarters on the eve of Operation Ajax, the 1953 coup d’état, organized by the American CIA and British MI6, which deposed the democratically elected government of Iran.
19 The opening of the Suez Canal allowed for the extensive import of cheap European goods into certain parts of Iran while providing the means for an unprecedented amount of exports of the region’s main cash crops (cotton, opium, grapes, etc.) to the Persian Gulf region and India.
20 This is especially true of those ulama of Najaf and Karbala who had financial links with the tobacco producers of Lar and surrounding Iraqi regions.
21 Grand Ayatollah Shirazi’s letter is quoted by Amanat as saying: “…why is it that today, for the sake of trifling profits and regardless of its corrupt consequences, they allow the infidels to dominate the livelihood and commerce of the Muslims? It was uncharacteristic of your royal prudence to permit your subjects, who indeed are abundant treasures of the state, to be deprived of controlling their own trade and business, and instead to bow to the subjugation of foreign pressures” (Amanat 1988,120).
22 Iranian merchants are said to have shut down numerous bazaars in response to Shirazi’s fatwa. Only following the Shirazi’s legal ruling did the boycott receive broad-based support from the population. A popular myth claims that even the women of the Shah’s harem quit smoking and refused to prepare him his water pipe (Nasr 2007, 122).
23 In a matter of days, Shirazi issued a fatwa repealing his initial one and permitting the
COERCION, CAPITAL AND SOCIETY: CLERICAL ELITES AT THE BEGINNING OF THE TWENTIETH CENTURY

By the end of the nineteenth century, the Iranian ulama had consolidated their position in society and had forcefully demonstrated their capacity to influence political debates and policies (Keddie 1966; Moaddel 1985, 529-552). Economic ties often led them to alliances with urban merchants, the petty bourgeoisie, and, far less frequently, the peasants. The ulama and the merchants each possessed substantial power bases that were largely independent from the Qajar monarchy with “religious direction and economic enterprise tending to complement each other” (Algar 1980, 236). Due mainly to their social position and relative levels of capital accumulation, both the ulama and the merchants could periodically challenge a monarchy that had little coercive power and an even larger need for the tax revenues to which both groups contributed.

By employing Tilly’s conceptual framework regarding the state-building process, it becomes apparent why both the ulama and the merchants were hostile to both a centralized state apparatus that would tax them or a less intrusive apparatus that would grant generous concessions to foreigners. Clerical elites feared encroachments on their judicial power, educational functions, and economic position while the merchants were in the business of avoiding further taxation and regulation of their commercial affairs. For numerous reasons, both groups also saw European interference in Iran as inimical to their interests (Algar 1980, 234–237).

Much in the same way as the bourgeoisie and landed gentry joined forces in England to contest the power of the monarchy, the dárs of tobacco among the Shiite population (Mottahedeh 1985, 215–18).

24 This occurred in Iran for numerous reasons. The two most important factors are the ulama’s control of the educational and juridical apparatus. The former infused many merchants and bazaaris with Shiite religious values. Control over the judiciary meant that commercial practices were regulated by ordinances of the sharia that were, in effect, formulated and interpreted by the mujtahids.

25 The ulama feared the impact of European cultural norms on Iran’s religious values and the societal esteem in which the ulama were held. Merchants feared the typical repercussions that came with opening trade and granting concessions to wealthier, more economically developed states.

In questioning both the historical emergence and the territorial location of the broad alliance that supported the Iranian Revolution, it is important to remember Iran’s status as the country with the largest proportion of Shiite Muslims. The Iranian ulama were able to gain the upper hand in both the prelude to and aftermath of the Revolution because they could draw on a super-majority of Iranians citizens (in contrast to Iraq’s sixty to sixty-five percent or Lebanon’s thirty to forty percent) to disseminate their political goals. Still, it is important to theorize why exactly Khomeini’s radically innovative theological position came to the fore at the time it did in the Iranian socio-political milieu and not, as a counter-factual, in nearby Iraq. By relying on a close historical examination of mujtahid power and influence in Iran, we can better comprehend both the social origins, theological and political motivations , and goals of the opposition movement that emerged in the late 1970s.

26 Ironically, Khomeini’s theory of the rule of the jurisprudent is said to have developed during his time in exile (1965–1978) in the Iraqi holy city of Najaf. Given my analysis of the mujtahids position in this country, it is by no means surprising that Khomeini’s doctrine had little if any purchase in the country in which it was developed intellectually.
and eventually succeeded in dethroning the Shah.27

After the ouster of Prime Minister Mossadeq in 1953, the newly restored Muhammad Reza Shah Pahlavi initiated social and economic programs substantially different from those pursued by his father. Like many developing countries in Latin America at the time, the Shah's policies consisted of a combination of import substitution industrialization—mostly making consumer durable goods—with a strong bias towards the large industrial and commercial classes (Moaddel 1986, 539–540). This move towards large-scale industry increased the economic marginalization of the traditional bazaar and undermined the small-scale industry of the petty bourgeoisie. Mounting state involvement in economic affairs, particularly in the agrarian sector, caused the emergence of a state-dependent bourgeoisie that was heavily tied to foreign capital. By 1963, land reform, along with the expansion in state credit institutions and the growth of a modern financial sector, further alienated the bazaar and landed elites (Graham 1979, 221).28 Some accounts even trace the reemergence of the senior mujtahids on the political scene to their unyielding opposition to a land reform bill that had little real bite to it (Lambton 1969, 58; Moaddel 1986, 541–546).29

This resistance is exemplified in the declarations of Grand Ayatollah Burujirdi at the time, and, ultimately, his successful efforts to block implementation of the effort until 1961. Under Reza Shah, the petty bourgeoisie was the class most antagonized by governmental policies. Under Muhammad Reza Shah and his post-1953 policies, it was not only the petty bourgeoisie but also the whole bazaar, especially consumers, and many landowning elites who felt increasingly disempowered (Pesaran 1982, 501–522). The marginalization of these social classes should be seen as a decisive factor in the growth of the motivational power and influence of the Shiite mujtahids, especially in the period following the White Revolution.30 Many proto-Weberians, Seymour Martin Lipset chief among them, would have predicted a substantial decline in the power of traditional institutions and leaders, such as Shiite religious elites. However, in the aftermath of considerable economic development, it is increasingly clear that so-called “traditional” forms of authority may persist in situations where they are supported and linked to an array of social classes and interests.

Following the explosion in state revenues after the 1973 oil crisis, Muhammad Reza Shah further repressed his most fervent enemies while incorporating possible opponents in the growing professional classes into the state apparatus. For example, the Shah, along with his Prime Minister Amir Abbas Hoveyda, created the Rastakhiz party in early 1975 as the single political organization to which all Iranians had to belong. The party’s compulsory dues and aggressive interference in the socio-economic spheres of many Iranians antagonized a number of social groups, especially Iranian bazaaris (Abrahamian 1982).31 The only two social spheres that can be said to have retained a certain, although minimal, degree of organizational and mobilization-related autonomy during this autocratic period were Shiite religious institutions and the local markets or bazaars. Following further repression and corrosion of their social position, the bazaaris and Shi'ite clergy responded to the grudging measures of political liberalization the Shah was pressured to adopt in early 1977 (Zubaida 2009, 64–81). In addition, the economic austerity measures taken to fight inflation and the flight of foreign investment by the Shah in mid-1977 disproportionately affected the lower classes of Iranian society, providing oppositional groups with a core of citizens which it could easily

---

27 Muariq Al-Karaki (d. 1534), an ‘Amili cleric of the early Safavid Empire, is the only other Shiite scholar said to have had anything remotely similar to Khomieni’s position on the clerical role in governance. Recent findings make even this association highly doubtful as al-Karaki is now said to have wanted only a beefed up legislative function for the ulama.

28 Interestingly, the Shah went so far as to endorse the physical demolition of the traditional bazaar in an attempt to establish newer, more ‘modern’ shopping centers modeled on those in the United States.

29 Ann Lambton called the land reform bill little more than a “window dressing.” We must remember that the senior ulama comprised a significant portion of the Iranian landowning class.

30 The White Revolution is the name given to an extensive series of ‘modernizing’ reforms launched in 1963 by Muhammad Reza Shah.

31 Iranian merchants were especially repulsed by the state’s interference in the market and the anti-profiteering campaigns led by Rastakhiz officials.
mobilize against the regime. In light of a common assault on their combined societal supporters, the ulama, for the first time in Iranian history, turned almost unanimously against the state.\textsuperscript{32} The consequences of this novel opposition for the Shah and his cohort would be disastrous.

A number of events further eroded the Shah’s position. The death of Ali Shariati, which was blamed by many of his followers on the Iranian secret service (Savak), further undermined the ruling elite while removing a popular figure who could have possibly counter-balanced Khomeini’s influence in the oppositional movement. The subsequent death of Khomeini’s son, Mustafá, put the clerical leader back in the political spotlight while fueling the first truly militant anti-Shah demonstrations (Taheri 1986). By the summer of 1978, continuing protests, bolstered by the government’s cut in spending and resulting increases in unemployment, had further weakened the Shah, undercutting the value of his appeals to the citizenry and moderate members of the clergy like Ayatollah Shariatmadari (Kurzman 2004). A general strike in early October succeeded in destabilizing the country’s economy and the regime, leading the Shah to establish a military government that would have little hope of restoring order. Events climaxed during the month of December’s Shiite celebrations of Ashura\textsuperscript{33} when at least two million protesters came out onto the streets to commemorate Imam Husayn’s martyrdom and call for the end of the Pahlavi monarchy. Many clerics participated openly in the protests, successfully mobilizing large swaths of the Iranian population through the potent use of Shiite religious symbols and discourse. Following a number of futile attempts to incorporate liberal opponents into a new government, and rejections of any compromise by opposition leaders, the Shah and the Empress left Iran on January 16, 1979, never to return. Thus, the stage was set for the triumphal return of Ayatollah Khomeini on February 1\textsuperscript{st}, in front of a welcoming crowd composed of over one million Iranians.

Khomeini and his cleric-led movement had achieved hegemony over the vast array of oppositional factions. Indeed, some scholars go so far as to claim that he had, by this point, already become the undisputed leader of the revolution (Taheri 1986: 144–146). This is by no means to say that I underestimate the contributions of secular intellectuals, student organizations, traditional left-wing political parties, and non-religious actors more generally to the oppositional movement. Yet, as some scholars have done, it is important to recognize and emphasize the ways in which these marginalized groups lacked organizational cohesion, a coherent political program, and charismatic leaders who could mobilize large swaths of the Iranian population (Abrahamian 1982; Moaddel 1986; Zubaida 2009). Consequently, by early 1978, the Shiite religious clergy had taken the reigns of large swaths of the movement hostile to the Shah. They employed and appropriated an innovative and dynamic blend of ideological tendencies—nationalism, populism, republicanism, and Shiite religious discourse, among others—in an attempt to monopolize the leadership of what were undoubtedly diverse oppositional factions. Supported by well-calibrated populism and radicalism, Khomeini and his followers were able to dominate the wide array of forces eager to overthrow the Shah. As Zubaida reminds us, “it was that political conjuncture which raised Khomeinist Islam to the leading position in the revolution, providing common symbols and slogans for diverse forces” (2009, 81).

In total contrast, and as an interesting counter-factual, Iraqi mujtahids had been largely bereft of power, both social and coercive, since the inception of the Sharifian monarchy in 1920. They were even more suppressed under the Ba’athist regimes beginning in 1968. With little independent economic base or intimate ties to important social classes, the Iraqi religious elites were either co-opted by the state (resulting in a position of political quietism, the route most often taken), sent into exile, or killed (as in the case of

---

\textsuperscript{32} As proof of their heterogeneous ties to different social actors, the unity of the numerous ulama became fragile almost exactly following the downfall of the Shah and Khomeini’s monopolization of power. Key issues over which the ulama clashed included: land reform, labor laws, nationalization of foreign companies and the roles of the public and private sector (Akhavi 1980, 96–103; Moaddel 1986, 542–546).

\textsuperscript{33} Celebrated during the Islamic month of Muharram, Ashura is commemorated by Shiite Muslims as a day of mourning for the martyrdom of Husayn ibn Ali, the grandson of the Prophet Muhammad.
Muhammad Baqir al-Sadr and Muhammad Sadeq al-Sadr, among others). Iraqi governments accumulated the coercive means necessary to crush numerous independent social actors, such as non-Sunni merchants, on whom they did not need to depend for tax revenues.

The shah of Iran was by no means bereft of coercive power. Alliances with the United States and Israel along with increasing oil revenues provided him with ample space to purchase the most novel weaponry of the time. Indeed, it could be convincingly argued that the shah’s accumulation of coercive means prolonged his rule for at least an extra two to five years. Nevertheless, the social position of the ulama, their alliance with other important class actors, and the political conjuncture that furthered their emergence at the head of the oppositional movement made any reliance on pure coercion all but impossible as a long-term strategy of survival for the tyrant.

After a careful historical examination, it is increasingly clear why, as a result of a distinct set of government policies and social conditions, the Iranian ulama had the ability to assert themselves in the political arena at a particular historical moment in the late 1970s. The Islamic character of the Iranian Revolution did not arise solely from the inherent religious outlook or opinions of particular social classes but rather from the “peculiar historical position of the religious institutions in Iran,” which were themselves a product of social structures and the state’s use of coercion and capital (Zaubaida 2009: 64–81). Khomeini’s whole notion of governance, embodied in the Shiite theological-political doctrine of vilayat-i faqih, is indeed a result of conditions particular to Iranian history. After a disillusioning attempt to institutionalize a role for the ulama in the legislative process during the Constitutional Revolution of 1905–1911, Khomeini had historical and political reasons to be suspicious of both constitutional guarantees for oversight and the real weight of legislative power (Keddie 1983, 579–598). He thus introduced a theory of government that would for the first time provide for ulama control of the executive branch while acknowledging, albeit hypocritically, the long-standing democratic pretensions of the Iranian population (hence the title Islamic Republic). While repressed by the Shah and engaged in competition with secular nationalists and radical leftists for public support, Khomeini initially decided to promote a notion of government ambiguous enough to engender cross-cutting societal support against the Shah and popular enough to succeed in a referendum on the establishment of an Islamic Republic. On March 30, 1979, the referendum successfully secured the support of 98.2 percent of the people.

While most popular accounts portray Khomeini as a religious fanatic, the theory of governance underlying the Islamic Republic and its practical application over the past thirty years contradict this myth. It is true that vilayat-i faqih was revolutionary as a religious doctrine in Shiism. Yet Khomeini followed a long tradition of Shiite scholars who had proven to be ideologically malleable, intellectually flexible, and concretely responsive to the political, economic, and social milieu that surround them. Contra Hegelian-inspired notions of ideological and historical development, it is increasingly clear that Shiite religious doctrines were far from emerging or operating in a historical, social, or economic vacuum. Weber’s “disenchantment” was neither followed the onset of ‘modernity’ nor eroded the power of so-called “traditional” forms of authority.

**RELIGIOUS IDEOLOGY IN RELATION TO COERCION, CAPITAL, AND THE SOCIAL STRUCTURE**

In focusing on the ulama’s role in society, I have attempted to challenge those theorists who view theological or ideological developments as independent of political, economic and social realities. I have also tried to push against the cultural essentialism that has recently reemerged to explain the persistence of religious actors compatibility with Islamic law. Indeed, the provision was never put into practice and, even though the original framers of the constitution designed for real power to reside in the parliament, the legislative branch was never quite able to wrest power from the Shah.
and ideologies in Muslim countries. This study argues that Moore and Tilly’s theoretical frameworks are relevant to the Middle East and challenges those who identify religious exceptionalism as a viable explanatory variable in events such as the Iranian Revolution of 1978-79.

Although he ulama defended their divergent political positions through rich interpretations of Islamic law, it would be unfair to claim that they consciously interpreted these religious texts with the interests of certain class actors in mind. Although values, even religious ones, may be strongly influenced by material interests, individuals and groups may ultimately come to believe in the righteousness, internal logic, and specific discourse inherent in their value systems. For this reason, Shiite theological interpretations must be based on a shared set of ideological premises that place several limits on the range of positions that ulama may consciously defend. Eschewing approaches to comparative history that privilege uniform or universal paths to ‘modernization,’ this work shines light on the importance of studying the position of intellectuals or class representatives in relation to coercion, capital, and the social structure, especially during the state-building process. Relative economic independence and freedom from the costs of coercion of alienated social classes permitted the stridency of members in a heterogeneous clerical class, in which each mujtahid and his influence was disparately affected by these factors. The Iranian Revolution should be seen as an attempt by a variety of domestic social actors to institutionalize their power and to develop an equitable society consistent with Iranian nationalism, Shiite religious identity, and democratic aspirations.

Works Cited


35 Meaddel (1986, 546–547) asserts that the nature of Shiite religious discourse precludes any meaningful, long-lasting alliance between the ulama and the proletariat or the peasants. The hadiths and Qur’anic verses supporting private property and the right to profit make alliances with these classes difficult and extremely rare although not impossible.
The rise of the American education finance reform movement began after a series of state supreme court cases challenged the constitutionality of school funding systems and the inter-district inequalities they propagated. Since its inception, the reform movement has focused on improving educational quality for needy students across the United States—a movement that eventually lead to the practice of targeting state educational aid to individual school districts on the basis of demonstrated need. Aid-targeting targeting policies have been met with controversy, however, since financial redistribution tends to face opposition in parts of the country with greater financial and political capital.

Outside of arguments for and against targeting policies, literature related to the education reform movement rarely explores the extent to which states truly target their aid to school districts and under what conditions (either at the local or state level) they do so. The study aims to reveal patterns by which states distribute funds to school districts. While results from this study address a number of related questions, its central focus is to assess the extent to which U.S. states—relative to how they have in the past—distribute their education aid on the basis of school district enrollment and district poverty. Other queries are addressed in this study. If states are not targeting their funds on the basis of school district enrollment, on what basis are they doing so? Do certain conditions or characteristics determine the amount of aid school districts receive? Can the presence or absence of some statewide characteristics—like which political party controls the state or whether the state supreme court has issued a ruling on the state's education system—make a state

Bill DeBaun is a 2009 graduate from American University, where he majored in political science and completed coursework in education studies. He is currently working as a Research Assistant with Westat.
Bill DeBaun is a 2009 graduate from American University, where he majored in political science and completed coursework in education studies. He is currently working as a Research Assistant with Westat.

The rise of the American education finance reform movement began after a series of state supreme court cases challenged the constitutionality of school funding systems and the inter-district inequalities they propagated. Since its inception, the reform movement has focused on improving educational quality for needy students across the United States—a movement that eventually lead to the practice of targeting state educational aid to individual school districts on the basis of demonstrated need. Aid-targeting targeting policies have been met with controversy, however, since financial redistribution tends to face opposition in parts of the country with greater financial and political capital.

Outside of arguments for and against targeting policies, literature related to the education reform movement rarely explores the extent to which states truly target their aid to school districts and under what conditions (either at the local or state level) they do so. The study aims to reveal patterns by which states distribute funds to school districts. While results from this study address a number of related questions, its central focus is to assess the extent to which U.S. states—relative to how they have in the past—distribute their education aid on the basis of school district enrollment and district poverty. Other queries are addressed in this study. If states are not targeting their funds on the basis of school district enrollment, on what basis are they doing so? Do certain conditions or characteristics determine the amount of aid school districts receive? Can the presence or absence of some statewide characteristics—like which political party controls the state or whether the state supreme court has issued a ruling on the state’s education system—make a state.
more or less likely to engage in targeting practices? Some of these questions have been considered in the scholarly literature, but the age of most studies hinders their contemporary relevance.

This study investigates the above questions to determine states’ current practices of targeting education aid to school districts. The rest of this paper proceeds in the following order. Section II outlines previous research related to the educational finance reform movement. Section III outlines data and empirical model used in the study. Section IV presents the study’s main findings, and Section V concludes.

EXISTING RESEARCH

Most states’ school districts fund their schools predominantly through the use of local property taxes (Sadker and Zittleman 2007). The considerable variation from district to district in terms of local property values—and, subsequently, tax revenues—causes schools’ resources to reflect their communities’ relative wealth (Biddle and Berliner 2002). This finance system, which creates a wide range of resources in school districts, has produced a movement whose aim is to equalize school resources by examining and reforming the practice of financing schools with property taxes (West and Peterson 2007). This movement is notable for a few reasons. First, local responsibility for funding schools is a standard that dates back to the earliest colonial classrooms; changing this system would represent a significant break from the fabric of the American educational system (Sadker and Zittleman 2007, 235). Second, the movement to equalize school district resources makes a tacit presumption that school funding and school quality are linked; both sides of this issue will be examined later. Third, how states are distributing their money is important because nationwide, state education expenditures account for approximately fifty percent of the dollars spent on education annually in the United States.1 Finally, one of the proposed mechanisms for equalizing school finance is targeting resources to disadvantaged districts. This in turn has raised questions about both the constitutionality of using property taxes to fund education and whether education is a constitutionally guaranteed right—questions that were most notably raised in the U.S. Supreme Court case Serrano v. Priest. The degree to which the practice of targeting resources to disadvantaged districts has been initiated by the states, and on what basis, is the focus of this study.

A wide body of literature examines the determinants of education finance reform. Topics include the history of this movement, whether or not a school district’s resources affect the education its students receive, and the effects of state intervention on the local level. Education finance issues are complex, interrelated, and continuously developing. While financial disparities between school districts are evident and commonplace throughout the United States, there is contradictory evidence in the literature that these disparities have any effect on student achievement or education quality. This contradiction makes it difficult to establish consensus on what course education finance reform should take.

Tiebout (1956), in his seminal work, “A Pure Theory of Local Expenditures,” described an economic model where every community provides services in accordance with its citizens’ demands and taxes them accordingly. Tiebout’s model is often cited in education finance research as an explanation of inter-district resource inequality. However, his model assumes that citizens at the margin are fully mobile and can vote with their feet and move to a community that satisfies their demands. The reality is that citizens in the poorest communities do not have the resources to meet the market conditions for obtaining better education. Left without options, many poor families are forced to stay in areas where their children’s educational needs may not be adequately met. This reality of inadequate schooling for some students was the impetus for the school finance reform movement. The legal claims and challenges that have driven and continue to drive this movement mostly push for state interventions to provide for more equal inter-district resources in order to counter achievement disparities. However, the effects of

---

1 The National Center for Educational Statistics reports that in FY2002, 49.3 percent of the average school district’s funding came from state sources. In FY2002, approximately $420 billion was spent on education nationwide.
district resource equalization are a topic for which the literature provides contradictory evidence.

Inter-district Resource Inequalities and Possible Effects on Student Achievement

In education, production function research can be used to examine the connection between a school’s resources (inputs), financial or otherwise, and the educational achievement of that school’s students (output). The first systemic evidence on this connection was the 1966 Coleman Report “Equality of Educational Opportunity,” which found that minority students in the United States both achieved less than their white peers and tended to be more affected by school quality. The Coleman Report also used wide-ranging data on race, achievement, and school resources to conclude that minority teachers and students were isolated away from white students into minority-majority enclaves within the educational system and relegated to schools and districts with fewer resources (Coleman et al. 1966).

The literature in the production function field since the Coleman Report has often been inconclusive, able to be divided into two sets of findings. The first and most common finding is that a school’s resources positively influence student achievement. Card and Krueger (1992) and Payne and Biddle (1999) convey this principle. They found that the higher-quality education provided by schools with adequate resources yielded higher student achievement or economic returns later in life or both. Payne and Biddle succinctly summarized their conclusions by saying, “surely it is time to put to rest the absurd myth that level of funding does not matter for public schools in America” (Payne and Biddle 1999, 12). The second finding in production function research is that a school’s resources have very minimal or no effect on student achievement. Hanushek (1986, 1989) provided examples of this “no-connection” school. He wrote, “detailed research...provides strong and consistent evidence that expenditures are not systematically related to student achievement” (Hanushek 1989, 49). Instead, Hanushek (1986, 1989) and Wenglinsky (1998) argue that contextual factors like race, parents’ education, and socioeconomic status are driving achievement disparities. Socioeconomic status has an effect both on where individuals are able to live and the quality of schools that they have access to. Additionally, historical racial achievement gaps create parental education disparities that are then passed on to subsequent generations of children (Roscigno and Ainsworth-Darnell 1999; Duncan and Magnuson 2005). The idea that out-of-school contexts also affect student achievement is one that dates back to the Coleman Report, which indicated that minority students seemed to be more affected by the neighborhoods surrounding their schools than white students. Proponents of school-finance equalization have subscribed to the belief that a school’s financial resources do affect the achievement of its students and that more money toward disadvantaged students can overcome the obstacles of poverty and parental education gaps. This belief has led proponents of reform to pursue redress through both legislative and judicial means.

State Legislative and Judicial Interventions in Education Finance

The education finance reform movement, starting with the Serrano case in California, took root and grew in a number of states, and continues to grow in other states. Many legislators and courts see education finance reform as a means of addressing educational inequality. However, targeting policies remain both controversial in the literature and, in many states, unimplemented.

The 1971 Serrano v. Priest case in California started the wave of school finance litigation that would change the face of American education expenditure. Litigants in these cases, charging that the education being offered in their schools was inadequate, unequal,
or in violation of a state’s constitutional guarantees, sued their local school districts and, in some cases, the state education agency, for redress of the educational conditions. Twenty-six state supreme courts have ruled that their state education finance systems are unconstitutional. In these states, changes in school funding formulas were most often enacted. In the states whose finance systems have not been overturned by the courts or where no judicial challenge has been raised, legislatures have often taken up school finance reform. Evans et al. (1997) and Murray et al. (1998) examined school finance litigation’s effects on both legislation implementation and per pupil expenditures. They found that when state courts ruled local tax revenue-based education finance policies unconstitutional, states shouldered a larger percentage of school funding. After such a ruling, states were also more likely to redistribute funds from richer to poorer districts. They also found that states that tried solely to legislatively mandate increased expenditures for disadvantaged districts or expenditure equity between districts were less likely to achieve these ends than states that also had a judicial mandate for reform. This finding may be attributed to the fact that a judicial mandate acts as a grim trigger and forces a state legislature to act on an issue that it otherwise would not have to; without the backing of a judicial mandate, a legislature may be less politically willing to reform their school finance system. Additionally, a judicial mandate provides cover for politicians in state legislatures; by blaming the court decision, legislators can deflect responsibility for resource redistribution away from themselves.

Mintrom (1993) discussed an area of education finance reform that other authors had not. He examined the political obstacles that stall the implementation of targeting policies. Redistributive targeting generally involves taking a percentage of tax revenue from each relatively wealthy school district and then using those funds to offer supplements proportional to other districts’ financial need. Mintrom said that impoverished municipalities are faced with a tough choice: “tax their poor constituents at higher rates or devote fewer resources to education than rich districts” (Mintrom 1993, 847). He also noted that the political economy in a state rarely makes targeting policies politically popular, and this in turn reduces their political viability. Mintrom also discussed two other local-level obstacles to targeting. The first is that parents are generally concerned with getting their children the best education possible; for parents in wealthy districts, redistributing money to poorer schools, and thus away from their children, is not viewed favorably. The second obstacle to targeting is that because communities and individuals with resources predominantly hold political control and influence, districts with fewer resources lack the political clout to push for finance reforms. Kozol (1991) also wrote extensively about this dilemma using case studies, which revealed that parents, students, and administrators in wealthy districts were unwilling to see their resources go elsewhere. These discussions of the political obstructions that impede school finance reforms in state legislatures lend credence to the Murray, Evans, and Schwab studies that found that school finance reforms without judicial weight behind them were less successful than those with them.

Arguments For and Against Targeting Policies

Redistributive targeting practices sprang up in many states following their supreme court’s mandating new education funding formulas. These practices are supported by various pieces of education finance literature. De Bartolome (1997) and Murray, Evans, and Schwab (1998) found that targeting policies and state supreme court cases have led to an increase in overall state funding for education but that this additional financial support has been offset by increased taxes. De Bartolome—by controlling for other variables—found that between 1970 and 1990, state court decisions accounted for 18 percent of state education aid increases. His findings lend support to Murray, Evans, and Schwab’s research that state supreme court cases can increase a state’s overall level of education funding.

Fernandez and Rogerson (1996, 1997) are the leading voices in the “rising tide” school of thought that supports targeting practices by individual states. They found that “a reallocation
of resources from resource-rich children to resource-poor children will mean greater future aggregate income for society as a whole,” and counter criticism of redistribution by noting that “the increase in future earnings for the children who gain resources more than compensates for the lost future earnings from those children who lose resources” (Fernandez and Rogerson 1997, 69). While Bourguignon and Rogers (2007) found this benefit as well, they added a call for more progressive taxation and argued for more efficiency in how education resources are used. Specifically, they called for “policies that improve incentives, information, and accountability [that] can raise the quality of schooling” (Bourguignon and Rogers 2007, 744).

Supporters of targeting policies highlight the rise in aggregate societal wealth stemming from increased education funds. But while there is a large body of literature showing support for targeting policies, there is an equally large body opposing them. The argument against targeting policies consists predominantly of findings that increased state intervention in local affairs causes apathy, inefficiency, and even decreases in student achievement. Berger and Toma (1994) conducted a nationwide examination of SAT achievement data. They found that students in states that required master’s degrees for teacher certification (used in this study as an indicator of increased state education regulation) generally perform worse on the SAT than their peers in states with fewer teacher certification regulations. Husted and Kenny (2000) built from this research in their study that used 34 states’ SAT achievement data as the dependent variable and the amount of state-level school funding as the independent variable. Their findings are in the same vein as Berger and Toma’s: increased state intervention (increased state share of school funding) leads to inefficiency, indicated by lower SAT scores. The authors also found that state equalization of education expenditures lead to decreases in overall educational achievement but did not change the distribution of achievement. Husted and Kenny found that overall when states mandated equal inter-district spending all schools performed less well, with poor school districts performing the worst. Husted and Kenny also noted that an increased state role in local education leads local constituents to hold schools less accountable for their achievement. This causes inefficiency that is also addressed in the literature.

Fischel (2002) offers a number of arguments against state involvement in the education system. He contended that redistributive policies can lead to “taxpayer revolts” like California’s Proposition 13 in 1978. Furthermore, Fischel notes that state intervention in local schooling can lead to apathy or revolts at the local level as citizens become increasingly disconnected from the funding process and feel unable to influence it. On top of this, Fischel writes: “the measurable educational outcomes have either declined or not changed. No social science study persuasively connects the school finance litigation with better outcomes for children from disadvantaged homes” (Fischel 2002 115). In a particularly damning conclusion, he summarizes by saying “at its worst, school finance litigation has engendered dog-in-a-manger equality of low-quality education. At its best, it seems to have done little more than shift tax burdens and property values in ways that offer no systematic benefits to the poor” (Fischel 2002, 115).

To What Extent are States Employing Targeting Practices?

While there is an abundance of literature drawing contradictory conclusions on the value of implementing targeting practices, there is a dearth of research about the extent of state targeting. Morgan and Pelissero (1989, 1992) are the leading scholars in this field and offer the most comprehensive research on state-level targeting practices. The authors found in their studies that the distribution of aid, either by the federal government or the states, was tied most closely to school district enrollment and not to school district resources. They also found that states were more effective targeters

---

3 Taxpayer revolts like Proposition 13 are instances where states and communities choosing to pay less for education overall in an effort to keep resources from being redistributed to more impoverished districts. This is a hurdle to the implementation of targeting policies that Mintrom (1998) also mentions.
than the federal government. According to their research, which examined school districts in forty-four states, the non-enrollment-based aid allocations that occurred were often targeted to districts in financial need. However, there was a lot of interstate variation in targeting practices. Morgan and Pelissero identified nine states as simply “better targeters.” The authors admit that more research needs to be conducted to find out what policies and practices make these states more efficient and willing to target than others. Still, their research fails to address whether a state’s school finance formula or other state-level variables affect a state’s targeting practices. Another weakness of these two studies is that they were both based on data from 1982. Since that data comes only a decade after the inception of the push for education finance reform, it does not provide a contemporary view of targeting by the states.

Some Evidence of What Affects School Targeting Practices

Overall, the literature on what factors affect targeting practices is scant. It appears that only Figlio, Husted and Kenny’s (2004) research utilized multivariate regression to see what variables affected the distribution of education expenditures. They found that, in accordance with the Tiebout model, greater within-state wealth and a larger number of school districts lead to more disparities in wealth between districts. They also found that Democratically controlled states had more spending equality because of that party’s tendency to propose and support policies addressing the needs of the poor. They found significant evidence that a state supreme court ruling declaring a state’s education finance system unconstitutional caused a decrease in school funding inequality. Finally, they found that more heterogeneous populations in a state contributed to school district financial inequality; this finding also supports the Tiebout-sorting model, which describes how diverse school districts cater to a wide spectrum of demands for school quality and level of taxation.

Other research in this area has been narrowly tailored. An example is Fletcher and Kenny’s (2008) research on the effect the elderly have on school spending. They held that a drop in school funding in areas with large elderly populations was generally offset by a per pupil funding increase derived from the decreased number of children in these same areas. De Bartolome (1997) found in his research that state education aid increases were attributable to increasing amounts of wealth disparity within a state.

Evidence that the elderly affect school finance exists in the literature in a number of places. Poterba (1997) concluded that increasing percentages of the elderly resulted in lower per-pupil spending, especially when the elderly and the students were from different racial groups. Harris, Evans, and Schwab (2001) discovered a small but significant negative effect on funding at the district level that was caused by the elderly; this effect was larger at the state level. Ladd and Murray (2001) determined that the direct effect on school spending caused by the elderly was indistinguishable from zero but that the elderly may have indirect effects on school spending. Brunner and Balson (2004) surmised that the elderly were generally more supportive of local spending than statewide spending on education. Finally, Fletcher and Kenny (2008) used the median voter model to find that the effect of the elderly on school spending was both negative and significant, but they did note that states and school districts should not be fearful of precipitous drops in funding because of an aging population.

The research in this area, especially Figlio, Husted, and Kenny’s, was a significant contribution to understanding how and why states target. However, many questions remain unanswered about state education aid targeting practices. These questions—and why their answers are important—are discussed below.

STUDY DESIGN

The Need for a Contemporary Study

This paper is partially based on Figlio, Husted, and Kenny’s (2004) study, which addressed some of the glaring gaps in the literature on education state resource inequalities. Redistributive tar-
geting practices are used by the states to address these inequalities. But today—more than twenty-five years after the period in which Morgan and Pelissero’s research was based—to what degree are states targeting aid to school districts on a non-enrollment basis? In states that do target, which characteristics render a district more likely to receive targeted funds? What role does a state’s aid formula have on the degree to which it targets its funds? This paper addresses these questions using variables that have been examined in the literature and others, like student poverty, that are absent. Morgan and Pelissero recognized these things as questions for further research at the end of their 1989 study. By building on Morgan and Pelissero’s research on targeting practices by the states and on Figlio, Husted, and Kenny’s research on the variables targeted by states, this paper will offer insight into how far the education finance reform movement has come and, potentially, how much farther it has to go to deliver on its promises to students in impoverished districts nationwide.

Theory and Explanations

As described in the literature review, education finance reforms that originated because of state supreme court decisions have been changing funding formulas and the way state aid is distributed to school districts. Those that originated from legislative decisions have had less impact. This study will primarily consider the connection between school district enrollment and state aid per pupil in an attempt to determine if states are targeting to school district enrollment at a greater or lesser rate than they were in 1982. In their 1989 and 1992 studies, Morgan and Pelissero found that states were primarily distributing aid to school districts on the basis of enrollment, not financial need or other district characteristics. All of these other district characteristics being equal, however, we would anticipate that the districts with the most students would receive the most state aid; the fact that these other characteristics are not equal is what makes the question of states’ redistributive targeting practices relevant. No other comprehensive study examining the connection between school district enrollment and per pupil state aid has been conducted since the Morgan and Pelissero studies, creating both a gap in the literature and a need for a more contemporary examination of this topic. This study will provide such an examination.

The secondary mission of this study will be to examine determinants of state aid other than school district enrollment. Figlio, Husted, and Kenny (2004) examined a number of these determinants including the black population in a state, the number of school districts per student (where a low number indicates monopoly and the absence of Tiebout choice), the percentage of people living in metropolitan areas, state median and mean income, Democratic party control of a state’s government, the presence of a state supreme court ruling overturning a state’s education finance system, and the degree of a state constitution’s mandate for school funding equity, among other variables. While Morgan and Pelissero’s and Figlio, Husted, and Kenny’s works described how these factors affected the allocation of state educational resources from state to state, neither study provided evidence of whether or not targeting practices had shifted from primarily focusing on school district enrollment to focusing on a district’s financial need. None of the studies discussed earlier included an examination of a state’s funding formula as a determinant of state aid to school districts. Additionally, many of the variables utilized by Figlio, Husted, and Kenny (2004) are based at the state level of analysis. This study will incorporate both district and state level variables.

This study will advance a few hypotheses: When comparing school districts, enrollment, not need, continues to be the main de-

4 It is important to note that school district enrollment, in addition to being an important determinant of state aid, is politically important as well. Both legislative districts’ boundaries and number of legislators are based on population. Legislative districts with a large number of representatives will have more political capital to push for funding for their school districts. Similarly, governors can exercise their veto power on measures that do not distribute funding to districts where their political base resides.
terminant of state education spending.

- If enrollment is still the predominant determinat of state aid to school districts, states may not be employing redistributive targeting practices that provide additional aid to school districts on the basis of financial need.
- If states’ per student aid to school districts is significantly tied to non-enrollment determinants, the evidence would suggest that there has been a shift toward targeting practices since 1982—the last year that targeting was examined nationally. These determinants include either district levels of need or other characteristics, like racial composition, or both. This finding may represent an attempt by the states to close achievement gaps between different socioeconomic or ethnic groups.

Coding and Measurement of Concepts

Because this study will examine how states target aid, school districts will be the primary unit of analysis. There are approximately twenty thousand school districts in the United States according to the National Center for Education Statistics. This study will not sample school districts that have incomplete data for any of the variables it considers. In order to capture any changes in targeting practices by the states, this study will utilize data from three school years: 1995–1996, 2000–2001, and 2005–2006. This method will provide a longitudinal view that has not been previously examined in the literature.

This study will focus primarily on three variables but will additionally consider others. All of these variables appear in Table 1. Except as otherwise specifically noted these variables are obtained from the Core of Common Data from the National Center for Education Statistics.

Primary Dependent Variable

- Per student state aid (STATEAID): coded as the total amount of aid given to a school district divided by the district’s reported enrollment.

Primary Independent Variables

- School district enrollment (ENROLL): indicates the total number of students from pre-kindergarten to grade twelve in a given district.
- Federal free lunch eligibility (FREELUNCH): represents the number of students in a school district who are eligible for federal free lunch (FREELUNCH) programs; if states are targeting their aid to districts on the basis of financial need,5 I would expect this variable to be positively tied to state aid because higher values of FREELUNCH indicate higher poverty in a district’s students.

Secondary Independent Variables

Each of these variables will provide insight into district-level characteristics through which states gain or lose aid support. If any or all of these characteristics has a significant effect on per student state aid, it will support the idea that states are targeting their aid toward local conditions and needs rather than solely toward school district enrollment.

- Own-source revenue per pupil (OWNSOURCE): represents the total amount of local education revenue divided by a district’s enrollment. This is a measure of a district’s effort, i.e., how much it invests in schools from local revenue. It is unclear how this variable will change in relation to the state aid a district receives; it could have a positive relationship with state aid if a state decides to reward districts for their local effort, or it could have a negative relationship if a state views

5The federal free lunch program is often used as an indicator of student, school, or district poverty status. Some examples include Krueger (1999), Lambin (1996), Caldas (1993), and Caldas and Bankston (1997). Sirin (2005) notes, “School SES is usually measured on the basis of the proportion of students at each school who are eligible for reduced-price or free lunch programs at school during the school year. Students from families with incomes at or below 130 percent of the poverty level are eligible for free meals (419).”
the capacity to raise local revenues as an indicator of a district not needing additional financial aid.

- *District debt per pupil (DISTDEBT)*: coded as a school district’s outstanding debt at the beginning of a given fiscal year divided by the district’s enrollment. If states are targeting their aid on the basis of financial need, DISTDEBT should vary positively with per student aid.

- *District percentages of black (BLACK) and Hispanic (HISPANIC):* these variables are calculated by taking the number of each group of students in a district and dividing it by total district enrollment. These variables will test whether states are directing funds to districts with high percentages of two groups on the lower end of the achievement spectrum.

This study will also examine five state-level variables to see what effect they have on aid distribution.

- **Political party control (PARTY)**: this variable will be constructed by assigning each state a value between 0–3. A coding of 0 indicates that neither chamber of the legislature nor the governorship is controlled by the Democratic Party. Codes of 1, 2, or 3 correspond to the number of aforementioned bodies controlled by the Democratic Party in a given state. This variable will examine if party control at the state level shifts states toward reducing spending inequality. These data will be obtained from the National Council of State Legislatures.

- **The presence of a state supreme court case relating to the state’s school finance system (SUPREMECOURT):** a 0 value will indicate the absence of such a case overturning the state’s school finance system; a 1 value will indicate the presence of such a case. These data will be obtained from the National Center for Education Statistics’ education finance division.

- **Percentage of residents in the state over 65 (ELDERLY):** I anticipate that states with higher percentages in this variable would target less due to a small, but general, decrease in school funding in areas with high proportions of elderly.

- **States’ Gini coefficients for 1999 (GINI):** the Gini coefficient is a measure of a state’s wealth distribution. States with a Gini coefficient closer to one (less equal distribution) are anticipated to distribute aid based on enrollment more than on other factors.

- **Presence of one of two state funding formulas (FORMULA):** the final state-level variable is one that has not been examined in the literature. These data will be obtained from Education Week’s EdCounts.org datasets, as well as from the National Access Network, a project housed at Teachers College, Columbia University. Two different types of funding formulas will be examined: foundation formulas and foundation/local-effort equalization formulas. Foundation formulas are expected to increase the effect of enrollment on state aid because every additional pupil will yield additional funding under the foundation formula. Foundation/local-effort equalization programs may have a similar effect that will be tempered by this formula’s incentives to raise money locally. In this study, a dummy variable will be used for each formula to determine if either corresponds to more aid targeted on a non-enrollment basis. Because no research has examined how the distribution mechanisms for a state’s aid affect its targeting practices, this variable’s incorporation will look for each formula’s relative ability to get state aid to districts on the basis of local characteristics.

---

6 This is a variable that Figlio, Husted, and Kenny (2004) examined; in their study they found that states with Democratic Party control had more spending equality than Republican controlled states because Democrats tend to represent the poorer populations in a state.

7 This variable will be incorporated to re-examine Evans, Murray, and Schwab’s (1997) findings that after such a supreme court decision occurred, states were more likely to redistribute funds from richer districts to poorer.

---

8 The foundation formula is described by EdCounts.org as “a set amount of funding that the state deems necessary for a basic education. [In this formula,] responsibility for providing this funding is shared between state and local districts.” Foundation formulas are most prevalent nationally and are found in 33 states.

9 Local-effort equalization “rewards districts for efforts to raise revenue through additional local taxes” by providing “an additional level of state aid” according to EdCounts.org. This hybrid formula is found in eight states.
help to linearize the data. Using a log-log model of analysis reveals the relative importance of the independent variables on the dependent variable. This method of analysis will answer questions like ‘for every additional percent increase in enrollment in a school district, by what percent does per student state aid increase?’ by creating log-log elasticities. Answering these kinds of questions will show the variables most closely tied to state aid. For example, if a marginal increase in enrollment results in less subsequent per student aid than a marginal increase in another variable, like percentage of black students, this will provide evidence that a state is targeting its aid more toward districts with high percentages of black students than toward districts with high levels of overall enrollment.

Six separate models will be estimated in this study: one model with and one model without state fixed effects for each of three academic years. The variables that will be controlled for in this study are own-source revenue per student, district debt per student, the percentages of black and Hispanic students in a district, the presence of one of two funding formulas in a state, the presence or absence of a state supreme court case overturning the state’s education finance system, the political party in control in the state, the state’s population percentage of people over the age of sixty-five, and the state’s Gini coefficient. All of these variables are included to test if—as a result of the education finance reform movement—the United States has moved away from using school district enrollment as the primary determinant of aid. If this movement exists, the models will identify whether states are now targeting aid to districts with financial need or targeting aid to districts with high populations of traditionally underachieving groups of students.

Morgan and Pelissero’s research (1989, 1992) revealed school district enrollment to be the predominant determinant of state aid. I expect that this study will find that enrollment is still the main determinant of state aid but that it will be less influential than it was in Morgan and Pelissero’s studies. If states have not shifted their targeting practices, I will find a strong relationship between enrollment and per student state aid. This would be evidence that targeting practices are not being employed by the states. If this re-

### Table 1: Coding Variables Used

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>State or District Level</th>
<th>Logged?</th>
<th>Independent/Dependent</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per student State Aid</td>
<td>STATEAID</td>
<td>District</td>
<td>Y</td>
<td>Dependent</td>
<td>Amount of Gross State Aid/ENROLL</td>
</tr>
<tr>
<td>School District Enrollment</td>
<td>ENROLL</td>
<td>District</td>
<td>Y</td>
<td>Independent</td>
<td>Total number of students K-12 in a district</td>
</tr>
<tr>
<td>Number of Free Lunch</td>
<td>FREELUNCH</td>
<td>District</td>
<td>Y</td>
<td>Independent</td>
<td>Number of Free Lunch Eligible Students</td>
</tr>
<tr>
<td>Own-Source Revenue Per Pupil</td>
<td>OWNSOURCE</td>
<td>District</td>
<td>Y</td>
<td>Independent</td>
<td>School revenue raised locally in FY/ENROLL</td>
</tr>
<tr>
<td>District Debt Per Pupil</td>
<td>DISTDEBT</td>
<td>District</td>
<td>Y</td>
<td>Independent</td>
<td>District debt at beginning of given FY/ENROLL</td>
</tr>
<tr>
<td>Percentage of Black Students</td>
<td>BLACK</td>
<td>District</td>
<td>N</td>
<td>Independent</td>
<td>Number of Black Students in District/ENROLL</td>
</tr>
<tr>
<td>Percentage of Hispanic Students</td>
<td>HISPANIC</td>
<td>District</td>
<td>N</td>
<td>Independent</td>
<td>Number of Hispanic Students/ENROLL</td>
</tr>
<tr>
<td>Political Party Control</td>
<td>PARTY</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>Scale 0-1 based on party control of legislative chambers and government</td>
</tr>
<tr>
<td>Supreme Court Ruling</td>
<td>SUPREMECOURT</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>Code 0 if no ruling exists, 1 if ruling exists</td>
</tr>
<tr>
<td>Presence of Foundation</td>
<td>FOUNDATION</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>1 if state has a foundation system, 0 otherwise</td>
</tr>
<tr>
<td>Presence of Local-Effect</td>
<td>FOUNDATIONLE</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>1 if state has a foundation/local-effect equilibrium system, 0 otherwise</td>
</tr>
<tr>
<td>Percentage of Elderly in State</td>
<td>ELDERLY</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>Number of people 65+ in state/state population</td>
</tr>
<tr>
<td>State’s 1999 Gini Coefficient</td>
<td>GINI</td>
<td>State</td>
<td>N</td>
<td>Independent</td>
<td>State’s 1999 Gini Coefficient</td>
</tr>
</tbody>
</table>

**Research Design**

This study will primarily investigate, through the use of ordinary least-squares (OLS) regressions, the relationship between school district enrollment and free lunch eligibility, the two main independent variables, and per student state aid, the dependent variable. In order to examine the relative importance of these variables, they will be converted into logarithmic measures. This will
relationship has weakened, I will find that non-enrollment characteristics in school districts have partially or completely supplanted enrollment as the predominant determinant of state aid. This would be evidence that redistributive targeting practices are performing as intended in an attempt to reduce inter-district resource disparities. If neither of these results is found, it will mean that enrollment is no longer the predominant determinant of aid but neither are any of the other variables that this study examines. Any of these three results would raise important questions for education finance reform policies in the United States.

FINDINGS

This study yields a number of interesting and, in some cases, unexpected results. Before discussing these, I have a few notes. First, though a district’s per student debt had been an intended variable, the data set from the NCES was incomplete and greatly reduced the number of observations (from 8,000-11,000 to under 2,000). Because FREELUNCH is already an indicator of poverty in a district, DISTDEBT was dropped to allow for more observations. Second, this study’s intention was to replicate Morgan and Pelissero’s 1989 and 1992 work as a model. This replication permitted a valuable comparison, on a one-to-one basis, of the results from Morgan and Pelissero’s 1982 research with data from one to two decades later. However, replication became impossible because I dropped DISTDEBT from my model. Additionally, Morgan and Pelissero’s 1992 research does not combine enrollment and other factors (i.e., race, district poverty) into one model but instead keep those variables separate. Their model examines enrollment used total state aid, not per student state aid, as the dependent variable. This method makes it impossible to compare on a one-to-one basis the models in this study and those found in Morgan and Pelissero (1992). Their 1989 study, however, provides a valuable regression; and it is with this regression that the analysis of this study’s results begins.

A look at the state of targeting from an earlier period is in order to see this study’s results about contemporary state practices. Pelissero and Morgan (1989) offered the following regression based on data from 1982:

\[
\text{State aid per pupil} = 1,810 + (0.01 \times \text{Enrollment}) + (-.17 \times \text{per student Income}) + (0.41 \times \text{Own-Source Revenue Per Pupil}) + (0.66 \times \text{District Debt Per Pupil})
\]

\[(R^2 = .11) \ (N=13,654)\]

This regression indicates a positive relationship between enrollment and state aid. All other factors in school districts being equal, a theoretical expectation is that school districts with the largest enrollment would receive the largest amounts of state aid. All other factors in school districts, however, are not equal, which led Pelissero and Morgan to examine other variables in their model. The negative relationship between per student income (the authors’ district wealth indicator) is potentially explained as states being disinclined to target money to wealthier districts. In 1982, it also appears that states were apt to reward districts for effort based on the $0.41 coefficient on own-source revenue. Pelissero and Morgan’s last variable, district debt per pupil, has a positive relationship that fits the theoretical expectation of states targeting aid to impoverished districts that carry more debt. Examining Pelissero and Morgan’s model is a valuable contrast to this study’s results.

Table 2: Effect of Enrollment on State Aid

<table>
<thead>
<tr>
<th>Variable</th>
<th>1995</th>
<th>2001</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Enroll.} \times 10^{-3}$</td>
<td>0.253&lt;sup&gt;**&lt;/sup&gt;</td>
<td>0.019&lt;sup&gt;**&lt;/sup&gt;</td>
<td>0.012&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>$\text{Income} \times 10^{-3}$</td>
<td>7.53&lt;sup&gt;**&lt;/sup&gt;</td>
<td>8.53&lt;sup&gt;**&lt;/sup&gt;</td>
<td>8.35&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>$\text{Debt} \times 10^{-3}$</td>
<td>1.91&lt;sup&gt;**&lt;/sup&gt;</td>
<td>1.70&lt;sup&gt;**&lt;/sup&gt;</td>
<td>0.81&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Enrollment was arguably the most important variable in this study.
because state aid has been so closely tied to it in the past. First, the regression of the log of per student state aid on the log of district enrollment without any other independent variables was examined—the results of which are reported in Table 2.11 In both 1995 and 2000, the regression estimates positive coefficients (0.025 and 0.031 respectively) that are in line with Morgan and Pelissero’s results; these support the assertion that enrollment is at least partially driving the amount of state aid a district receives. However, in 2005, the coefficient changes to −0.013, which would indicate that larger district enrollments do not drive increased levels of state aid. All of these values were significant at the $p < 0.01$ level.

Second, six different regressions were run involving the full model: both with and without state fixed effects for each of the three academic years examined. The results of the extended model are presented in Table 3. More surprising than the coefficient change on enrollment in the 2005 partial model was that in all six full models, school district enrollment varied negatively with per student state aid. Five of these coefficients were significant at the $p < 0.01$ level. Only the coefficient in the non-fixed effects 2000–2001 model returned insignificant results. The implication of this unexpected finding is that enrollment no longer drives allocated state education aid.

If increased enrollment is not yielding additional state aid for school districts, what factors are? This study’s hypothesis was that if states are not targeting aid on the basis of enrollment, the states might instead be targeting to districts that are impoverished or have large populations of students from historically underachieving groups (in this study, black and Hispanic students). The results here were mostly supportive of that hypothesis. The free lunch eligibility variable, which was expected to vary positively with aid, did so in five out of the six models. However, the coefficients were positive and significant in only three of the six models. This result is important because it shows that state aid is being targeted to factors other than enrollment.

Results from the two variables measuring percentages of minority students in districts were mixed both directionally and in terms of significance. Though percentage of black students should vary positively with state aid if states are targeting to districts with large African-American populations, it did so (and was statistically significant) only in the year 2000 model with state fixed effects. The percentage of black students varied negatively and significantly in the models without fixed effects from all three years; however, the results from models without fixed effects are unlikely to be as valid as the estimates that include fixed effects. The percentage of Hispanic students varied positively in five models (three of which were statistically significant and included fixed effects), but it also had a statistically significant negative coefficient in the 2005 model without fixed effects. The implications of the findings related to districts’ percentages of black and Hispanic students on states’ targeting practices are many, and they will be discussed in-depth in the next section.

With respect to own-source revenue per pupil, states could potentially react in one of two ways. States could either reward districts for their relative local financing effort or take funding away because of perceived affluence or self-sufficiency. In 1982, own-source revenue varied positively with state aid, but in all three years examined here, own-source revenue per pupil varied negatively with state aid. The coefficients on own-source revenue per student are consistently negative, and they are significant in four of the six models. This finding indicates that nationwide, states’ attitudes have shifted away from rewarding local financing effort and instead have cut funding from districts that were able to raise money locally. The problem that these results present is that if either of these postulates were true, districts would be expected to begin modifying their behavior in order to maximize the amount of state aid received. This conclusion introduces the problem of endogeneity because the direction of causality becomes ambiguous; we cannot know if the districts’ behaviors are affecting states’ tendencies to

11 While I performed the regression with both the variables and their logs, only the results of the logged regressions are discussed here. This is because the unlogged regressions did not return significant results and were an inferior fit to the logged regressions.
Table 3: Results Obtained – Full Model

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>1993</th>
<th>EL</th>
<th>2003</th>
<th>EL</th>
<th>2003</th>
<th>EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>logENROLL</td>
<td>0.0065</td>
<td>0.0066</td>
<td>0.0067</td>
<td>0.01071</td>
<td>0.0074</td>
<td></td>
</tr>
<tr>
<td>logFREE LUNCH</td>
<td>0.1029</td>
<td>0.1026</td>
<td>0.1030</td>
<td>0.01074</td>
<td>0.0073</td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td>-0.3230</td>
<td>0.1642</td>
<td>0.1642</td>
<td>0.1642</td>
<td>0.1642</td>
<td></td>
</tr>
<tr>
<td>HISPANIC</td>
<td>0.0557</td>
<td>0.0549</td>
<td>0.0549</td>
<td>0.1549</td>
<td>0.1549</td>
<td></td>
</tr>
<tr>
<td>logPINCOME</td>
<td>0.1444</td>
<td>0.1449</td>
<td>0.1445</td>
<td>0.01071</td>
<td>0.01071</td>
<td></td>
</tr>
<tr>
<td>PARTY</td>
<td>-0.0281</td>
<td>0.0151</td>
<td>0.0151</td>
<td>0.0151</td>
<td>0.0151</td>
<td></td>
</tr>
<tr>
<td>SUPREME COURT</td>
<td>-0.0247</td>
<td>-0.0247</td>
<td>-0.0247</td>
<td>-0.0247</td>
<td>-0.0247</td>
<td></td>
</tr>
<tr>
<td>FOUNDATION</td>
<td>0.3410</td>
<td>0.3410</td>
<td>0.3410</td>
<td>0.1549</td>
<td>0.1549</td>
<td></td>
</tr>
<tr>
<td>FOUNDATION LEE</td>
<td>0.0045</td>
<td>0.0045</td>
<td>0.0045</td>
<td>0.0045</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>ELDERLY</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0071</td>
<td></td>
</tr>
<tr>
<td>GIRL</td>
<td>1.2670</td>
<td>1.2670</td>
<td>1.2670</td>
<td>1.2670</td>
<td>1.2670</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>3.5190</td>
<td>3.5190</td>
<td>3.5190</td>
<td>3.5190</td>
<td>3.5190</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>8224</td>
<td>8224</td>
<td>8224</td>
<td>11201</td>
<td>11201</td>
<td></td>
</tr>
<tr>
<td>N-squared</td>
<td>0.191</td>
<td>0.191</td>
<td>0.191</td>
<td>0.191</td>
<td>0.191</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Results Obtained – Full Model

shift money toward or away from own-source revenue raising or vice versa.

Results from the other state-level control variables were contradictory. The signs and significance levels were not consistent as evidenced by Table 2. These fluctuations made it difficult to draw meaningful or clear results; this problem may have been because this study’s unit of analysis was the school district.12

Discussion

This study’s results support the hypothesis that states are targeting their education aid to districts that need it based on poverty. That practice of states targeting money toward impoverished districts is one that education finance reformers, for decades, have been fighting for judicially and otherwise (West and Peterson 2007). Though enrollment was anticipated to be the predominant determinant of a district’s level of per-student state aid, regressions from the 1995–96, 2000–01, and 2005–06 school years show that enrollment, contrary to findings from previous studies, varies negatively with aid. These same regressions showed that per-student state aid varied positively with both the number of students in the district eligible for free lunch programs and the percentage of Hispanic students in the district.

At first glance, these results appear to represent a victory for advocates of redistributive policies across the country. These results bear some examination, however. First, though a shift away from distributing state aid based on enrollment does represent a trend toward targeting based on district-level characteristics, it is

12 State level variables like FOUNDATION, FOUNDATION LEE, SUPREME COURT, and PARTY proved unwieldy in analyses primarily conducted at the district level. Though a state may have a general type of funding formula, the way that that formula actually operates is often unique to an individual state. Supreme Court decisions are similarly unique, and their impacts are difficult to ascertain through binary variables. Lastly, state party politics are better measured through a system other than a 0–3 scale. GINI was dropped because of collinearity with the fixed effects models. ELDERLY encountered the same inconsistencies as the other control variables. These variables are better suited to a model examining what conditions promote more total state education spending.
important to remember that many of our nation’s most troubled school districts are also the most heavily populated (Biddle and Berliner 2002, 54). For these districts, many of them urban, shifting funds away from enrollment may inadvertently end up taking money away from the same students that advocates of redistribution are trying to help through school finance reform. This is especially true because in none of the six models is the positive free lunch eligibility coefficient larger than a negative district enrollment coefficient which means that the net effect is a decrease in funds to poor districts with high enrollment.

Similarly disconcerting, the coefficients on the percentage of black students variable indicate that districts with higher percentages of African-American students see less state education aid directed toward them. This conclusion is troubling because black students are still achieving less than students of other races and ethnicities. African-American students, who lag the farthest behind other races in terms of the achievement gap, appear to be a financial detriment to districts trying to receive state education aid funds. Federal Title I funds that are directed to high-poverty school districts likely make up some of the decreases caused by the negative coefficients found here on both district enrollment and percentage of black students. Another silver lining in these results is, of course, the mostly positive coefficients seen on the percentage of Hispanic students variable. Though the negative coefficient on the 2005 non-fixed effects model is significant, the three more accurate fixed effects models show significantly positive rates of targeting to Hispanic students.

The amount of unexplained variance in this study’s models lends credence to the idea that there are variables affecting targeting that have not been examined by this study. The largest R-squared obtained here is 0.243, comparatively small when that leaves 75 percent of the variance unexplained. It begs the question of what other factors could be affecting the ways that states distribute their aid. Student achievement, as measured by performance in state standardized assessments, could be driving aid distribution if funds are being allocated to either reward high-achieving districts or help to raise achievement in underperforming schools. Determining what these other factors is a task for future research.

This study expands the literature of education finance targeting. It also updates a previous finding in this area. But more questions remain. First, in a future analysis, more district level variables could be examined to determine more specifically which district-level characteristics are being targeted toward. There are many examples of these to consider. A few are district population density (to see if money is being targeted to rural or urban districts), district percentage of special education students, or district percentage of ELL students. Additionally, achievement variables could be incorporated to determine if districts that are not meeting achievement goals have additional funds targeted to them. Finally, it would be valuable to revisit the federal targeting question in Morgan and Pelissero’s 1992 study. That study showed that there were no discernible targeting practices at the federal level, but since the implementation of NCLB, and with accountability an increasingly important part of the current domestic education culture, it would be interesting to see if districts are getting additional federal aid based on achievement levels. It would be equally interesting to see if the federal government tends to target to other characteristics, or if it is still not engaging in targeting practices. This field is full of potential future inquiries—in large part to the ways that NCLB has affected federal-state, state-local, and federal-local government relationships. With these changing dynamics, it is important to examine how fiscal policy is being affected.

**CONCLUSION**

This study’s results are notable because they differ from those found in the most recent study of state aid targeting. The school finance reform movement appears to have gained ground in its effort to redistribute aid to poor and traditionally underachieving
students. Regressions conducted in this study covering three academic years show with statistical significance that school district enrollment, a historically dominant and positive determinant of where aid is allocated, now varies negatively with per student state aid. These regressions also show that school district poverty, as examined through student eligibility for federal free lunch programs, is being targeted toward nationally as evidenced by significant and positive correlations with aid. Hispanic students are being similarly targeted, but coefficients for populations of black students vary negatively with state aid, which is a surprising finding that merits special attention in future research.

This study is valuable in the context of today’s education system for a few reasons. First, the American economy is in danger of being dragged down by a workforce that is not as educated as it could be. The school districts that need to have educational aid targeted to them the most, because of poverty or populations of students from historically underachieving groups, are often also the school districts that are failing to produce well-educated and well-informed citizens. Fortunately, this study’s results indicate that states are shifting their funding practices to try to ensure that districts that are in need are targeted with more funds. Unfortunately, this funding change is neglecting at least one notable group, black students, who continue to achieve less than other racial groups and this funding change is neglecting at least one notable group, black students.

Finally, this study leaves the door open for new research in a number of directions. First, the low levels of explained variance in these models means there could be other variables affecting targeting and the distribution of state aid that were not considered in this study. Additionally, further examination into state behaviors toward district own-source revenue, NCLB’s effect on federal targeting practices, or the apparent shifting of funds away from black students that is revealed by this study would all be valuable contributions to this area of research. Such research is indispensable for understanding how education revenues can be spent to best aid the students who need the most assistance.

Works Cited

While American dependence on energy imports has been extensively analyzed and critiqued, the United States has done relatively little to help address its European allies’ reliance on imported energy. The European Union (EU) is the “world’s largest importer of oil and gas,” buying eighty-two percent of its oil and fifty-seven percent of its natural gas from abroad. Half of EU oil imports and a third of EU gas imports come from Russia (Cohen 2007, 2). This dependence, in conjunction with an EU energy market that remains divided along national borders and a pattern of bilateral deals between European national energy monopolies and external suppliers, endangers not only cooperation within Europe but also the reliability and independence of European diplomatic and political action with regards to U.S. interests (Noël 2008, 8). Unable to form a common position based around a unified energy market, EU member states must increasingly choose between stable energy supplies and siding with the United States on key diplomatic issues ranging from sanctions on Iran to reciprocal investment rights.

This paper aims to address this issue by analyzing the roots and effects of the disunity of European energy markets, noting the U.S. interest in an integrated, reformed European energy market, outlining policies that mitigate dependence on imported energy, providing two case studies which illustrate the drawbacks of current policy and potential of new proposals, and finally suggesting three major platforms of reform to be pursued jointly by the United States and the European Union. I argue that a European Union featuring a more competitive private energy market operating in the context of robust institutional oversight would enjoy significantly greater

Andrei Stetsenko is a senior at Princeton University, where he focuses on global energy markets at the Woodrow Wilson School of Public and International Affairs.
While American dependence on energy imports has been extensively analyzed and critiqued, the United States has done relatively little to help address its European allies’ reliance on imported energy. The European Union (EU) is the “world’s largest importer of oil and gas,” buying eighty-two percent of its oil and fifty-seven percent of its natural gas from abroad. Half of EU oil imports and a third of EU gas imports come from Russia (Cohen 2007, 2). This dependence, in conjunction with an EU energy market that remains divided along national borders and a pattern of bilateral deals between European national energy monopolies and external suppliers, endangers not only cooperation within Europe but also the reliability and independence of European diplomatic and political action with regards to U.S. interests (Noël 2008, 8). Unable to form a common position based around a unified energy market, EU member states must increasingly choose between stable energy supplies and siding with the United States on key diplomatic issues ranging from sanctions on Iran to reciprocal investment rights.

This paper aims to address this issue by analyzing the roots and effects of the disunity of European energy markets, noting the U.S. interest in an integrated, reformed European energy market, outlining policies that mitigate dependence on imported energy, providing two case studies which illustrate the drawbacks of current policy and potential of new proposals, and finally suggesting three major platforms of reform to be pursued jointly by the United States and the European Union. I argue that a European Union featuring a more competitive private energy market operating in the context of robust institutional oversight would enjoy significantly greater

Andrei Stetsenko is a senior at Princeton University, where he focuses on global energy markets at the Woodrow Wilson School of Public and International Affairs.
commercial and diplomatic leverage – particularly with regards to supplier states such as Russia.

BACKGROUND INFORMATION

EU Energy Vulnerability

Recent European energy crises, including natural gas pricing disputes between Russia and Ukraine in January 2006, March 2008, and January 2009, have highlighted the vulnerability that results from Europe's disunited market (Bely 2009, 3). Slow to respond to crises and its energy transmission infrastructure fragmented, the EU suffers from a “solidarity gap” between its new and old members (Tcherneva 2009, 1). For example, a dispute in June 2007 between Germany and Poland, ostensibly about European parliamentary representation, was also an opportunity for Poland to express frustration at “the perceived reluctance of Berlin to stand up to Moscow on a whole host of issues,” including energy (Baran 2007, 134). ¹ Russian energy firms, aware of this “lack of cohesion,” play their hand to further increase European dependence (Baran 2007, 131). Russia's gas export monopoly Gazprom does this by pursuing lucrative bilateral relationships directly with Western European nations (Leonard and Popescu 2007, 16).

The world's largest gas producer, Gazprom is the keystone of Russia's energy industry (De Leon 2010, 1). Established in 1992 from the foundation of the Soviet Ministry of Gas (Ahrend and Tompso 2005, 802), it produces 70 percent of Russian natural gas and represents virtually all of Russia's gas exports (Roberts 2004, 17). Of the 47 trillion cubic meters of gas reserves in Russia – roughly a quarter of estimated worldwide supply (Nicoll and Delaney 2007, 1)—Gazprom owns licenses to 60 percent (International Energy Agency 2006, 15). Since 2005, 50.002 percent of Gazprom shares have been owned by the Russian state, giving the Kremlin majority control over the corporate board (Heinrich 2006, 8).² The remaining 49.998 percent of shares are owned by a combination of of Russian banks, energy firms, European energy distribution firms, and smaller private investors (OAO Gazprom 2008, 80). Through downstream acquisitions, Gazprom has shifted policy from volume-maximization to profit-maximization, emphasizing added-value income over simple sales of gas at the border (Mitrova 2008, 2–3). Today, over seventy percent of Gazprom's profits are derived from sales of gas terminating in EU states (Cameron 2009, 26).

Gazprom's bilateral contracts with major consumer countries such as Germany, France, Italy, and Austria were renewed in 2006 for periods of between fifteen and twenty-five years (Stern 2007, 3). Both Gazprom and the European energy companies benefit from the “stable flow” of both revenue and energy supplies they provide in the context of an otherwise volatile price environment (Wu and Cavusgil 2006, 87). Both Gazprom and these European firms find the prospect of a reformed European energy market unappealing for many of the same reasons. The Commission envisions a market that is integrated, transparent, and highly competitive. Were the continent's energy market ever to actually become liquid and amalgamated, both Gazprom's and European monopolies' profit spreads would inevitably decrease as competition eroded their ability to capture price differentials (Westphal 2006, 52; Checchi 2009, 1; Finon and Locatelli 2007, 29).³

EU Energy Markets and Infrastructure

Collectively, the twenty-seven member states of the European Union form the world's largest import market for natural gas, consuming 530 billion cubic meters (bcm) of gas in 2008 (Roberts 2004, 2). However, the European Union's large western members—

¹ Certain Polish politicians argued representation should account for Polish population losses during World War II.

² Specifically, the Federal Agency for State Property Management owns 38.373 percent, state oil company OAO Rosneft owns 10.740 percent, and state holding company OAO Rosgazifikatsiya owns 0.889 percent, for a subtotal of 50.002 percent.

³ Western Europe's energy monopolies have countered that their large size enables them “to compete in” a highly competitive “global market place.”
notably France, Germany, and Italy—import large volumes of gas constituting relatively smaller proportions of their overall gas supply, in contrast to eastern members such as Poland and Bulgaria which import lower volumes constituting dramatically higher proportions of their gas supply (Gelb 2007, 2–3). See Table 1 for a visualization of this distinction.

Table 1: Comparative Dependence on Russian Gas (Gelb 2007)

<table>
<thead>
<tr>
<th>Country</th>
<th>Gas imported annually</th>
<th>% of Domestic Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>44.5</td>
<td>29</td>
</tr>
<tr>
<td>Italy</td>
<td>44.2</td>
<td>31</td>
</tr>
<tr>
<td>France</td>
<td>11.5</td>
<td>24</td>
</tr>
<tr>
<td>Poland</td>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>Slovakia</td>
<td>6.4</td>
<td>99</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>2.8</td>
<td>99</td>
</tr>
</tbody>
</table>

The “new” member states of Eastern Europe, which remain tethered to Soviet-era gas distribution systems, are almost three times more dependent on Russian gas than Western European nations such as France, Germany, and Italy (Noël 2008, 1–2; European Commission 2008, 9). Whereas Germany and Italy represent nearly half of total imports of Russian gas, countries including Estonia, Latvia, Lithuania, Bulgaria, Slovakia, and Finland, while proportionally much smaller importers, individually draw nearly one-hundred percent of the gas they use from Russian supplies (Checchi, Behrens, and Egenhofer 2009, 19). These highly dependent eastern members are truly the EU’s weak link in terms of energy security. They “lack modern, well-financed companies that can secure more favorable terms with Russia” and typically do not possess large storage facilities (Smith 2008, 18). In comparison, both France, Germany, and Italy all enjoy significant storage capacity of “20 percent or more of annual consumption” (Westphal 2009, 24).

Figure 1 provides an additional illustration of the variation in European states’ dependence on Russian gas supplies.

Figure 1: Disparities in Dependence

The current European Commissioner for Energy, Guenther Oettinger, confirmed this continental bifurcation as the decisive factor behind the Commission’s failure to implement the unbundling of energy monopolies, which would require the splitting of production and distribution arms of current energy firms and an end to monopoly firms’ domination of national energy markets. Member states such as France and Germany, cognizant of the domestic significance and economic weight of their national monopolies, have, according to Mr. Oettinger and others, consistently demurred on the topic of unbundling, leaving such proposals all but dead in the water (Pop 2010). This policy inaction is invaluable for Gazprom, as the Commission’s unbundling proposal would “effectively bar” the company from exercising any control over the wholesale and retail distribution, or “downstream,” segment of the European energy market (Perovic 2008, 3).

Barring a dramatic change in its energy policy, the European
Union can expect to import eighty percent of its natural gas by 2030 (European Commission 2006, 3). By then, the EU is projected to consume between 700 (Roberts 2004, 2) and 816 bcm (Goknuel 2009) annually, of which only 163 bcm will be produced within the European Economic Area, principally in Norway (International Energy Agency 2002, 189). Even in the extremely unlikely case that the EU’s gas demand does not grow over the next two decades, it will require significantly increased levels of imported gas to compensate for declining internal production (Godzimirski 2009, 4; Checchi, Behrens, and Egenhofer 2009, 14–15).

Within Europe, the rift over how to react to this growing dependence on external energy suppliers has been labeled “the most contentious issue for EU governments since Donald Rumsfeld and the Iraq War” (Leonard and Popescu 2007, 26). Nations such as Germany and Italy have concluded bilateral natural gas deals with Gazprom, for which such large, well-paying markets are essential to maintaining profitability (Cohen 2007, 1). Meanwhile, countries such as Slovakia and Poland (which import smaller amounts of gas) are less diversified in terms of energy supply and thus have suffered the brunt of recent gas shutoffs, including a two-week standoff in January 2009 between Russia and Ukraine, Gazprom’s most important transit country for supplying EU customers. During that crisis, leaders of large Western European EU members, including President Sarkozy, Chancellor Merkel, and Prime Minister Berlusconi, were noticeably silent (Tchernaeva 2009, 1).

The infrastructural isolation of the newer EU member states—particularly Lithuania, Latvia, and Estonia—further complicates Europe’s physical and institutional energy disunity. To an even greater extent than many other former members of the Soviet bloc, the Baltic states represent an “energy island” almost completely cut-off from gas markets in the rest of Europe (European Commission 2006, 6). The Commission’s proposed “solidarity rule,” intended to institutionalize the transfer of gas from less to more affected EU member states in the event of a crisis, would be of little use without the necessary infrastructural connections to countries such as these (Le Coq 2009, 37; Finon and Locatelli 2007, 26–27).

While the EU has delineated in great detail a proposal for improved inter-member state energy infrastructure called the Trans-European Energy Networks (TEN-E) and has even outlined proposed mechanisms to compensate transmission operators for cross-border electricity transfers via Regulation 1228/2003/EC, a lack of necessary “financial and political instruments” has stood in the way of implementation (Westphal 2009, 16). As a result, the Baltic states and Finland, which are not currently connected to the continental gas network, are unable to implement even nominal diversification away from Russian gas (Nyquist, Egenhofer, and Legge 2001, 20).

Background on the EU Internal Politics

Brussels is acutely aware of the price the European Union pays for its disunited energy market. The European Commission, the major EU policymaking body, is conscious of the fact that a “common external dimension to its energy policy” is impossible without an integrated, collaborative market (Youngs 2007, 5). A large part of this problem pertains to a lack of institutional jurisdiction. Energy is primarily the responsibility of the Commission, with its “energy technocrats” working together with national-level energy experts (Youngs 2007, 7). Yet European member states are not empowered to apply Community (EU-wide) law (Council of the European Union 2003, 2); moreover, the European Commission has “no authority” to regulate the most visible manifestation of the current system: the long-term contracts between Gazprom and Europe’s national energy monopolies (Stern 2007, 2).

6 One example of successful European infrastructural diversification occurred outside the scope of EU institutions. In 2006, eight Central European nations, including EU members such as Austria but also non-members including Croatia, met following that year’s Russo-Ukrainian gas crisis and were able to provide the collective political backing for a Hungary-Croatia gas pipeline tying the region to a planned LNG deliquefaction terminal in Adria, Croatia.

7 The Council of the European Union, meanwhile, takes a more political role.
In the context of a liberalized Union-wide market in almost every economic sector, with the prominent exception of agriculture, energy policy remains a field viewed in Paris, Rome, and Berlin as a “national prerogative.” The old member states have been “extremely reluctant to cede any sovereignty to the EU in this sphere” (Mangott and Westphal 2008, 169). In particular, the western European energy monopolies (most notably France’s GDF Suez, Germany’s E.ON Ruhrgas, and Italy’s ENI) have largely stood in the way of establishing greater cross-border transmission interconnections. This situation exists because such interconnectors would have little value unless national utilities were able to sell energy across borders throughout the Common Market, which is a fundamental platform of proposals to “unbundle,” or split up, these very same national monopolies.

At the same time, the EU bureaucracy has produced no shortage of memoranda, directives, and regulations discussing “effective legislative and regulatory frameworks” for reforming the EU’s energy sector (European Commission 2006, 3). The Commission’s Directive 98/30/EC, unanimously approved by EU member states in June 1998, as well as 2003/55/EC five years later, were intended to move Europe towards an integrated market (Lecarpentier 2005, 1) by mandating “third party access to transmission and distribution networks” (Delegation of the European Union to Ukraine, Moldova, and Belarus 2009, 13–15). EU green papers, policy outlines, and regulatory briefs consistently emphasize the value of a more liberalized and more competitive European energy market, one diverse in terms of “source, supply, transport, and sales” (Cameron 2009, 23). They note the challenge of accomplishing diversification in the context of projected rising overall demand and declining indigenous production (Mangott and Westphal 2008, 175). Some even urge, obliquely, active diversification away from Russia: “increasing imports from an established source should be avoided” (Egenhofer et al. 2004, v).

Yet EU documents do not delineate any path towards implementation of mechanisms intended to counter third party supply monopolies. As an example, Directive 2004/67/EC only requires that each member state establish “minimum security of supply standards” (Delegation of the European Union to Ukraine, Moldova, and Belarus 2009, 15). Critics of current EU policy argue that handling of energy issues in Brussels has been “too rhetorical and too visionary,” devoting relatively little attention to pedestrian but crucial issues such as pipelines and market regulation (Hill 2005, 6). Yet even EU diplomats acknowledge that member states are likely to maintain their own foreign policy approach, particularly with regards to energy. Ursula Plassnik, a former Austrian foreign minister, noted that she expects member state priorities to prevail for some time even as the newly established European External Action Service draws upon staff from the Council Secretariat, the Commission, and member state foreign ministries to play a more active role in Europe’s external policy (Plassnik 2009).

Granted, Europe-wide energy policy has seen some limited success. For example, in 2003, the Commission forced Gazprom to drop “destination clauses” prohibiting downstream resale from its contracts with European firms (Lecarpentier 2005, 6). However, the EU bureaucracy has at times provided a sort of scapegoat for avoiding real liberalizing work in the energy sphere; in an interview, former President of the European Commission Romano Prodi noted member states’ tendency to play the “it was Brussels” card in order to defer responsibility for energy reform (Prodi 2009). Notably, the ultimately unratiﬁed European constitution would have explicitly established energy as an area of “shared competency” between member states and Brussels (Westphal 2006, 52). The Lisbon Treaty, which replaced the constitution, meanwhile, while continuing to formally refer to energy as an area of “shared competency,” speciﬁcally reserves decisions regarding “the general structure of [. . .] energy supply” for member states (Office for Ofﬁcial Publications of the European Communities 2007, C 306/88).

The U.S. Interest in an Integrated European Energy Policy

Officials from Brussels to Bucharest, unable to form a common position based around a unified energy market, often must
“choose between an affordable and stable energy supply and siding with the US on some key issues” (Cohen 2007, 12). Aware of the possibility of a European “strategic drift” resulting from energy-related conflicts of interest exacerbated by the lack of a unified European energy policy or transmission network (Cohen 2007, 1) the United States can move proactively to ensure that Russian energy does not indirectly limit the economic and foreign policy independence of the United States’ European partners.

The United States can work to encourage the implementation of the policy options detailed below, particularly by influential EU members such as France, Germany, and Italy—the so-called “Strategic Partners” for whom special energy relationships with Russia have the potential to “[undermine] common EU policies” (Leonard and Popescu 2007, 2). Conspicuously silent on subjects ranging from Russian nuclear collaboration with Iran to the lack of investor rights for Western firms in Russia, these nations have the potential to stand as unencumbered partners to the United States and to each other.

U.S. Interest in a Liberalized Transatlantic Energy Structure

By simultaneously encouraging European integration on energy matters and cooperating with the European Union on the creation of a liberalized transatlantic framework for energy independence, the United States can help enact a more sustainable energy policy for both itself and its allies. Together, the European Union and the United States represent forty percent of world energy consumption (Belkin 2008, CRS-18–CRS-26); as major consumers, Europe and the United States can jointly “promote more market-based approaches to the development, extraction and trade of key energy commodities such as oil and gas” (Union of Industrial and Employers’ Confederations of Europe 2007, 4). In other words, work to liberalize a market in which suppliers currently tend to be monopolistic by creating a single, competitive buyers’ market to counter the sizeable leverage enjoyed by suppliers such as Gazprom. This transatlantic effort should focus first on integrating markets for natural gas. For producer countries, oil is a much more difficult commodity to use for geopolitical purposes due to the fact that there exists a liquid international market for oil offering multiple, transparently priced supply, transport, and trans-shipment options (European Commission 2008, 4). This characteristic in and of itself is evidence of the benefits to be reaped from creating a comparably open market for natural gas (Noël 2008, 2).

Limiting and countering the effects of energy import dependency is critical to ensuring U.S. interests ranging from the security of the oil and gas supply to NATO bases in Turkey to preserving European diplomatic leverage vis-à-vis Russia (Cohen 2007, 7). Such goals can be more effectively achieved through better exchange of technology and information and through a coordinated diplomatic effort to mitigate import dependency. By pooling their diplomatic and economic influence, the United States and European Union can ensure that producer leverage is not excessive and that the integrity of downstream distribution networks remains intact.

Other Factors to Consider

While this cooperative approach is not intended to antagonize any third country, Russia will not react benignly to an overt attempt to reduce dependency on its energy. Already, it has shown a willingness to disrupt any activity that mitigates this dependence, as evidenced by Russian efforts to suppress pipeline plans it does not control and embargo refineries that resist Gazprom acquisition. For example, Russia has succeeded in stifling the EU-supported Nabucco pipeline from Turkey to Austria in favor of its own plan, South Stream, achieved through intense lobbying of transit countries such as Hungary. In Latvia and Lithuania, Russia ceased oil shipments to the Ventspils Nafta and Mazeikiu Nafta refineries, respectively, drawing a “muted” reaction from Western European EU members unwilling to protest at risk of revealing their own vulnerability (Baran 2007, 5). Through more vigilant enforcement of existing business regulations such as the aforementioned ener-
gy directives outlined by the European Commission, the United States and the European Union can help ensure that monopolistic moves do not go uninvestigated and that a collective strategy is adopted before it is too late. Without such a strategy, the European Union—by far the largest, most lucrative, and most accessible consumer of Russian energy—will continue to one-sidedly assume the vulnerabilities of an energy relationship characterized by mutual dependence (Bustin 2009).

POLICY OBJECTIVES

An Integrated Energy Market

Europe can only leverage the scale of its energy market by uniting it. By implementing the regulatory changes and developing the infrastructure necessary for this to happen, the European Union will give its eastern members the flexibility they currently lack due to their high individual levels of import dependency. The contemporary European energy grid and market are a legacy of national utility operators and their 1970s-era bilateral agreements with the Soviet Ministry of Gas (European Commission 2003, 1; Noël 2008, 8–9). Some names have changed—Italy’s Eni, the Netherlands’ Gasunie, Gaz de France, etc. now face Gazprom (Baran 2007, 8). However, these European companies continue to purchase gas using the same patchwork of bilateral deals that allow Russia to divide the continent into a series of weak segments (Korski 2008, 2). The alternative is a single European market that would require Russia to sell gas at the EU border, rather than to national firms. In this type of new system, the opportunity cost of geopolitical uses of energy would be raised; instead of shutting supply off to one country, supplier states would only be able to decrease supply to the market as a whole—significantly disincentivizing supply cuts (Noël 2008, 2).

By integrating their energy distribution and consumer networks, EU member states would confer on the energy sector the benefits accrued to virtually every other European industry through the Single Market: increased competition, lower prices for consumers, and most importantly, dramatically increased bargaining strength vis-à-vis third countries. A single European market for gas, and eventually for electricity as well, would leave Russia with one western export option: Europe (Noël 2008, 9), where a concession to any national distributor would immediately become available to consumers across Europe (Soros 2009, 2). The resentment and tension currently bred by eastern EU member states’ high dependence on gas imports would be replaced with the solidarity and flexibility of an integrated grid.

By lobbying its allies, in particular Germany, France, and Italy—the governments most reluctant to liberalize gas distribution, largely due to their powerful national energy monopolies (Noël 2008, 12)—the United States could help jumpstart the creation of a new European transmission and pipeline network in simultaneity with its own planned energy sector reforms (Obama for America 2008, 8). Through the transatlantic technology and investment initiative recommended below, this effort would allow for easier substitution of energy sources, a critical step for integrating renewable technologies undergoing expansion in both Europe and the United States. The United States can benefit by working with the EU on establishing funding mechanisms for energy infrastructure modernization. This coordination could take the form of a U.S. counterpart to the EU’s public infrastructure investor, the European Investment Bank (EIB) (European Commission 2008, 8–12). In fact, the Obama administration has already voiced plans for a comparable U.S. institution designed to provide low-cost financing for electricity transmission and pipeline projects (Leftly 2009).

---

8 Sixty-five percent of all Russian export revenues over the past few years have come from sales of energy to the European market. Russia is dependent on the openness of the European market; in fact, European profit margins finance profit-destructive energy subsidies to domestic Russian consumers.

9 Western and Central European countries began importing Soviet gas in the 1970s. These imports received a significant boost as the Iranian Revolution largely removed Iranian gas from the continent’s potential supply mix and indigenous production in the North Sea began a steady decline.
Unbundled, Liberalized Distribution

An integrated European market for energy would still not fully solve the issue of collusion between Gazprom and the so-called “national champions,” such as Germany’s E.ON Ruhrgas and Gaz de France. Thus, a comprehensive energy solution would require that any firm engaged in the distribution of energy in Europe, be it Russian, Danish, or American, be split from energy production (Hanson 2008, 121). Such unbundling is already in place in the United States, where it has successfully increased competition in the energy sector (Pollitt 2007, 28). In Europe, this policy would forestall geopolitically motivated down-market moves. In addition, it would complement an integrated energy grid by ensuring that gas and oil flow to distributors, not to the monopolies that currently segment the market along national borders.

The relationship between these monopolies and Gazprom has directly contributed to the division between eastern and western EU members. For example, Polish politicians infamously referred to Nord Stream, an undersea pipeline currently under construction in a partnership between Gazprom and German firm E.ON Ruhrgas, as “a new Molotov-Ribbentrop pact” (Leonard and Popescu 2007, 10). The pipeline, laid along the Baltic seafloor, bypasses an overland route through Poland that would have cost an estimated one-third as much (Baran 2007, 135). The costs to Europe include higher energy prices for consumers and a free rein for Gazprom to foment friction and tension among EU member states (Cohen and Szaszdi 2009, 9).

Energy sector unbundling would “Europeanize” the relationship between Gazprom and the European Union (Noël 2008, 3). Gazprom realizes that downstream activity, including wholesale and retail distribution, holds the greatest long-term profit potential, especially if Europe succeeds in crafting a single energy market (Locatelli 2008, 11). Yet Russia is aware that the only way to gain real access to these downstream assets is through reciprocation, that is, by allowing upstream access to European firms in exchange. For both Russia and the European Union, decoupling between production and distribution would defuse security worries by ensuring that no enterprise, domestic or foreign, controls the full vertical chain of energy production and distribution. Unbundling would better equip the energy industry for dealing with dwindling supply and more widespread sources of investment capital (Youngs 2007, 13).

Reciprocal Investment Rights

The benefits of an unbundled European energy sector are dependent on reciprocity between the European Union and Russia. Just as Europe will remain dependent on Russian energy for the foreseeable future—these policy proposals are meant to restructure, not necessarily eliminate, this dependence—Russia will remain dependent on European demand and on Western capital to finance reinvestment in upstream infrastructure (Jackson 2006, 6). Yet Russia cannot expect to benefit from this capital as long as it engages in tactics such as those carried out at Sakhalin and Kovytka gas fields, where Royal Dutch Shell and BP, respectively, were forced to sell their stakes in developing projects to Gazprom in what could be termed benign expropriations (Dellecker 2008, 6; Leonard and Popescu 2007, 23).

Only by taking the “fear tax” out of investments in energy production can Russia ensure the same rights for its own investments downstream in Europe and the United States (Egenhofer et al 2006, 4). The United States has a vested interest in this sort of reciprocity. Whereas Lukoil, a Russian oil firm, was allowed to purchase one hundred percent of U.S. firm Getty Petroleum and its 1,500 service stations, American and European firms are restricted to an apparent ownership ceiling of twenty percent on their investments in Russia (Smith 2008, 7). This lack of reciprocity is also

---

10 In an interview, Fyodor Lukyaonov, a prominent Russian journalist, referred to these national monopolies as “allies with Russia against liberalization.” Lukyaonov, Fyodor. Editor-in-Chief of Russia in Global Affairs. Videoconference Interview. 27 March 2009.
11 The 1939 agreement between Nazi Germany and the Soviet Union that included a secret protocol for the two sides’ partitioning of Poland.
of direct concern to European consumers of Russian gas; without sufficient capital investment in production infrastructure, Russia will not be able to sustain current levels of production in coming years, creating the possibility of future energy shortages (Baran 2007, 11). The missing factor is investment, not the availability of capital. Foreign investors cannot be expected to pour the enormous amounts of capital required for the continued exploitation of Russian resources without a significant solidification of investor and property rights (Monaghan and Montanaro-Jankowski 2006, 20).

The United States can pursue investment reciprocation through forums such as Russian WTO accession talks. It can also push for Russian ratification of a new treaty that would be acceptable to all parties replacing the moribund Energy Charter Treaty (Emerson 2006, 7). A new treaty which secures investor rights and reduces transit risks would be advantageous not only to Europe but also to Russia, which would benefit from a clear legal mechanism to arbitrate energy disputes with transit states such as Ukraine (Bely 2009, 6). The American role in securing passage of such a treaty could also involve exerting diplomatic pressure on certain new EU members, such as Poland and Lithuania, which have in the past obstructed further negotiations with Russia (Velyaminov 2009).

Focus on Liquefied Natural Gas

Liquefied Natural Gas (LNG) currently represents fifteen percent of EU gas imports. Easier to move and store than ordinary gas, it would be a useful complement to an integrated continental distribution grid (Belkin 2008, CRS-18–CRS-19). In an integrated market, LNG sells much like oil, in the sense that it is highly mobile and difficult to manipulate geopolitically (Noël 2008, 6). Moreover, producers including Norway, Algeria, Nigeria, and Qatar have begun selling LNG, offering Europe an immediate means for reducing dependence on Russian pipelines (Council of the European Union 2007, 17). Russia also has a vested interest in LNG, which would allow it greater export flexibility (Kupchinsky 2007, 1). The United States can encourage the development of LNG infrastructure in Europe, particularly in the pipeline-dependent Baltic states and Poland (European Commission 2008, 9). Whereas LNG was constrained by prohibitive costs only a decade ago, a recently concluded boom in the construction of liquefaction and de-liquefaction plant infrastructure has made it “competitive with pipeline gas.” Indeed, the growth rate in LNG production has for several years exceeded the growth in overall natural gas production capacity (Dahlman Rose & Co. 2009, 6). Through a partnership for LNG investment and technology transfer, the United States and Europe could further encourage the transformation of the gas market into an oil-like market, with spot purchase and arbitrage gas availability taking the place of contract-driven pipeline-delivered gas (Day 2009).

Focus on Nuclear Energy

In reducing dependence on imported sources of energy, the European Commission has chosen to emphasize the role of renewable energy sources such as wind and solar power. However, its own analysts admit that “[e]ven when the renewable energy policy goals are reached, Europe is likely to be dependent on more imports than today” (European Commission 2008, 2). Thus, if Europe is to make progress on reducing dependency on imported gas and oil, projected to reach seventy and ninety percent respectively by 2020, other sources must be considered (European Commission 2008, 18). Currently, nuclear energy is Europe’s biggest single non-
imported source of energy (European Commission 2008, 9–10), supplying fifteen percent of total EU energy consumption needs and thirty-one percent of electricity needs (FORATOM 2007, 1). Europe’s nuclear industry enjoys a global leadership, with 148 reactors in fifteen of twenty-seven member states easily outproducing the United States’ own nuclear sector (European Commission 2008, 1). Yet attitudes towards nuclear energy differ widely; whereas France satisfies nearly three quarters of its electricity demand through nuclear power (Belkin 2008), Germany and Sweden have only recently begun reconsidering plans to phase out nuclear power generation altogether (The Economist Intelligence Unit 2009, 1–2). The United States, meanwhile, has not built a new reactor in decades, and tentative plans for a new generation of American nuclear plants are based on a French design (Wald 2008).

Obviously, the choice of pursuing nuclear power cannot be forced onto any EU member state, or the United States for that matter (European Commission 2008, 4). However, given the fact that simply maintaining nuclear generating capacity in both the United States and the European Union will require extensive new construction (FORATOM 2007, 2), a transatlantic nuclear collaborative effort would make long-term energy independence all the more feasible. This collaborative effort could come under the auspices of the IAEA in the form of a new transatlantic agreement for nuclear cooperation, comprising the United States, the European Union, and Russia. Indeed, nuclear energy cooperation offers an opportunity to defuse some of the tension that may result from overtly acting to reduce dependency on Russian gas. Current nuclear sector collaboration with Russia, such as the Rosatom-Siemens joint nuclear power engineering enterprise in the Kaliningrad enclave of Russia or the Department of Energy-Russian Ministry of Energy common protocol (U.S. Department of Energy 2003, 2), can serve as a model for pooling American, European, and Russian efforts towards common goals such as avoiding proliferation and ensuring nuclear safety (The Economist Intelligence Unit 2009, 1–2).

The European Union’s indigenous energy production is de-pleting partly due to reluctance to maintain or expand generation sources perceived to be polluting or unsafe, such as coal and nuclear energy (European Commission 2008, 8). By 2020, nuclear power’s contribution to European energy production is projected to fall unless a significant number of new plants are built to meet new demand and replace obsolete reactors (European Commission 2008, 15). The United States faces a similar problem, and has a similar aversion to increasing carbon or import dependency. A joint partnership could serve as a model for a world whose energy use is projected to rise 50 percent by 2030 (Barden 2008, 34).

The stalled U.S.- and EU-backed Nabucco pipeline emblemsizes the disadvantages of the European Union’s current approach to energy policy. Intended to carry Caspian and possibly Iraqi natural gas westwards through Turkey to a hub terminal in Austria, it has been overshadowed by the Russian-backed South Stream project. South Stream, like its Baltic cousin Nord Stream, expensively cuts across the sea floor (the Black Sea between Russia to Turkey) in order to avoid the costs (in the form of annual transit fees) and the risks (in the form of the constant risk of contract renegotiation) of transit countries such as Ukraine. Owned by a consortium comprising both Gazprom and Italian utility ENI, the sixty-three bcm pipeline is set to be completed by 2015 (Mankoff 2009, 20). It is, in the words of Hungarian Prime Minister Gyrúrsány, “backed by a very strong will and a strong organizational power,” contrasting clearly with the institutionally fragmented EU (Baran 2007, 140).

To set South Stream in motion, Russia picked out one country, Hungary, for a lucrative bilateral deal. By guaranteeing Prime
Minister Gyurcsány that South Stream would terminate in Hungary, giving it partial ownership of the terminal storage facility. Gazprom—with support from the Kremlin—effectively co-opted one EU member at the expense of the entire Union’s energy independence (Dempsey 2009). South Stream, like Nord Stream, is an overtly political project: it simply diverts gas that currently flows through Ukraine. As part of the deal, Gazprom agreed to construct a spur off the pipeline into Serbia, gaining control of Serbian oil company NIS in exchange (Pop 2008). Romano Prodi, the former Italian Prime Minister, was offered chairmanship of the project, in a parallel to former German Chancellor Gerhard Schröder’s current chairmanship of the Nord Stream consortium, for which he is paid an annual salary of one million euros (Cohen and Szaszdi 2009, 9; Osipov 2007, 9).

South Stream, in other words, perfectly demonstrates the failings of Europe’s fragmented market. This situation allows for maneuvering by a producer country—Russia—that does not offer consumer countries the same investment rights it regularly exercises downstream. The competitor pipeline, Nabucco, is truly a “litmus test” of European institutional and political strength. For its own part, the United States did exercise extensive pressure on the Hungarian Gyurcsány government to back Nabucco (Baran 2008). However, in the end the prime minister agreed to the Russian proposal, after he realized that the EU– and U.S.– backed plan appeared to be crippled by political fragmentation (Cohen and Szaszdi 2009, 9).

Nabucco, if it were to be built, would break Russia’s monopoly on Caspian natural gas. By 2020, for example, it would allow Azerbaijan to export an amount of gas equal to one third of Russia’s current exports to the European Union (Baran 2007, 24). Yet even if it existed today, it would not break Russia’s control over the European market. This can only occur when the European Union implements a single energy market, a market immune to lucrative but ultimately one-sided deals that hamper European unity. In an integrated grid and market, Nabucco could then be tied into countries like Poland or even the Baltic states through proposed supplementary pipelines, including north-south links, finally installing a gas alternative where none exists today (Baran 2007, 24).15

The Caucasus and Caspian Sea Regions

The geopolitical importance of the Caucasus and Caspian regions, particularly in recent years, is hard to overstate. The area is an energy crossroads, home to pipelines, oil and gas fields, and clashing geopolitical ambitions. As a result of Europe’s energy dependence, its leaders are tempted to cast a blind eye when Russia bullies or hampers democratic efforts in nations such as Azerbaijan, Georgia, Kazakhstan, and Turkmenistan because of their energy producer or transit country status (Baran 2007, 133). Moreover, using its unchallenged energy muscle, Russia has successfully acquired substantial geopolitical influence in the region. This in turn has further weakened Europe’s position by denying it independent alternate energy suppliers (Baran 2007, 10).

The United States has previously succeeded in improving its Caucasus allies’ energy security. In the late 1990s, the United States assisted the Azerbaijani and Georgian governments in constructing independently owned pipelines from Baku to Turkey and on to Europe, free of Russian control (Baran 2007, 12–13). With cooperation from NATO ally Turkey, the Baku-Tbilisi-Ceyhan (BTC) oil pipeline connected Azerbaijan’s Caspian Sea shoreline to Turkey’s southern coast in 2006 (Baran 2007, 136). The United States also helped prevent Gazprom’s takeover of this and other Georgian energy infrastructure by offering Tbilisi a grant to maintain and upgrade its network (Baran 2007, 25). The BTC pipeline was briefly threatened during Russia’s 2008 invasion of Georgia, underlining the geopolitical and symbolic value of this independent

---

14 This would be an important achievement, even if political considerations are set aside. Currently Russia is using Central Asian and Azerbaijani gas to replace its own declining capacity. Forcing competition would be an added incentive for necessary Russian production investment.

15 With pipeline connections to coastal de-liquifiation terminals offering access to other gas alternatives in the form of LNG.
link, which nonetheless represents at most six percent (one million barrels of oil per day) of Europe’s overall crude oil demand (Cohen and Szaszdi 2009, 2). 16

Clearly, Russia is in a stronger position today than it was in the late 1990s. However, by integrating its capacity for political leverage, the EU could join the United States in ensuring that energy development in alternate producer regions, especially in the contested Caucasus and Central Asia, is cooperative rather than a zero-sum game (Baran 2007, 16). This could be accomplished through information sharing, investment reciprocity, and possibly through negotiations about Iran and missile shield proposals for Eastern Europe. Above all, the United States would need to make clear that it is not “surrounding” Russia (Smith 2008, 10), but rather insisting on reciprocal investment access to producer regions — an understandable goal given apprehension about monopolistic control from Gazprom (Bely 2009, 4). By giving its unified, “unambiguous support” to a U.S.-supported program for establishing both geopolitical stability and energy market competition in this region, the EU can avoid losing Azerbaijan and other alternate producer states to agile Russian diplomacy (Loskot-Strachota 2009).

POLICY RECOMMENDATIONS TO ADVANCE U.S. INTERESTS

In light of the cases and policy options presented above, the United States has several options for mitigating European institutional discord and energy dependence on Russia. Some policies, such as nuclear and Liquified Naval Gas cooperation or collaboration with NATO partners such as Turkey, can be accomplished directly with the collaboration of relevant U.S. departments (State, Defense, and Energy) and agencies. Others, such as European energy market integration and utility unbundling, could be better encouraged through primarily ministerial-level diplomatic pressure on member states and the European Commission in both bilateral meetings as well as roundtable meetings ranging from the Transatlantic Economic Council (TEC) to G8 and WTO. Additionally, ambassadors in Europe could express the United States’ recognition and support for European efforts to reduce dependence on imported energy through diplomatic contacts and opinion editorials in local newspapers. Through these forums, the U.S. could leverage the increasing willingness it has recently shown to pursue post-Kyoto climate change negotiation for a greater EU willingness to pursue energy strategy cooperation that would offer the European Union a plethora of additional intrinsic benefits, including greater supply security.

First, the Obama Administration should encourage European energy market integration. A major aspect of the EU’s energy insecurity involves its inability to transmit energy across national borders. Both Europe and the United States are presently engaged in planning for the development of a new generation of energy infrastructure. From European objectives such as “overnight battery recharging for plug-in electric vehicles” (European Commission 2008, 1) to American plans for dynamic transcontinental transmission lines (Joyce 2009), these plans are comparable and could be better coordinated. Currently, only about ten percent of European power generation capacity can be exported across borders (Belkin 2008); similarly, the Obama Administration believes that the current American power grid is not equipped to carry new sources of energy to points of need (Obama for America 2008, 8).

Implementation of a European energy grid could begin with a “warning system” of bilateral deals between European energy firms and foreign energy suppliers (Leonard and Popescu 2007, 59). Notified in advance of such deals, the Commission would have time to review and react to tactical moves by Gazprom. In the medium term, the European Union could continue implementing basic transmission and pipeline developments, such as the France–Spain and Poland–Lithuania electricity grid interconnections it currently coordinates (European Commission 2008, 5–6). In the long term, both gas and electricity could begin to be traded across an integrated grid, lessening the need for divisive national-level deals with

---

16 The BTC pipeline’s capacity is one million barrels of oil per day. A parallel pipeline, the Baku-Tbilisi-Erzurum (BTE) pipeline, supplies Azerbaijani natural gas to Turkey.
external suppliers (European Commission 2008, 10). The current project for integrating the European grid, Trans-European Energy Networks (the aforementioned TEN-E), does not appear to be getting the job done (European Commission 2008, 3). Thus, the best option for the United States may be to lobby major EU states, such as France and Germany, to explore successful existing models of grid cooperation. Existing initiatives, such as Baltrel, which combines energy producers and transit operators in Scandinavia, the Baltics, and Poland, could serve as an “intermediate step to” a larger European single market for both gas and electricity (Baltrel 2007, 3–5).

The U.S. role in this European project is twofold: coordinating it with its own push for a new generation of transmission infrastructure, and insisting that a real structure to counteract Russian leverage, rather than vague “crisis response mechanisms,” are put in place (Council of the European Union 2007, 18). In other words, the United States could push for a long-term solution to the issue of dependence rather than short-term responses to Russian action. If both the European Union and the United States work together, both private and public resources will be easier to find, and major efforts—such as connecting wind farms in the North Sea and solar installations in New Mexico to mainland grids—will be complementary and coordinated (European Commission 2008, 2–3). Finally, in contacts with Russia, the United States could make clear that an integrated European grid would better equip Russia to export energy alternatives beyond oil and gas, such as selling electricity directly to an integrated European grid (European Commission Green Paper 2006, 6).

Second, the Obama Administration should pursue transatlantic energy market liberalization. Liberalization of the transatlantic energy market would involve unbundling European monopolies, but would provide for the transparency and reciprocity that is necessary for constructing a single energy market and two-way flow of investment. This policy would benefit Europe by forestalling bilateral deals between Gazprom and national energy firms, and it would thus help remove the hidden cost paid by European consumers for the disunity of their energy market. The United States would benefit by finally obtaining the reciprocal right to invest in EU utilities as European firms have done in the deregulated U.S. energy market for years. Even Russia would profit from accession to this policy, which would give its unbundled firms full investment rights downstream and allow Western capital and technology to flow to fill critically needed upstream needs without fear of extortion.

The European Union has the power to implement this proposal under its power to enforce competition regulation (Gray 2009, 1). It will, however, be difficult to put into motion due to the resistance of powerful members such as Germany and France, both of which have resisted unbundling (Belkin 2008, CRS-9). However, the only way for an integrated European energy market to be fully effective is through the complementary effect of this policy. With the strength of a united, liberalized market, Europe could finally effectively pursue antitrust measures against Gazprom’s violations (Smith 2008, 2) of Article 82 of the European Community Treaty (Delegation of the European Union to Ukraine, Moldova, and Belarus 2009, 17–18; Baran 2007, 141). Furthermore, armed with the strength gained from this policy, the European Union would have more leverage in pursuing the expansion of the energy acquis to nations such as Ukraine, where a lack of unbundling and transparency has contributed to and exacerbated recent gas crises.

The United States has tools and resources to help a Europe that has integrated and liberalized its energy sector. In particular, the United States can play a role in prodding investigation of non-transparent entities such as the notorious RosUkrEnergo, a Switzerland-based firm that has allegedly served as a high-level money-laundering scheme for well-connected Gazprom and Ukrainian

17 It certainly is guiltier, and more of a threat to European commercial security, than Microsoft has repeatedly been found to be.
18 Article 82 forbids the “abuses of dominant positions” such as “dissimilar conditions for equivalent transactions”—an excellent description of Gazprom’s current strategy of price discrimination between national European customers.
officials (Leonard and Popescu 2007, 10–11; Wilson 2009, 1). The United States can also assist the effort to liberalize the energy market by incentivizing transparency in transit states, by contributing to funds that help transit states improve infrastructure and discouraging geopolitical manipulation, and by cooperating with the European Union on intelligence and law enforcement (Smith 2008, 10).

Third, the Obama Administration should establish a transatlantic partnership for nuclear and LNG collaboration. A new collaborative effort on nuclear energy and liquefied natural gas between the United States, the European Union, and possibly Russia, would help fill a “glaring omission” in many EU reports about energy dependence (European Policy Centre 2007, 2). Starting with regular consultations through the existing Strategic Energy Dialogue and culminating in a European Aeronautic Defense and Space Company (EADS)—style multinational partnership for the transatlantic community (perhaps under the auspices of the IAEA), this proposal would utilize technology to add stability for energy consumers and producers alike.

In the field of nuclear energy, this partnership would build off existing research collaboration such as Europe’s Energy Research Alliance, offering the United States, European Union, and Russia a chance to cooperate on improving safety and efficiency in a method of power generation all three powers utilize (European Commission 2008, 8). By standardizing and streamlining regulation, this policy would make it easier for French firms to import plans for so-called “Generation III” nuclear plants, or those incorporating latest safety and efficiency features, to the United States—where nuclear currently suffers from a regulatory process which is a “product of the 1970s anti-nuclear and post-Three Mile Island mindset” (Spencer 2009, 2; FORATOM, 8). Similarly, U.S. firms would find it easier to enter the European market, helping to expand an energy sector that produces energy “on a large scale, in a secure manner, at competitive costs and with respect for the environment” (European Commission 2008, 5).20 Existing U.S. legislative efforts at moderating the relatively high startup cost of new nuclear plants, such as a provision in the 2005 energy bill to lower the cost of capital for new reactors (primarily through debt insurance), could be extended in light of recent recommendations that underlie the extent to which scaled-up reintroduction of nuclear power generation could significantly reduce costs (Friedman 2007, 9).

In the field of liquefied natural gas, this proposal would help lessen European dependence on the long-term bilateral contracts that contribute to disunity among EU members. By increasingly transforming natural gas into an oil–like liquid market with widespread spot (rather than contract) availability, LNG would defuse the geopolitical aspect of gas importation (Delyagin 2006, 136-137). Already, expanded LNG capacity has helped move natural gas towards a price structure more like that of oil, a truly global commodity (Krauss 2009). The United States and European Union could join together in further promoting LNG through tax incentives and common regulatory streamlining. Extensive development of LNG would remove the need for highly expensive new pipelines, such as South Stream, by making gas available wherever a port exists to receive it (Belkin 2008, CRS-15).

CONCLUSION

The European Union is dependent on a small number of large suppliers for its energy. This dependency has severely exacerbated the fragmentation of its energy market along national lines and intensified contention between old and new members. This situation endangers both cooperation among EU member states as well as the reliability and independence of European diplomatic and po-

---

19 EADS, or the European Aeronautic Defense and Space Company, most famous for its Airbus jets, was founded in 2000 in a merger of several European aerospace firms. It has proven highly successful in terms of encouraging EU industrial and technology integration and cooperation in a way that might serve as a useful model for encouraging transatlantic energy collaboration.

20 Eurobarometer polls show that a plurality of opponents to nuclear energy base their opposition on dissatisfaction with current waste disposal practices. A possible solution is the approach currently taken by the United Kingdom, which restricts nuclear operators to land around decommissioned nuclear plants.
lical action with regards to United States interests. By diplomatically prodding the European Union to adopt both a more competitive private energy market and more robust regulation of its energy sector, the United States can help dampen the benefits enjoyed by suppliers such as Russia as a result of its monopopy power. By agitating for transmission infrastructure improvement, expansion of nuclear power generating capacity, and speeding up of LNG technology implementation, the United States can help lessen Europe’s overall dependence while increasing the economies of scale for its own comparable energy reforms (Cohen and Szaszdi 2009, 2).

Works Cited

_____. Testimony to the United States Senate Committee on Foreign Relations. 12 June 2008.
Bely, Andrei. “After the Oil Boom.” Russia in Global Affairs. 1. 8 March 2009.
Hanson, P. “Russia and Europe are Doomed to Cooperate.” Russia in Global Affairs, March 2008.
The Economist Intelligence Unit. “Germany nuclear: Russia, Germany join for nuke engineering.” The Economist, 10 March 2009.
In the Age of Discovery, European powers explored, conquered, and tied together the regional systems of the world. These states—Portugal, the Netherlands, France, and England—were the first world powers and influence events around the globe based on their power projection capabilities. In jockeying for position with other competing states, establishing colonies, and subjugating tribute states, the world powers relied on one particular form of power projection: naval power. Christopher Columbus, James Cook, Bartolomeu Dias, Francis Drake, Vasco da Gama, Henry Hudson, and Ferdinand Magellan all sailed from Europe and explored Africa, South Asia, the Americas, and Oceania. If a map does not show the obvious, these men’s stories do: Europe conquered the world because it could project naval power. Indeed, Europe’s use of land power on these newly-explored continents depended on the ability of its navies to bridge the oceans between the metropoles and the colonies.

Nearly seventy-five percent of Earth’s surface is water. In the international system, a blue-water navy is one of the necessary conditions to be the world’s dominant power (Organski 1968, 364). From time to time, the dominant state perceives a potential challenger, a rising state that is developing capabilities to rival or overtake those of the dominant power. Power transition scholars seek to shed light on the dominant state-rising state dyad to understand

1 A blue-water navy has the ability to undertake operations far from land for extended periods of time without support from or defense by forces based on land. This differentiates a blue-water navy from green- and brown-water navies: a green-water navy can project power up to a couple hundred miles from shore, but still relies heavily on land support, while a brown-water navy is a riverine navy, patrolling rivers, lakes, and coastlines.

Brian C. Chao graduated cum laude in 2009 from Dartmouth College, receiving a Bachelor of Arts degree in Government with High Honors as well as the Rockefeller Prize for the best thesis in international relations. Brian is currently a Chinese Government Scholar at Tsinghua University in Beijing, China.
Brian C. Chao graduated cum laude in 2009 from Dartmouth College, receiving a Bachelor of Arts degree in Government with High Honors as well as the Rockefeller Prize for the best thesis in international relations. Brian is currently a Chinese Government Scholar at Tsinghua University in Beijing, China.

In the Age of Discovery, European powers explored, conquered, and tied together the regional systems of the world. These states—Portugal, the Netherlands, France, and England—were the first world powers and influence events around the globe based on their power projection capabilities. In jockeying for position with other competing states, establishing colonies, and subjugating tribute states, the world powers relied on one particular form of power projection: naval power. Christopher Columbus, James Cook, Bartolomeu Dias, Francis Drake, Vasco da Gama, Henry Hudson, and Ferdinand Magellan all sailed from Europe and explored Africa, South Asia, the Americas, and Oceania. If a map does not show the obvious, these men’s stories do: Europe conquered the world because it could project naval power. Indeed, Europe’s use of land power on these newly-explored continents depended on the ability of its navies to bridge the oceans between the metropoles and the colonies.

Nearly seventy-five percent of Earth’s surface is water. In the international system, a blue-water navy is one of the necessary conditions to be the world’s dominant power (Organski 1968, 364). From time to time, the dominant state perceives a potential challenger, a rising state that is developing capabilities to rival or overtake those of the dominant power. Power transition scholars seek to shed light on the dominant state-rising state dyad to understand how these relationships unfold and what consequences they may have for the international system.

A blue-water navy has the ability to undertake operations far from land for extended periods of time without support from or defense by forces based on land. This differentiates a blue-water navy from green- and brown-water navies: a green-water navy can project power up to a couple hundred miles from shore, but still relies heavily on land support, while a brown-water navy is a riverine navy, patrolling rivers, lakes, and coastlines.

Brian C. Chao graduated cum laude in 2009 from Dartmouth College, receiving a Bachelor of Arts degree in Government with High Honors as well as the Rockefeller Prize for the best thesis in international relations. Brian is currently a Chinese Government Scholar at Tsinghua University in Beijing, China.
when transitions are attempted, how they are attempted, how each state reacts, and what factors aggravate or mitigate the attempted power transition. These two concurrent conditions—the dominant state’s dependence on its naval dominance and the periodic dominant-riser power transition dyad—lead to the question of naval power’s role in power transitions.

Today, the dominant state is the United States and its navy is the dominant navy in the international system. But the United States is not the only power, and the recent rise of the People’s Republic of China (PRC), its implications for the United States and the international system have been receiving much attention. As China rises, it may transform its current green-water navy into a blue-water navy, which would be China’s first in 600 years. This would add a new variable in the Taiwan scenario and could challenge U.S. naval hegemony in East Asia. Resource security, prestige, and establishment of a maritime defense barrier are all reasons for Chinese naval development. Whatever the reason, the United States has kept a watchful eye on the PRC’s navy. What effect would a Chinese blue-water navy have on the United States and bilateral relations? Would both U.S. perceptions of and policy toward China change appreciably and, if so, how?

This paper analyzes the naval variable in the current U.S.-China relationship and addresses these questions in five sections. First, I elaborate on the role of naval power and its enduring importance in the face of air power, nuclear weapons, and globalization. Second, I discuss what the current U.S. threat perception is and its contributing factors. I argue that, as of spring 2009, the People’s Liberation Army (PLA) Navy does not present a general naval threat to the United States. This is due to a PLA Navy modernization effort that has been partially ineffective and that is aimed primarily at Taiwan rather than global power projection. Third, I outline the steps China’s navy would take for the United States to develop a threat perception sufficient to alter its China policy. Fourth, I discuss policy implications for the U.S.-China relationship in the event of a naval arms race. Finally, I examine the theoretical implications that this chapter holds for the United States, China, and international relations and the power transitions that could occur therein.

NAVAL POWER’S ENDURANCE IN THE MODERN ERA

Five hundred years ago, naval power was indisputably important; however, its relevance is no longer obvious. In a world of airplanes, nuclear weapons, and increasing economic globalization, does naval power still occupy such a prime role? These issues raise broader questions about the salience of technology in understanding contemporary geography, geopolitics, and geostrategy. Although many argue that technology has figuratively closed the distance between peoples and rendered literal distance meaningless, I hold that naval power is still important in light of air, nuclear, and economic developments.

Air power has not made naval force obsolete. It may be quick, but it is temporary. Aircraft cannot stay in the air for months at a time; they must return to base to be refueled. Refueling in flight requires sending up a certain number of aircraft just to ensure that those already in the air stay in the air; it is not an economical process. Furthermore, while technology has allowed one airplane to connect a fuel hose to another and recharge that machine, no technology exists by which one pilot in a fighter plane can be relieved in mid-air by another pilot in the same machine. Thus, man and machine constrain the permanence and efficacy of air warfare, which raises a third issue in the form of basing requirements. Air power requires bases from which to project, but air force bases are immobile, requiring great geographic distribution for expedient global air support. The navy’s aircraft carriers mitigate this disadvantage by providing mobile air bases around the world; thus, global air power has actually made global naval power more, not less, important.

2 I define power transition to mean both successful and unsuccessful power transitions.
3 “Geography is the geological reality of the earth, composed of mountains, rivers, seas, climate, and so on. Geopolitics is a combination of geological features (e.g., natural resources) with human activity (e.g., production and communication technology) that alters the value of places. . . . [G]eostrategy describes where a state directs its military and diplomatic efforts” (Grygiel 2006, ix–x).
taint. Aircraft and airmen cannot stay in the air continuously with enough firepower to be effective without exacting great costs. What an air force calls “virtual presence” is, in reality, actual absence (China, Taiwan, and Mongolia Team [China Team] 2009).

The navy does not face the problems of the air force. Its vessels, especially those that are nuclear-powered, can remain at sea for extended periods of time. Blue-water navies, by definition, are self-reliant for both provisions and defense and do not require land support. Since blue-water navies are not tied to their shore support facilities, naval forces are free to sail to more places for longer periods of time in a more efficient way than air forces. This self-reliance ensures that the navy is not tethered to land as the air force is. Naval vessels also avoid the manning problem of aircraft. Ships and submarines are manned by crews in the hundreds and thousands, so just as the machine can operate for extended periods of time, so its large crew can operate for extended periods of time as well.

Nuclear weapons have not rendered naval power obsolete because the two are different types of weapons designed for different purposes. Nuclear weapons are meant to ensure a state’s survival, while naval dominance is meant to preserve the dominant state’s power projection capabilities. To argue that nuclear weapons have diminished the importance of naval power or any other conventional force would thus require one to assume that most, if not all, external threats are existential. This is not the case. Nuclear weapons are not used for power projection, but for existential continuation.

It can be argued that nuclear weapons render conventional warfare between nuclear-armed states obsolete because the adversaries cannot be sure that a conventional conflict would not escalate. This point is valid, but it implies that all conflicting naval power projections have a great probability of ending in war. This is not the case, as the United States and the Soviet Union did not go to war in their naval arms race. The Soviet Union did not respond to initial U.S. naval superiority by going to war with Washington; rather, it built a blue-water navy. Likewise, the United States did not respond to this buildup with war against Moscow; it respond-

ed with renewed naval shipbuilding and a forward defense strategy. The nuclear-weapons-as-conventional-war-deterrent argument also cannot explain why the United States chose a horizontal escalation strategy in the 1980s that was deliberately confrontational and that may have raised the chances of war through its geographic expansion, creating more points for potential U.S.-Soviet conflict.

It is certainly possible for states to perceive nuclear weapons as more important than they are or important in more ways than they are. Policymakers may choose to fall back on nuclear weapons as a guarantor of existence and of power. The danger then becomes one of allowing conventional capabilities to diminish on the grounds that nuclear weapons make up the qualitative deficit, only to be caught out when the opposing state reaches nuclear parity. One’s existence may still be ensured, but one’s freedom within the international system is greatly restricted.

There remains the argument concerning economic globalization and the technology on which it depends. This argument does not so much attack naval power as it does geography; if geography is unimportant, then so are forms of power defined by it. The economic globalization argument asserts that a free global market diminishes the incentives to exercise direct control over natural and economic resources. . . . [T]he prerequisite of power is not control over resources but the ability to purchase them” (Grygiel 2006, 165). This argument fails to consider that the very economic ties that bind geographically distant states ensure that a state in one part of the world becomes even more vested in the geopolitical situation of another part of the world. “Because of limited resources, states have to discriminate according to some geographic criteria” (Grygiel 2006, 165).

The economic globalization argument also ignores the reality that “globalization does not mean that trade occurs in cyberspace.” Trade is facilitated through geographically-defined routes. The United States relies on ships and open sea lanes for forty percent of its foreign trade (Grygiel 2006, 165–66), while the proportion of global commerce that travels by sea stands at ninety percent. Colin S. Gray writes, “The strict limitations of air and space vehicles as
carriers of bulky or heavy cargo translates [sic] as a permanent significance for U.S. maritime power as a servant of foreign policy and national strategy” (1994, 162–63). Gray concludes that any future world order must also have a “nervous system that is principally maritime” (1994, 165). The sea lanes vital to international trade are kept open by U.S. naval power, not by a U.S. cyber monopoly, U.S. air power, or the U.S. nuclear arsenal. Recent Somali pirate attacks on merchant vessels in the western Indian Ocean affirm the signal value of the freedom of the seas to global commerce. Despite the technological advances of the modern era and technology’s pervasive influence in daily life, the importance of spatial considerations has not fundamentally changed.

THE PRESENT SITUATION

In this section, I show that Chinese naval development has not been uniform and still lags far behind U.S. abilities in key categories such as training, expertise, and quality of matériel. These factors prevent China from seriously and consistently projecting power beyond the East Asia region and limit the quality of the PLA Navy. I then contend that the PRC’s naval development is not geared toward blue-water power projection, but is instead a reflection of Beijing’s real defense priority: deterrence of de jure Taiwanese independence and, in case deterrence fails, the ability to establish de facto control of the island, with or without U.S. intervention. For geographic reasons, Chinese military preparations for a war involving Taiwan do not need a blue-water navy. Taiwan is, in a sense, a distraction for China that prevents Beijing from fully devoting itself to funding and constructing a navy with oceanic abilities.

Chinese Naval Development

What is the state of the PLA Navy today? Is it a blue-, green-, or brown-water navy? What is the trajectory of its development? Avery Goldstein argues that the PLA Navy, while preparing for a localized war involving Taiwan, is simultaneously pursuing a blue-water navy (2005, 60n36). His unconvincing argument cites certain acquisitions that the PLA Navy has made, such as guided missile destroyers and transport ships, but then explicitly connects their acquisition to use against the United States in a Taiwan conflict (Goldstein 2005, 61). His mention of improved PLA Navy amphibious capabilities undermines his blue-water hypothesis, as the utility of amphibious craft is obvious in a conflict over the island of Taiwan, but less so for oceanic power projection (Goldstein 1997/98, 48). Furthermore, the sources that he draws upon are out of date, some having been written in the early 1980s (Goldstein 2005, 60–62nn36-42). If the PLA Navy is pursuing a blue-water navy that projects power around the world, it has not done so in any serious or sustained manner.

Size, expertise, and operational effectiveness are necessary in determining a navy’s strength (Bush 2009). The fleet must be large enough to consistently project power at will, its officers and crew must have the training and experience to project power, and the ships, submarines, support vessels, and shore facilities must be have enough capacity to sustain power projection. An examination of the PLA Navy shows that Chinese naval development inconsistent with global capabilities and Goldstein’s argument. While the size is certainly increasing, expertise is minimal and operational effec-

---

4 The timing of this case study presents unique information-gathering difficulties. The most valuable documents are classified, so I must rely on scholars, policymakers, and the professional naval community. I draw heavily from four interviews I conducted for this case study: Professor William S. Murray, Dr. Richard C. Bush III, Professor Bernard D. Cole, and the China Team in the Office of the Secretary of Defense, U.S. Department of Defense. Murray is a former U.S. Navy submariner and an associate research professor at the U.S. Naval War College, where he also helped found its China Maritime Studies Institute. Bush is the Director of the Center for Northeast Asian Policy Studies and a senior fellow in foreign policy at the Brookings Institution. Cole is a professor of international history at the National Defense University and served with the U.S. Navy’s surface fleet before retiring with the rank of captain. The China Team is responsible for implementing and coordinating U.S. defense policy toward China, including managing the military-to-military relationship between Washington and Beijing. Murray’s views are his own and do not necessarily reflect the opinions of the U.S. Naval War College. Cole’s views are his own and do not necessarily reflect the opinions of the National Defense University. The China Team’s views are its own and do not necessarily reflect the opinions of the Secretary of Defense. None of my interviewees’ views necessarily reflect the opinions of the Defense Department or the U.S. Government.
tiveness is still nowhere near where it ought to be for a blue-water navy. Expertise and operational effectiveness retard PLA Navy effectiveness.

**Expertise**

It is apparent that the Chinese navy is, if not exactly a paper tiger, certainly far less capable than its fleet size would suggest. The level of expertise that the officer corps and crew of the PLA Navy possess is very low by U.S. standards. Inadequate naval education, insufficient training, and inflexible organizational mindset inhibit PLA Navy personnel from using their matériel professionally, efficiently and fully. Expertise, or lack thereof, is the single most important factor hindering Chinese naval advancement (Murray 2009; Cole 2009; China Team 2009).

The PLA Navy lacks a unified naval academy that offers a general course of education (Murray 2009). The naval schools are divided by region and specialization, which results in naval officers receiving educations of varying quality that emphasize different skills. Unlike Western navies, China’s navy lacks a professional non-commissioned officer corps, which serves as a link between commissioned officers and enlisted crew and is the primary source of naval and military instructors of recruits (Cole 2009). Some are not actually recruits, but conscripts, whose terms of service usually last only two years, resulting in high turnover, little skill development, and low returns on investment (Bergsten et al. 2008, 202). The PLA Navy has begun trying to recruit people with technical degrees into its officer corps, as well as trying to set up university programs similar to the Naval Reserve Officers Training Corps in the United States (Murray 2009). These are beginning steps, so their levels of success will not be known for some years.

A second problem, which is related to poor education, is the low level of training. Only ten percent of PRC warships make overseas deployments, doing so once a year (Murray 2009). Cole holds a higher opinion than Murray of PLA Navy officer training, but still finds it lacking: individual sailors’ skills may be adequate, but officers and crew as a whole still face a steep learning curve because they lack of at-sea training and state-of-the-art equipment (Cole 2009). Training 245,000 sailors in a disorganized education system is exacerbated when the ten percent of warships that deploy can only take hundreds (International Institute for Strategic Studies [IISS] 2009, 383). The resulting lack of real-life exercises for many sailors degrades their abilities. Though the Chinese are placing great emphasis on submarines, their sailors’ anti-submarine warfare (ASW) training is “rudimentary”: the tool is sharper than the mind using it (Cole 2009).

The PLA Navy is neither nimble nor flexible. The biggest hurdle for the PLA is the organizational culture and a lack of creative thinking. It is a service with a fear of failure, which is seen as simply unacceptable (China Team 2009). Chinese officials have no enthusiasm for anything that involves free play. In short, there is no questioning, not enough critical thought, and an environment that is not conducive to testing new ideas.

In a ship-to-ship communication exercise to mimic a scenario where a U.S. vessel and a PRC vessel meet in the night and can only communicate by radio, the China Team wanted the United States and PRC commanders to simply talk to one another. This would increase each side’s ability to communicate and decrease chances of misunderstandings or mistakes. It would allow each side to think on its feet. The PLA Navy insisted on scripting out not only the steps of the exercise, but the actual words that each commander would say despite U.S. protestations that this would not resemble reality and would defeat the purpose of the exercise. The PLA Navy’s fear of failure constrains it from trying anything that is not planned in detail as a result, it favors form over substance.

Poorly-trained sailors lack the experience to effectively deploy at great distances and Chinese command, control, and communications capabilities are not comparable with those of the United States. In reference to the recent deployment of three Chinese vessels to East Africa, the China Team notes that it is dealing with a navy that congratulated itself because the three ships it sent half-
way around the world managed to survive the voyage (2009). A PLA Navy Ming-class submarine was not as fortunate in 2003; it was found floundering in the Bohai Sea in northeast China, with the entire crew on board dead at its posts. The belief among the U.S. naval community is that the submarine’s diesel engine was not shut off when the submarine submerged, sucking oxygen out of the vessel in a couple of minutes. That any navy in the twenty-first century could have fatally failed in such an essential action astounded U.S. naval professionals (Murray 2009). That the PLA Navy did indeed fail to execute this elementary precaution reveals just how far it still has to go to school its sailors in the operation of its ships and submarines.

Operational Effectiveness

What can a navy really accomplish if it has a large number of ships manned by poorly educated and inadequately trained crews? All of my interviewees think that PLA naval capability is lacking, and I discuss the constraints that hold the PLA Navy back from operational effectiveness in this section. The first constraint is the fleet’s composition. China does not have a blue-water navy today. It has no aircraft carriers, no overseas bases, no forward-deployed fleets, and no sufficient support infrastructure. Cole (2009) notes that there have been discussions of the PLA Navy constructing an aircraft carrier for some years and the U.S. Department of Defense (DoD) estimates that, if given permission and money, the PLA Navy could build one in the next decade. Yet the PLA Navy faces major hurdles before it can build one. The technology, expertise, and resulting cost required to build the ship itself are immense, but the technology and expertise needed to maintain the shore facilities that would support the carrier are just as complex, particularly so for a navy that would be building its first one.

Carriers are the most complex man-made vessels afloat in history, not only in and of themselves, but for all the unattached

5 Cole strongly disagrees with this “insulting” assessment of the mission, which appears to have been a success.
on more patrols in 2008 than in any previous year: twelve. Moreover, Hans Kristensen notes in "Chinese" that none of the PLA Navy’s three nuclear-fueled ballistic missile submarines (SSBNs) has undertaken a patrol. This is significant because it speaks to China’s nuclear deterrent. A major component of the U.S. nuclear deterrent is that it has nine SSBNs operating around the world at any given time, with four on alert and ready to fire their nuclear warheads on command. Their locations are classified, unlike the locations of other U.S. ships and submarines. This ensures the United States of a second-strike capability in the event of nuclear attack, since no enemy knows where the SSBNs are. In China’s case, no such problem exists, since its SSBNs are never deployed; in fact, they can be viewed dockside via commercial satellite.

Cole cites maintenance as the most important problem facing the Chinese fleet, which is the result of operational deployment and opportunities for training. The Ming-class vessel that sank in 2003 had just come out from shore-side maintenance (Cole 2009). Once at sea, the Chinese navy cannot fix things that the United States does on a weekly basis (Murray 2009). In particular, damage control systems are of poor quality, especially in fire suppression. When Chinese vessels sailed across the Pacific Ocean to visit Naval Base San Diego, U.S. sailors noticed that the PLA Navy had great difficulties with its onboard hydraulic hoses (Cole 2009). Avery Goldstein admits, “issues of training and maintenance set limits” on Chinese naval modernization (2005, 62).

Cole rates the quality of the ships themselves as average (2009). Murray believes that Chinese naval vessels constructed from 2004 onwards are of a much better quality, though still behind the United States, the United Kingdom, Germany, Japan, and South Korea; PLA Navy vessels built before 2004 are, in his opinion, obsolete (2009). Beijing’s navy is seriously constrained in its capabilities, especially when considered what is necessary to project and sustain global power. Were the PLA Navy to build aircraft carriers and requisite support ships and facilities, there would have to be an accompanying increase in education, training, and exercises for the new fleet to be effective.

A final constraint is the issue of command and control aboard PLA Navy vessels. Whatever doctrinal and educational reforms have been implemented, the Chinese military chain of command remains a largely Leninist apparatus. There are two parallel chains of command: operational and political. Cole observes that the shipboard relationship between the senior operational officer (usually a captain or senior captain) and the senior political officer (a commissar, who is equal in rank to the senior operational officer) often depends on the personal relationship between the two. Sometimes, the operational officer dominates; at other times, the political officer dominates. This discrepancy raises a whole host of questions over how much influence one has upon the other in both combat and non-combat situations (Cole 2009).

Despite having many ships and submarines, the Chinese fleet cannot project blue-water power, is not actually designed to do so, is of lower quality than other navies, is not taken out to sea for exercises, and suffers logistical and maintenance issues that impair effectiveness. Sailors have received different educations at different naval academies emphasizing different skills, who do not go to sea to put classroom knowledge to practice, and who operate in a risk-averse environment that discourages initiative. The PLA Navy may be numerically superior in principal combatants to every navy save that of America, but closer inspection reveals that these vessels and the crews that man them cannot compete with the U.S. Navy. The China Team points out that, after the 2004 Indian Ocean tsunami, the U.S. Navy was the first foreign force to arrive with assistance; the geographically closer PLA Navy could not contribute anything to the relief effort (2009). While China’s anti-piracy operations off the coast of Somalia are the first real example of naval projection, it is nowhere near to what would be sufficient. Indeed, the PLA as a whole “has not engaged in significant observable combat for almost 30 years (since 1979 in Vietnam)” (Bergsten et al. 2008, 201). Lack of expertise and low operational effectiveness tempers the numerical challenge that China can ostensibly throw at the United States. Quantity may have a quality all its own, but China has yet to reach

---

6 I thank Prof. Daryl G. Press of Dartmouth College for providing this source.
even that level.

Due in part to this contradicting mixture of quantity and quality, a general Chinese naval threat to the United States is presently non-existent. The matériel is partly there and what is missing can be built. China has purchased old aircraft carriers and, aside from turning them into amusement parks, has sought to learn about the intricacies of carrier construction (China Team 2009). The Pentagon estimates that China could build and operate a carrier before the end of the next decade (DoD 2009, 40). Discussion of when the PLA Navy fleet will gain a carrier has been ongoing for years, with experts each year predicting that acquisition or construction is imminent (Cole 2009). China still lacks a professional, expert officer corps to ensure that any power projection by the PLA Navy is not fleeting, weak and unsuccessful. Unless and until China does so, neither the U.S. professional naval community nor U.S. policymakers will see the PLA Navy as a great threat to U.S. global dominance.

THE TAIWAN QUESTION

The Taiwan question involves the world’s dominant state preventing another state from forcefully occupying an island ruled by a government that lost a civil war. The dominant state, the United States, is obliged by its own law in the form of the Taiwan Relations Act to actively provide weapons for Taipei’s defense. This issue predates the introduction of the power transition dynamic in U.S.-China relations; that is, the Taiwan issue was a thorn in U.S.-China relations before Washington ever considered China to be a potential challenger to U.S. global hegemony. This is important because Taiwan is an island ninety miles off the Asian mainland, so kinetic operations by either the United States or China would have a substantial naval component.

China does not need to project power globally to conquer Taiwan and it does not need to conquer Taiwan to project power globally. This means that the United States could find Chinese naval development threatening, not to general U.S. command of the oceans, but rather, to a specific regional conflict involving one party to which Washington has given a defense commitment. Herein lies the difficulty of gauging U.S. perceptions of Chinese naval power: does the United States believe that PLA Navy development is aimed at U.S. naval power in general, at U.S. naval power in a specific Taiwan conflict, both, or neither? In this section, I show that the current U.S. perception is that the PLA Navy is a force specifically designed to deter local U.S. naval power in the event of a Taiwan conflict. I will also discuss how Taiwan may alter U.S. threat perceptions.

The Impetus for PLA Naval Development

It is the assessment of U.S. policymakers, naval strategists, and most scholars that the PLA Navy’s modernization is solely designed for kinetic operations involving Taiwan. Murray notes that the Chinese navy has built and “bought with thought to accomplish goals,” these goals being, in order of importance: successful kinetic operations against Taiwan’s armed forces, successful deterrence or neutralization of U.S. kinetic operations in theater, and the general defense of the PRC. In Cole’s view, the PLA Navy has, since 2000, focused almost exclusively on Taiwan. Bush states, whatever naval intentions China may have had before the mid-1990s, it quickly directed its focus on Taiwan after the Third Taiwan Strait Crisis, when China fired missiles in 1995 to protest Lee Teng-hui’s visit to the United States and his remarks at Cornell University, and again in 1996, in the run-up to Taiwan’s first direct presidential election. The United States responded by dispatching two carrier strike groups to regional waters, a show of force that revealed, in an eye-opening way, China’s shortcomings vis-à-vis U.S. naval power. The Pentagon states that “a potential military confrontation with Taiwan and the prospect of U.S. military intervention remain the PLA’s most immediate military concerns” (DoD, 2009, 9).

Susan L. Shirk writes that Taiwan is the motivating factor, attributing the increase in PRC destroyers and submarines to a “humiliated” Jiang Zemin’s decision to take a harder stance on Taiwan
Jeffrey W. Legro’s examination of Chinese power leads him to conclude: “it is unlikely that the current scale of China’s military modernization signals more than a desire to protect its version of autonomy, which problematically includes Taiwan” (2007, 518). Alastair Iain Johnston concurs, “There is good evidence as well that military modernization programs, training exercises, and doctrinal innovation . . . are aimed to a large degree at dealing with Taiwanese separation.” He also traces this to the Third Taiwan Strait Crisis (2003, 27). An analysis by Ellen L. Frost, James J. Przystup, and Phillip C. Saunders asserts that the Chinese are focused primarily on “deterring possible U.S. intervention in a Taiwan crisis” (2008, 4). M. Taylor Fravel notes that PLA Navy modernization is not only concerned with “delaying or slowing the deployment of U.S. forces to the theater and potentially frustrating U.S. military operations around the island [Taiwan] if a conflict erupts,” but area denial would also protect “China’s wealthiest provinces and cities such as Guangdong and Shanghai, which could become military targets in a conflict over Taiwan” (2008, 131).

Consensus holds that Chinese naval modernization and Chinese armed forces development are concentrated on the Taiwan issue: preventing both Taiwanese independence and U.S. intervention. From a U.S. perspective, then, the nature of a Chinese naval threat is largely contained to a regional, ad hoc conflict. There is a threat posed to U.S. naval power, but it is not posed to U.S. naval power beyond East Asia. Threat perception, a subjective phenomena, is even harder to measure when the threat itself is limited in geographic and policy scope.

PLA Naval Strength

The PLA Navy’s order of battle suggests that its fleet is geared for a Taiwan conflict and the deterrence of U.S. intervention in such a conflict. Chinese concern over Taiwanese independence (and U.S. support) has led Beijing to develop its naval policy of anti-access and area denial. The PLA Navy has been developing and acquiring ‘mines, submarines, maritime strike aircraft, and modern surface combatants equipped with advanced ASCMs [anti-ship cruise missiles]” to bolster its sea denial capabilities (DoD 2009, 21). The DoD believes that China is “prioritizing the development of measures to deter or counter third-party intervention in any future cross-strait crisis,” developing “the capability to attack, at long ranges, military forces that might deploy (anti-access) or operate (area denial) within the western Pacific” (2009, 20).

The IISS places submarines “at the heart of the [PLA N]avy’s goals, even while debates take place over aircraft carrier capability” (2008, 360). Cole believes that China’s submarines are specifically targeted at U.S. Navy aircraft carriers in the China Seas. Lyle Goldstein notes that Beijing’s focus on submarines results from its intense study of the Falklands War: “The key enabler for almost all other types of operation in the Falklands was underwater warfare, a fact not lost on Chinese observers. The importance of underwater warfare is the primary principle emphasised [sic] in Chinese analyses of sea combat in the Falklands” (2008, 72). The Falklands present interesting comparisons to a cross-strait war with U.S. intervention: a geographically proximate continental state tries to take an island that is backed by a more distant yet more powerful state (L. Goldstein 2008, 66). From the time of the Third Taiwan Strait Crisis through 2009, the PLA Navy’s submarine service has seen net growth of 30 percent from fifty submarines to sixty-five, almost all of this accounted for by the construction or purchase of fifty-four patrol submarines with ASW capabilities and ASCMs. It has also constructed eight ASW support vessels (IISS 2009, 385).

Beijing’s acquisition of twelve Kilo-class patrol submarines with ASW capabilities from Russia are “measures to acquire the capability its armed forces would need to deny the United States Navy access to waters adjacent to the Chinese coast and surrounding Taiwan” (Howarth 2006, 2). He also notes the PLA Navy submarine service-leading role:

[The submarine service] provides Beijing with its most effective sea-denial instrument, well suited to preventing the U.S. Navy from approaching the Chinese littoral
In any crisis or conflict involving the United States and its allies, the PLA Navy’s submarine fleet would play a leading role in an anti-access strategy designed to keep United States naval forces away from the Chinese coast. (Howarth 2006, 97)

U.S. surface ships deployed to waters around Taiwan would operate in a more restricted environment and be more vulnerable to attack.

Granted, finding an enemy ship (even a 96,000-ton carrier) in open waters is quite difficult (Howarth 2006, 102), but it raises an unpalatable scenario for the United States: if China were to sink a U.S. carrier in a war over Taiwan, what would the United States do? Would the U.S. Navy be willing to send another carrier, knowing that the transit time would only give the PLA Navy even more time to bolster its theater area denial capabilities, or would the United States try to defend Taiwan with its remaining submarines, destroyers, frigates, and cruisers in theater?

This focus on submarines merits a brief discussion on diesel submarine capabilities.7 Chinese diesel submarines, which comprise sixty-two of the sixty-five in service, derive distinct geographic benefits from a Taiwan operation. The PLA Navy’s diesel submarines are particularly suited to operating in the East China Sea, which is warmer than most waters. It is also frequently the depositor of sand, mud, and silt brought by the Yangtze and other rivers draining into the ocean. Together, the “surface layer of warm water can trap the signals of hull-mounted sonars [sic], reducing their capability to negligible range against a submarine even at modest depth,” while the silt, sand, and fresh water mix with the salt water to create “varying layers of salinity” that “reflect or refract the sonar beams” (Howarth 2006, 94). The East China Sea’s warmer climate and marine sediment distort sonar beams and prevent other vessels from making accurate readings. Warm, murky waters make an already difficult task—the detection of sub-surface vessels—even harder.

Anti-submarine warfare still relies on acoustic detection methods, which means that diesel-powered submarines can operate “in virtual silence over the full band of sonic frequencies” (Howarth 2006, 93). Murray observes that one could very well be on top of a diesel submarine before detecting it (2009). Diesel-powered submarines can also rest on a seafloor less than 400 yards below sea level, turn off all machinery, and emit no sounds for detection. Nuclear-powered submarines cannot take similar maneuvers: they cannot turn off all onboard machinery and would clog vital condenser inlets if they settled on the East China Sea’s muddy seabed (Howarth 2006, 93). Nevertheless, the benefits of operating diesel submarines have yet to be exploited, given the PLA Navy’s reluctance to deploy them.

Since the PLA Navy is spending its resources on a Taiwan scenario, it has little to devote to building a blue-water fleet. Cole argues that the reason a much-discussed Chinese aircraft carrier has not been built because it is incompatible with the PLA Navy’s primary objectives. Carriers are unnecessary to defeat Taiwan, to deter the United States in a Taiwan conflict, and to ensure national security; the costs of acquiring a carrier and building the shore facilities to maintain it are hard to justify (Cole 2009). The costs may be particularly high for China, considering its lack of expertise in building any (Cole 2009; and China Team 2009). There have been reports of PLA Navy surface warfare officers receiving carrier training, but it has come to light that this carrier training is for helicopters, not fixed-wing aircraft (Cole 2009). The Taiwan contingency appears to actually be holding China back from devoting itself wholeheartedly to a blue-water navy (Murray 2009). U.S. policymakers’ and naval scholars’ opinions are unsurprisingly inclined to the belief that China is building for a Taiwan contingency, i.e., challenging the United States in the East China Sea and the Taiwan Strait.

The irony of the Taiwan situation is that U.S. policymakers and naval experts are sure that China is building its navy against the United States, but in the specific event of a Taiwan crisis, whose unique and self-contained nature makes extrapolation to a general

---

7 The following two paragraphs are from Chao 2009, 32.
dominant state-rising state relationship more difficult. The United States can appreciate why China would build a more powerful navy. This complicates an analysis of U.S. perceptions of China and what role certain variables play in forming those perceptions. That Beijing’s naval modernization threatens Washington in some form is admitted: to deny Washington the operational ability to support Taiwan in a cross-strait war. However, one must ask how much importance the United States should place on a Chinese naval threat that is essentially an ad hoc response to a self-contained dispute that does not threaten vital U.S. interests.

The Taiwan issue also highlights the capabilities-intentions debate. The PRC has had the intention of invading Taiwan for sixty years; the United States has had no reason to spend those years worrying about the PLA Navy. That U.S. naval professionals had not previously considered the PLA Navy a threat suggests that increased PRC capabilities are responsible for U.S. threat perceptions. Ambitions without commensurate capabilities are the classic formula for a paper tiger, be it Washington’s Monroe Doctrine in the 1830s or Beijing’s One-China Policy in the 1970s.

While China can build many ships and submarines, it has not been building a force that could project power around the world in any sustained and effective manner. Furthermore, its ability to operate those vessels is questionable. U.S. policymakers and navalists thus see many hurdles for China to overcome before they will seriously consider a Chinese threat to global U.S. naval dominance. The make-up of the fleet suggests that the PLA Navy is preparing for a war against both Taiwan and the United States. This is inherently a naval threat to the United States, but not an overarching one. It is self-contained to a small island located on the other side of the Pacific Ocean. China’s navy does not threaten the U.S. globally, but it is a local concern in the Taiwan Strait. China’s navy has the second largest principal combatant fleet in the world, but it consists of poorly maintained ships and submarines manned by poorly trained sailors. The PLA Navy is not a threat to U.S. global naval dominance.

In this section, I outline what actions China would need to take for the United States would find threatening to its global naval dominance. I posit that the two primary signs would be the acquisition of matériel and logistics to project power continuously over long distances and the effective and frequent use of those new acquisitions to raise the PLA Navy’s operational effectiveness. The naval power necessary to not only unseat the current dominant naval power, but to take its place, is quite different from the naval power necessary to invade and hold a small island 90 miles offshore. Chinese preparations for a Taiwan contingency could actually serve as a baseline comparison for U.S. perceptions of Chinese naval development. If China were to develop a navy that was not only capable of invading Taiwan, but of Hawaii, for example, then the United States would take notice. As of yet, the Chinese navy’s geographic extension cannot encircle the globe. The PLA Navy has the size to challenge the U.S. Navy, but it lacks the expertise and capabilities. It is clear from my interviews that these two areas hold the PLA Navy back and depress any alarming U.S. threat perception in terms of global naval power. In this section, I discuss the specific changes the PLA Navy would need to make for the U.S. Navy to take a Chinese naval challenge seriously.

Cole sees no evidence that China is building a world-class navy (2009). The United States will not see a Chinese threat to its blue-water fleet unless and until China builds a blue-water navy. The distinguishing attribute of a blue-water navy, as opposed to green- or brown-water navies, is its ability to undertake operations far from land for extended periods of time without land support, i.e., its ability to project power. It operates on the high seas, over the ocean deep; it is capable of being out of the sight of land for weeks at a time. At present, the PLA Navy has no such capability, and not until it gains those capabilities will the U.S. Navy find something to be threatened about.

The most obvious course for the PLA Navy is to pursue an
aircraft carrier program. Murray and Cole both think that China will inevitably acquire one, primarily for the sake of national prestige. In essence, China will get an aircraft carrier because the other major powers do too: China is the only member of the United Nations (UN) Security Council’s permanent five to not have a carrier. Cole believes that, within ten years, China should have a ski-jump aircraft carrier similar to the Royal Navy’s Invincible class, which displaces 21,000 tons (compared with the U.S. Nimitz class’ 96,000 tons). A carrier, as the largest and most advanced weapon of war ever built, affects perceptions in a greater way than any other type of naval vessel.

The PLA Navy could not keep its carrier(s) in port, as it does its SSBNs. U.S. policymakers and naval strategists would see through such inaction. The carrier must be both deployable and deployed and its crew must have the ability to maintain and operate the carrier and its accompanying strike group. The carrier must be deployed in open ocean; sailing around the East China Sea does not demonstrate blue-water capabilities, with land so near. A Chinese aircraft carrier and its support group would have to be able to operate in open water for extended periods of time with no land support; this is what the U.S. Navy expects of its own carrier strike groups and this is what would get the U.S. Navy’s attention.

A strike group would need to accompany the carrier and would consist of cruisers, destroyers, frigates, submarines, and support ships. The PLA Navy would have to build more of these vessels to offer credible protection to a deployed carrier. The PLA Navy has guided missile destroyers and guided missile frigates, but Murray notes that only 10 percent make an overseas deployment. PLA Navy warships have “a really long way to go.” To worry the United States, the PLA Navy would have to deploy its principal surface combatants on a regular basis and deploy them far from home waters. More of the PLA Navy’s submarines would have to deploy more often.

China would need to build a logistics tail with tenders, auxiliary ships, supply ships, and shore support facilities to credibly project power and affect U.S. threat perceptions (Murray 2009; and China Team 2009). Refueling and replenishment ships and capabilities would have to be improved, especially since China has no RAS logistics or support vessels as of 2009 (Murray 2009; and IISS 2009, 385). China would need to demonstrate an ability to keep its ships seaworthy, an ability to feed and supply its officers and crews for long periods of time away from land, and an operational effectiveness that only training and professionalism can provide.

All this is dependent on the underlying assumption that China would first improve its educational and training programs. Sailors need to receive some uniform instruction and need the opportunities to go to sea and actually practice their skills. Academic education and operational effectiveness must be developed; currently, the Chinese navy has too little of both. The atmosphere must be changed, and a professional officer corps must be developed that embraces initiative, creativity, professionalism, and risk-taking. China could very well build a carrier, but if its sailors do not know how to properly use it, it will be of no use. China could send its SSBNs out to sea as strategic deterrents, but if their crews suffer the same fate as the Ming-class vessel of 2003, then the matériel would be useless.

A second way for the PLA Navy to negatively affect the United States’ views would be to establish a network of overseas bases (Murray 2009). Unlike the United States, China does not have territories scattered around the world or defense commitments that would allow it to station naval assets far from home waters. If China were to build up a blue-water fleet, it could obtain bases or basing privileges with strategically positioned states. To secure those bases now would signal great optimism in China’s self-assessment of its naval modernization; it would also be of no practical utility, as the PLA Navy would not have the power projection to use those bases. It would also provide advance warning to the United States that the PRC was seriously developing power projection capabilities.

Power projection is important for China not simply because it is what separates a blue-water navy from a green- or a brown-water navy, but also because it is the only way that the United States
would find naval power threatening. Washington would need to see Beijing gain the capability to reach before the United States could feel a blue-water threat. If the Chinese navy could do that, then it could probably go to many other places that U.S. naval power would find to be an encroachment. China could build a blue-water fleet and keep it close to shore, but that would not threaten U.S. command of the seas.

What China needs to do to negatively alter U.S. naval threat perceptions is at once simple and difficult. It is not too difficult to build ships and submarines, but it is more difficult to build them to a quality matching or exceeding that of the U.S. Navy. It is not too difficult to improve naval education, adopt a uniform curriculum, or make training more rigorous, but it is more difficult to overturn an organizational mindset that is risk-averse and would be loath to overturn much of anything. It is not too difficult to send a ship or submarine out to sea, but it is more difficult to send them out self-reliant, over long distances, and repeatedly. Building for Taiwan is not the same as building for the world.

THE U.S. RESPONSE: A POLICY RECOMMENDATION AND IMPLICATIONS

In this section, I discuss what strategy the United States should pursue if it becomes clear that China is transitioning to a blue-water navy. I argue that the United States should pursue a “capabilities gap” strategy (hereinafter shortened to “cap-gap” strategy) designed to make futile any rising state’s attempt to challenge the dominant state for command of the seas. Rather than close the gap in mutual intentions, the United States should widen the gap in mutual capabilities. I then briefly consider the state of the resulting U.S.-China bilateral relationship and its effects on regional and international systems.

The Cap-Gap Strategy

There is one way out of the security dilemma that would lead to an outcome in the dominant state’s favor: if the potential challenger never develops the potential to challenge the dominant state. China should not build a blue-water navy to contest command of the oceans with the United States. It might do so anyway; it might start, then stop; it might forego the naval challenge altogether. Robert S. Ross notes that continental states that have tried to challenge naval powers “have failed throughout history.” Ross thus raises one reason why China would not challenge the United States: the futility of competition with America (2005, 87).

Military buildup will be met with either a reciprocal military buildup or no military buildup. If there is no Chinese military buildup, then it is possible that China does not seek to dethrone the United States from the dominant position in the world and no attempt at power transition is in the offing. If there is a military buildup, however, then the United States needs to win it. The United States will not win by simply maintaining a small superiority over China; this is reactive and would constantly put the United States on the defensive. Ross may or may not be right; China may decide not to challenge the United States. Whatever Beijing’s ultimate decision, the United States needs to be proactive and make it clear that military competition with the dominant state in the world would be futile. Shaping Chinese capabilities is more important than shaping Chinese intentions: instead of trying to close the two states’ intentions gap, the United States ought to widen the two states’ capabilities gap. This would bring military superiority into the U.S. shaping strategy and add another dimension through which the United States could, if it chooses, persuade through a position of strength.

The cap-gap strategy can break out of the security dilemma. Beijing might respond to Washington, but a reciprocal response does not also mean an equal response. The strategy I propose is concerned precisely with unequal response. Cap-gap explicitly recognizes the security dilemma; it accelerates the security dilemma for the dominant actor, creating asymmetric capabilities so weighted in its favor that the rising state will find the costs exceed the benefits. The challenger could fight with its military inferiority, but
it would be crushed. The security dilemma would be broken when one side realizes that its reaction would be too costly and would not matter.

What does this mean in practice for the United States? The United States must maintain its forward-deployed naval forces, which are the power projection of the U.S. Navy and of the United States as a whole. The United States must maintain its Fifth Fleet at Manama, Bahrain, a symbol of both U.S. power and of U.S. commitment to the oil-rich region. Recently, the DoD stood up U.S. Africa Command, seen as a response to China’s growing (but still relatively small) influence on the world’s second-largest continent. A lesser priority would be the maintenance of the Sixth Fleet at Gaeta, Italy, its prioritization perhaps dependent on regional conflicts in the Middle East and Russian activities in the theater. The most important would be the maintenance of the Seventh Fleet at Yokosuka and Sasebo, Japan, along with the U.S. naval base at Guam and the U.S. presence in South Korea, which is becoming more naval and air-oriented. A Chinese blue-water navy that projects power would contest the United States; therefore, so long as the financing is available, the United States should maintain its worldwide network of naval bases and keep watch over the vital chokepoints and regions of the globe.

The changing U.S. force posture toward a Pacific orientation should continue. Positioning U.S. forces closer to the Chinese homeland would force Beijing to concentrate on self-defense (as well as the balance in Taiwan). China’s two self-defined defensive perimeters include the South China Sea and the East China Sea in their entirety and extend as far east as Guam. This suggests the desired minimum geographic extent of Chinese power projection. The United States is already within both defensive perimeters and there is no suggestion of U.S. forces moving outside of those perimeters. A bolstered U.S. presence in the Pacific, projecting power from Yokosuka, Sasebo, Guam, and Honolulu, could pin China to home waters and keep Beijing’s attention on self-defense. Such a task would be aided by China’s geography: its coastline is unbroken and extends along the east and southeastern edge of the country. The United States could thus place a great concentration of force with relatively few men and matériel.

The purpose of increasing the capabilities gap is two-fold: it is to bolster the U.S. military in general so that China will be dissuaded from competing, but it is also meant to remain relevant should China not be dissuaded. In order to prevent China from exercising worldwide naval dominance, the United States must be able to present a threat in East Asia and the western Pacific, so that if China does develop a blue-water navy (and preventive attack is not an option), then it would be forced to operate on the defensive in home waters against the fait accompli of U.S. command of the seas. Even if China develops a blue-water fleet, the PLA Navy may find itself constrained in projecting power globally, consistently, and credibly. The U.S. Navy would have the opportunity to contain the PLA Navy in the China Seas.

This strategy should be adopted once the political will exists for it but cannot be implemented beforehand, for the same arguments that are made against preventive war can also be made against cap-gap.

Limitations and Drawbacks

Enhancing the capabilities gap is much easier said than done. The costs of building and maintaining a larger U.S. military are very high, especially since U.S. military technology is so far advanced. The United States may face the financial risk of having to temper its military buildup. Current U.S. defense policy seems to be that, in financial straits, geographic breadth is sacrificed for maintenance of overall qualitative superiority. At the expense of other fleets, the U.S. Navy is prioritizing and redirecting new resources to the Pacific theater, specifically, to the forward-deployed Seventh Fleet, though it is probable that the San Diego, California-based Third Fleet will receive a matériel boost as well. This could result in the United States drawing down its forces in the Atlantic Ocean, Central and South America, and the Mediterranean Sea.

The political and strategic implications of such a move are
unknown, but may include steps by other states to fill the power vacuum. The United Kingdom might decide to revive its naval force if it thought the United States had become too preoccupied with East Asia. In Africa and the Middle East, where no state is building a navy or seeking maritime hegemony, one might see a security vacuum remain unfilled. If maritime piracy were to remain or re-emerge as a problem during a U.S. drawdown, it would be up to other states, most likely Russia, the wealthy European states, and states in the vicinity of the piracy to respond effectively without heavy U.S. contribution or cajoling. Reduced naval capabilities would also affect U.S. humanitarian relief efforts, citizen evacuation efforts, and general power projection. Of course, if the U.S. economy grew at such a rate that the United States and its navy could increase their superiority over China and its navy while simultaneously maintaining, even increasing, its commitments elsewhere, then this is a moot point.

The U.S. Navy has proposed a 313-ship navy that is supposed to be ready in less than thirty years. At current shipbuilding rates, this target will not be met. This highlights another problem for the U.S. Navy: timely, efficient, cost-effective execution. The U.S. Navy’s recent shipbuilding projects have been “embarrassing” fiascoes: the Zumwalt-class guided-missile destroyer’s costs have soared so far beyond projections that the Navy’s original order for thirty-seven Zumwals has been cut to three; the littoral combat ship (LCS) has experienced dramatic cost overruns as well; and the newly-completed United States Ship (USS) New Orleans amphibious transport dock failed its inspection because it was deemed unfit for transporting men and matériel (China Team 2009; and Kreisher 2009, 41). One problem is that two companies hold a monopoly on U.S. Navy shipbuilding contracts, while another problem is that ships are being built before the designs are even finished (Taubman 2008). The solutions are to break up the monopoly and to actually plan vessels out before starting to build them. These are simple solutions, though not necessarily easy ones.

Money is the foundation of military might. China, with an economy that can potentially challenge that of the United States both in size and dynamism, could very possibly present a more potent challenge than the Soviet Union did in terms of financial strength. The U.S.-Soviet contest was an endurance race; it would appear that the U.S.-China case could, in the economic sense, prove to be more of a sprint. If this is the case and if the U.S. economic foundation were to be sufficient, then a cap-gap strategy implemented at the right time could leave China in the dust. If the economic foundation were insufficient, the United States would exhaust itself and collapse, regardless of whether or not it crossed the finish line before China (if it crossed the finish line at all). A strong economy is necessary for a strong military, especially one used to project power. The cap-gap strategy is predicated on the U.S. ability to maintain this balance; otherwise, it would be utterly pointless for the United States to pursue a strategy that so profoundly disrupts the economic-military balance as to render that strategy’s outcome irrelevant by sealing the dominant state’s fate regardless.

Alternative Policies and Scenarios

The United States should not launch a preventive attack. Such an operation would be extremely risky, with a less than desirable chance of success and the very great possibility of blowback that would affect U.S. positions throughout East Asia and could potentially destroy the regional allies that Washington would rely on for support if and when relations with Beijing deteriorate. The political repercussions would be massive and, if the Chinese government does not respond with a force commensurate to the nationalist outrage of its citizens, China may be plunged into chaos. The possibility of a hard-line, jingoistic military leadership supplanting civilian officials is not palatable. The United States should be careful of the regime change it sometimes wishes to see.

The United States should not adopt a containment strategy at this time. For one, it would be very difficult, as regional allies seem reluctant to give up their own engagement or hedging strategies. As noted previously, the dominant state has unique fears from a challenger that smaller states either do not worry about or lack
the means to solve. It would be very difficult for the United States to ask other states to join it in containing China when the threat is unclear and when regional states’ geographic proximity means that they would bear the brunt of any hostile Chinese response to a containment coalition.

The United States should not adopt a policy consisting solely of appeasement or engagement. The United States derives great benefits from the international system that it created and its trade with the world depends overwhelmingly on its control of the oceans, sea lanes, and chokepoints of the world. If the United States were to appease or solely engage China, it would find its position in and benefits from the international system decreasing at an alarming rate and through its own doing. While the current international system is one of America’s making, the world is not. This is a world in which other states have conflicting goals and varying abilities to achieve them. While goals may conflict with those of the United States, capabilities should not.

The United States should not disregard geopolitics, power projection, and military superiority because economic integration has rendered major war highly unlikely in the globalized world. This is conjecture and has not yet been proven. Global power projection demands prudence and caution, not untested faith. Beyond this, dominant state-rising state relations do not need to worsen to the level of war for the dominant state to feel pressured, disadvantaged, and weaker, perhaps to a simultaneously unacceptable yet irreversible level. The United States would find that an unpalatable military balance could have repercussions on other issues, including diplomatic initiatives, alliances, and general influence. Paper money and copper coins, while (and because they are) valuable, ought to have some support by firepower.

**WIDER IMPLICATIONS OF HOSTILE RELATIONS**

In early 2009, China completed its first successful overseas deployment of naval forces since the fifteenth century. Shortly after and closer to home, China harassed the United States Naval Ship *Impeccable*, an oceanographic surveillance vessel, operating in the South China Sea. The United States dispatched the USS *Chung-Hoon* to escort the otherwise defenseless *Impeccable*. In late April, the PLA Navy celebrated its sixtieth anniversary with an international fleet review, featuring twenty-five PLA Navy vessels and twenty-one vessels from fourteen other navies. Events of 2009 suggest that the naval variable will continue to figure prominently in U.S.-China relations. But what is the overall relationship to look like? What is the endgame for Washington and Beijing in the twenty-first century?

The U.S.-China relationship will be hostile, though not necessarily warring. As China continues its military buildup and develops abilities far too capable for its stated goal of self-defense (assuming Beijing continues to state this justification), the United States will adjust its China and defense policies in an attempt to maintain superiority. Outright hostility might be tempered with practical negotiations and diplomatic exchanges, or careful hedging could not temper a perception of unabated advances. Relations between Washington and Beijing may carry on as a necessity, though probably not out of any feelings of goodwill.

A souring U.S.-China relationship would affect each state’s relations with regional powers. The United States may opportunistically use a regional dispute involving China to constrain Beijing. The most obvious case is Taiwan, but the United States could also encourage Japanese militarization. Should China become involved in a war in East Asia and should America decide to intervene, Washington may enter on the anti-Chinese side. It might actually be beneficial for the United States to encourage hostile relations between China and its land neighbors, as Paul Kennedy noted in the case of 18th-century Britain and France (1987, 88–89, 97–98). By keeping China preoccupied with a continental geostrategy, the United States may forestall, if not entirely scuttle, Chinese plans for a maritime geostrategy. At the least, China would have to try and execute both geostrategies simultaneously. Depending at what stage of PLA naval development the United States reacts, there is the potential for reciprocal meddling by China in the Americas.
The broader international system may see a worldwide battle for power and influence. This might not be the Cold War of the twentieth century, especially if the disparity in U.S. nuclear capabilities and Chinese nuclear capabilities is maintained, but it could take on the global dimension of the United States’ struggle against the Soviet Union. Global power projection must be combated on a geographic scale larger than most other arms competitions. The world could, as a result of this competition, be divided into U.S. and Chinese spheres of influence. If Beijing managed to send forth a navy to compete with Washington’s, one would see a more evenly divided international system; if Washington successfully destroyed a Chinese navy or prevented it from reaching a geographic breadth comparable to that of the U.S. Navy, one might see a twenty-first century geostrategically divided much like the twentieth, when the United States still managed to maintain overall naval superiority, leaving its adversary with a largely continental empire.

CONCLUSION

In this section, I offer some final thoughts on larger issues affecting and affected by the U.S.-China case study. I discuss the relationship between economic power and power projection and how the latter is simply impossible unless the former exists simultaneously. I discuss the elusive presence of preventive attack in power transitions and what role the nuclear balance plays in it. Finally, I address the inter-related issues of the security dilemma, hedging, and Taiwan in U.S.-China relations.

Economic Strength and Power Projection

A state’s military expenditures are constrained by the amount of money available and by competing priorities. Power projection is an expensive endeavor, and it is perhaps fitting that the dominant state in the international system, with its weighty economic power, should also be a formidable military power. The ability to maintain global naval capabilities is not cheap: the United Kingdom discovered in the years preceding World War I that there was a limit to how long it could withstand industrial Germany’s naval buildup, while the U.S. Navy, facing economic malaise in the 1970s, languished as its Soviet rival gained ground.

The current economic crisis may have great implications for U.S. defense expenditures, particularly since recent shipbuilding fiascos involving the Zumwalt-class destroyer and the LCS have not endeared the U.S. Navy to the Congress. While the PLA Navy’s budget appropriations have increased, they have not increased relative to other branches, but only as a reflection of across-the-board increases resulting from a booming economy (Cole 2009). It remains to be seen how the current global financial downturn will affect the U.S. Navy and the PLA Navy in the long-term. If the effects are largely similar in breadth, depth, and duration, one may see the United States emerge from the economic crisis with a relatively similar level of naval superiority over China as before the economic crisis. If China’s suffering is greater than America’s, then the mere possibility of a Chinese naval challenge to the United States might be postponed for some years. If, however, the United States emerges worse from the conflict than China does, one may see the United States respond to China with a greater lag. The issue would then become whether the United States could respond before the PLA Navy reached the point where any U.S. response would be deemed too expensive and a lost cause.

The United States is currently the only state that has true power projection capabilities. Murray notes, in his research, he has not come across anyone discussing, let alone suggesting, that the United States acquiesce to a rising PLA Navy. He personally conceives of only a domestic issue affecting U.S. force posture in East Asia. If the United States were to suffer a “terrible depression” or some other broadly-accepted “economic, political, or environmental crisis,” Washington would have to reorganize its fiscal priorities, at which point the U.S. Navy’s overseas commitments may be cut (Murray 2009). Cole (2009) also notes that the U.S. response would ultimately depend on its financial means, but that, so long as money was not a problem, the United States would respond with a
robust program to remain ahead of China.

This would greatly affect the United States and China in a competition for global power. Already, concerns over growing Chinese influence in places such as Southeast Asia, the Indian Ocean, and Africa have led to calls for a renewing of U.S. power and influence in those regions. If the United States were to face an economic downturn while China continued its growth, Washington would be unable to project its power and maintain its influence on a level that could compete with Beijing. The three forward-deployed fleets of the U.S. Navy — in Italy, Bahrain, and Japan — could be reduced in strength, if not scrapped altogether, leaving U.S. allies vulnerable to both attack and bandwagoning pressures from the rising state. A healthy economy at home is essential for a robust naval presence abroad.

The Nuclear Balance and Preventive Attack

Nuclear superiority increases the probability of success in a preventive strike, particularly if the target state’s second-strike capability could be neutralized. U.S. military officers argued for preventive war against the Soviet Union while they still possessed nuclear superiority. Nuclear parity made preventive war a less palatable option, but it did reinforce the importance of conventional forces. Keir A. Lieber and Daryl G. Press (2006), however, argue that the United States has such nuclear superiority today that it could neutralize China’s entire nuclear arsenal if necessary. Assuming that were true, why does the United States refrain from preventive destruction of the PLA Navy, especially since its three SSBNs do not operationally deploy? The answer may be as simple as that it is contrary to current U.S. policy; therefore, U.S. policymakers, who make and implement policy, will not undertake steps to undermine that very policy. Current U.S. policy is to hedge against and shape China precisely in the hope that the future will not end in war. Like détente in the 1970s, preventive war is simply one of many methods of policy execution that do not agree with the U.S. policy being executed.

As straightforward as the policy inconsistency argument sounds, there are other reasons. It is notable that the United Kingdom, in the pre-nuclear era, also refrained from launching a preventive war. There must be something more than a simple fear of mutually assured destruction at play. The first issue may be a threat-reaction gap. Preventive war, by definition, is aimed at something that is not currently threatening, but may become threatening in future. It would be psychologically difficult to both justify an attack on another state’s potentially threatening yet currently benign naval force and to justify doing so with a preventive attack. The human mind may simply find the concept too jarring and difficult. States usually do not engage in defensive wars against non-existent threats, so for a state to be told that it must go to war precisely because the threat it will face does not yet exist seems very presumptuous, if a bit nonsensical. Fighting a threat that does not (yet) exist is hard to justify.

The dominant state’s wariness in launching a preventive war may also be a product of the dominant state’s role in the international system. The dominant state represents the status quo and seeks to ensure stability. This may explain the relative rarity with which dominant states launch preventive wars. Since preventive war is a reaction to a currently non-existent threat, launching a preventive war is a proactive decision, not reactive. By definition, an extant threat can only provoke a reactive response. The dominant state is a reactive state: it has climbed to the top of the greasy pole and sees no reason why it should harass other states unless they challenge it.

There are many other difficulties that present themselves when planning a preventive attack: intelligence and counterintelligence capabilities, logistical coordination, operational mistakes and setbacks, other kinetic options, other grand strategies that do not espouse preventive attack, and political repercussions and blowback. If, despite all these considerations and the “fog of war,” a preventive attack on a growing navy were successfully executed, that would still leave the target state in existence, since destroying its navy would not entail destruction of the state. The target would
be free to retaliate, even without nuclear weapons. China could launch an invasion of Taiwan; create problems in North Korea; attack Japan, South Korea, and other regional allies; involve itself in Afghanistan; fund anti-American terrorists; commit state terrorism, such as killing Americans in China after the preventive attack; or do a number of other things that could harm the United States. Washington would have to be prepared to deal with any and all of these responses, perhaps simultaneously.

The Security Dilemma and Hedging

As the United States and China are unsure of the other’s intentions, but see growing capabilities, the result may well be the classic security dilemma: a spiraling armament race. Suspicions are already evident. Murray notes that the PLA Navy has been “cagey” in its interactions with the U.S. Navy and that it is unenthusiastic about naval understanding between the United States and China (2009). The Pentagon’s annual report emphasizes what it perceives to be a lack of Chinese forthrightness and what the U.S. response has had to be:

China has begun a new phase of military development, . . . but has left unclear to the international community the purposes and objectives of the PLA’s evolving doctrine and capabilities. . . . The limited transparency in China’s military and security affairs poses risks to stability by creating uncertainty and increasing the potential for misunderstanding and miscalculation. The United States continues to work with our allies and friends in the region to monitor these developments and adjust our policies accordingly (DoD 2009, 1).

The China Team also complains about the lack of transparency, trust, and openness from its Chinese counterparts, echoing Murray’s view that they are not genuinely interested in mutual understanding. Chinese officials go through the motions and undertake joint ventures for the sake of doing them. This is partly seen in the PLA Navy’s proposals, which usually suggest complex activities but provide little to no detail, so they are either not carried through or, if they are, it is because the China Team has pushed and prodded for the details necessary to execute the exercise. What U.S. policymakers view as strengths—U.S.-China cooperation and jointness—China views as weaknesses. For the China Team, one of its basic issues is not so much that China is acquiring matériel to “blow the United States out of the water,” but that China denies such an “obvious fact.” The 2009 edition of Military Power of the People’s Republic of China concludes that Beijing “has improved modestly the transparency of its military and security affairs,” but that it sees transparency as more of a “transaction to be negotiated” and less of a responsibility (DoD 2009, vii).

The security dilemma may play a very large role in future U.S.-China relations. Indeed, it may have already begun. Murray notes that the U.S. Navy has already begun to shift its focus from the Atlantic Ocean to the Pacific, with the three Seawolf-class nuclear-powered attack submarines already serving in the latter and the newer Virginia class also marked for service with the Pacific Fleet (2009). Kristensen writes in “U.S.” that the U.S. Navy’s SSBN patrols have continued at levels comparable to the Cold War, with the majority occurring in the Pacific Ocean. The U.S. Navy is also planning to transfer more carriers to the Pacific (China Team 2009). In 2009 testimony before the Congress, Adm. Timothy J. Keating, then-commander of U.S. Pacific Command, referred to Chinese naval development as the reason why ASW is the U.S. Pacific Fleet’s “number one priority” (U.S. Senate 2009, 25). As with carriers and submarines, each increased Chinese capability would be met with a U.S. response to counter the threat.

In the specific situation of Taiwan, the security dilemma is unavoidable, given U.S. strategic ambiguity. By deliberately avoiding a firm commitment either way to the Taiwan question, the United States not only compels China to build up capabilities for a Taiwan crisis, but is itself obliged to respond to China’s developments in order to maintain strategic flexibility. At the same time, by leaving
open the door for intervention in the Taiwan Strait, Washington
ties down Beijing by forcing it to devote an inordinate amount of
time, money, and matériel to an island one-half the size of West
Virginia. Were China free of the Taiwan issue, it might embark
wholeheartedly on a navy with power projection capabilities and
have already presented a challenge to the United States. One could
thus argue that strategic ambiguity, while fueling an arms race
across the Taiwan Strait, has dampened and delayed an arms race
on a wider, general scale.

There is, however, one problem. If the Taiwan issue were the
prism by which the United States or China viewed the entire bi-
lateral relationship—if Washington or Beijing were to project its
overall threat perception of the other based on what happens with
Taiwan—then strategic ambiguity, far from being a salve, becomes
an infection. It should not be a shock that China, with its extensive
coastline, should build some sort of navy, but its capabilities and
level of threat could be blown out of proportion, creating the pos-
sibility that it would not matter what China’s geostrategic focus re-
ally was because Taiwan would make it appear maritime in Amer-
ica’s eyes. Given the gross disparity between China’s resources and
Taiwan’s resources, it is simply a matter of time before the cross-
strait military balance tips in Beijing’s favor (if it has no already).
The United States’ role would then become even more vital, plac-
ing greater importance on U.S. capabilities in theater to credibly
uphold strategic ambiguity, while the perceived interest gap over
Taiwan between the United States and China may make Beijing
angrier at Washington’s continued refusal to retract military pro-
tection over something that the latter should care less about.

The true importance of the Taiwan question now comes into
focus: since China views Taiwan as a question of basic territorial
sovereignty and Chinese Communist Party legitimacy, it is very dif-
cult to see how Beijing would halt its military buildup in Fujian
unless Washington were to renounce its defense commitments.
There is thus the potential for a security dilemma (Taiwan) within
a security dilemma (overall U.S.-China relationship). Beijing has
not halted its military buildup with Ma Ying-Jeou’s accession to the
Taiwanese presidency (Bush 2009). The United States, to maintain
strategic ambiguity, must build up in response to Beijing and nega-
tively change its threat perception, which would lead Beijing to do
the same, and the spiral would take on a life of its own.

It is possible China may not escalate. Ross explicitly posits a
naval arms race as the reason for China’s reluctance to proceed with
a carrier program:

Thus far, China’s leadership has resisted the temptation
to acquire this prestigious symbol of great power status.
... China’s leaders are seemingly aware that a Chinese
carrier ... would challenge U.S. maritime interests and
induce an arms race with the United States that China
could not win. (2005, 86–87)

Cole (2009) also notes the possibility of the PLA Navy removing
itself from the security dilemma because a naval race against the
United States would simply be too hard, though he also thinks the
carrier delay is due to its operational uselessness against Taiwan.
This would signal that China would be content with a continental
sphere of influence in Asia, which would not clash with the largely
U.S. maritime sphere in the region (Ross 1999; and Ross 2003).
Although my interviewees see the existence of a Chinese carrier
program as inevitable, this does not mean that Ross is necessarily
wrong. One carrier is frankly not a security threat; multiple carri-
ers would be more threatening. If the PLA Navy were to construct
one carrier, see a robust U.S. response, and refrain from construct-
ing anymore, Ross’s assertion would still stand. Only his timing of
when China would realize a carrier’s dramatic effect on U.S. per-
ceptions would be wrong.

The security dilemma and its self-fulfilling nature have im-
plications for the current U.S. strategy of hedging, which is a com-
bination of engagement and dissuasion. Engagement on its own is
solely an attempt to influence and shape the rising state’s chang-
ing position in the international system. It is reflected in the U.S.
détente policy toward the Soviet Union, in which nuclear parity
was acknowledged and conventional superiority was allowed to deteriorate. The policy of dissuasion is the maintenance of military superiority with the purpose of persuading potentially hostile states to refrain from initiating arms races. Dissuasion is not deterrence. Whereas dissuasion is applied to states that may or may not be hostile, deterrence is applied to states that are assuredly hostile. The practical differences between the two lie in the breadth, depth, and intensity of the military buildup. Deterrence is also exclusive of engagement, for if shaping a state’s actions is still possible, then a simultaneous pursuit of deterrence is presumptuous and self-defeating. Dissuasion is thus a self-strengthening movement taken when another state’s hostility is possible, even plausible, but not definitive.

The dissuasion component of the U.S. hedging strategy is something not seen in the strategies of other states that may choose to hedge against China. Singapore, for example, could choose to pursue a hedging strategy against China, but regardless of whether China builds up a military or not, the Singaporean military could only be a small part of the entire hedging strategy. If China does not build up its military, then Singapore has no need to improve its military. But if China builds up its military, no plausible Singaporean military buildup could defeat the Chinese military in a confrontation. Thus, for many states, dissuasion would be either a useless or an unnecessary component to China policy.

For the United States, however, the situation is different. America is the dominant state in the world and possesses the world’s best military, including the navy. It is an inescapable fact that the dominant state of the international system, almost by definition, possesses such kinetic superiority. Assuming the rising state is building up its military simultaneously, any dominant state that pursues a hedging strategy against the potential challenger must place great importance on military armament in its hedging strategy if it wishes to remain dominant. The dominant state is forced to be a reactionary power: if the rising state builds a military, the dominant state has to respond, assuming it wants to stay dominant. This forced response, if met by the rising state with an escalatory response, initiates the security dilemma.

Engagement is the other half of the U.S. hedging strategy, and while the U.S. experience with détente shows that engagement by itself is a foolhardy venture, it would seem to be less so when tempered with dissuasion. The problem is that dissuasion does not just temper engagement, it kills it. When the dominant state and the rising state undertake reciprocal arms buildups, engagement is taken over by talk about the arms race, as early twentieth-century Anglo-German relations show. If two states can have an engagement, then dissuasion is the pre-marital agreement of international relations. Precautionary steps to guarantee one’s well-being in a potential worst-case scenario—while realistic, prudent, and far-sighted—may also be self-fulfilling. Hoping for the best and preparing for the worst keep the former from becoming reality and make the latter seem prescient. If the United States continues on its hedging strategy toward China, it will find that its attempt to shape Beijing into a responsible stakeholder in the international system will not succeed. The very dissuasion that will cause engagement to fail will also have prevent engagement from being seen as a greater failure than it could otherwise be, for dissuasion would mean that the United States will have maintained military readiness.

The driver behind the dominant-riser dynamic is the rising state. Recognition of this fact may explain why dominant states try to “shape” the rising state, in the hopes that the rising state will start to act in a manner more conducive to the dominant state’s desires, thus allowing the dominant state to stop reacting in ways unfavorable to international stability. This also means that, if China does challenge the United States for world power, then the reactive decision to hedge is unavoidable on the dominant state’s part. The United States would have chosen a failed strategy, but this failed strategy would be the most optimistic available when compared with outright dissuasion, deterrence, or containment. Hedging would still afford China the opportunity to halt its challenge and become a “responsible” (i.e., subordinate) stakeholder in a U.S. international system.

This suggests that a national policy of hedging is doomed
to fail in a dominant state-rising state dynamic because it all but guarantees a security dilemma; it is hard to see how the engagement component of the hedging strategy can overcome increasingly hostile military relations. The exception would be for one of the states to voluntarily halt the arms race regardless of how the other state responds: either the dominant state would take the chance of being overtaken or the rising state would cement its status as a potential challenger and not turn into a bona fide challenger. The United States has tied much of its changing force posture in the western Pacific Ocean to Chinese military development, implying that Washington is merely reacting to Beijing’s unsatisfactorily-explained actions.

The real danger of the security dilemma, aside from the dilemma itself, is the product. China may be spurred on to develop a blue-water navy precisely in response to U.S. efforts to prevent it from doing so. Even if China did not consciously seek a blue-water force, it is difficult to see Beijing get mired in an arms race with the United States and still be content to produce small patrol craft. As the arms race intensifies, the threat looms closer, and the need for larger and more capable vessels grows. It is possible that China could simply build up hundreds of patrol craft and other small vessels, but that is not what past ascendant powers have done. Major powers have major warships. The United States, in such an arms race, would have to consider its rate of construction as well, but it would face the problem of its fleet being dispersed around the world. China does not have such a problem and can concentrate whatever forces it builds. If the United States cannot keep pace with Chinese shipbuilding and its superiority erodes as a result, it may have to call on alliance partners, redistribute the fleet, or both.

For much of history, but before the United States achieved any status as a major power, China was a powerful state. Granted, much of its dominance was confined to Asia and it has never been the dominant state in the international system, but China’s past two hundred years are actually the exception, not the rule, in its unbroken history. China was, simply put, great. It may feel that it is time to be great again. China currently has a large navy: it is an unimpressive fleet. The vessels themselves are of varying quality and operational effectiveness, while their officers and crew lack high and uniform standards in professionalism, initiative, and training. These weigh down the PLA Navy more than any anchor could. At the same time, Chinese naval development has focused on confrontation with Taiwan and on keeping the United States out of that conflict. Clear signs that China wants to keep the United States out of more than just the Taiwan Strait have not been forthcoming: “The absence of a true expeditionary logistics capability . . . will limit the PLAs’ ability to project and sustain military operations at locations distant from the mainland” (DoD 2009, 38). Scholars such as Avery Goldstein who raise the apparition of a Chinese blue-water navy may be right someday, but not today. If there is a Chinese naval threat in the form of blue-water capabilities and global power projection, it is still a latent one.

The current U.S. strategy for China, while mixing optimism with caution, may find that optimism misplaced and that caution deepening to anxiety. For now, America does not face multiple challengers, as its British predecessor did. It does not face the mutually assured destruction that it did last century. It does not yet face the blue-water navy, the challenge to power projection, and the threat to international dominance that the Soviet Union presented last century. Will China try and succeed where the Soviet Union failed? The world waits and watches.

Works Cited


INTRODUCTION: A PORTRAIT OF OSH, KYRGYZSTAN

Osh, the “southern capital” of Kyrgyzstan located in the ethnically rich Ferghana Valley, is a network of streets and lanes gradually rising into a tree-lined urban landscape from surroundings of yellow hills and flatter expanses of farmland. Even from an initial cursory glance, unlikely juxtapositions perhaps only to be found in this region of the world leap out. Walking a straight course along, say Ulitsa Lenina or Kurmanzhan-Datka, one could travel from the busy bazaar where vendors sell melons, flat rounds of tandoor bread, and hard balls of dried yogurt, past Soviet blocks of apartments and smaller alleys leading into *maballas* of courtyard houses, to the main government “white house” across from an imposing statue of Lenin. Osh’s face seems to be that of “two cities” which “captures the coexisting presence of divergent orientations and aspirations within the city.” But even a two-sided characterization dividing this urban space into a Soviet sector and a “traditional Central Asian” sector is too simple (Liu 2007, 66). Seemingly contradictory beliefs, habits, and identities overlap and spill out of bounded domains, manifesting themselves in various aspects of everyday life in Osh, whether in terms of religious identification, ethnic connections, national sentiments, linguistic choice, or merely commonplace practices.

In these interviews, language reveals itself to be a window onto the forces that shape cultural and national identity: it illuminates dialogues of power within a society and shapes politics, builds national solidarity, and underscores group divides. It is malleable, and

---

Grace Zhou, a senior at Columbia University, is majoring in Linguistics and concentrating in Anthropology. She is interested in post-Soviet regions and greater Eurasia. This summer she will be journeying back to Osh, Kyrgyzstan, to learn Uzbek.
INTRODUCTION: A PORTRAIT OF OSH, KYRGYZSTAN

Osh, the “southern capital” of Kyrgyzstan located in the ethnically rich Ferghana Valley, is a network of streets and lanes gradually rising into a tree-lined urban landscape from surroundings of yellow hills and flatter expanses of farmland. Even from an initial cursory glance, unlikely juxtapositions perhaps only to be found in this region of the world leap out. Walking a straight course along, say Ulitsa Lenina or Kūrmānzhān-Datka, one could travel from the busy bazaar where vendors sell melons, flat rounds of tandoor bread, and hard balls of dried yogurt, past Soviet blocks of apartments and smaller alleys leading into mahallas of courtyard houses, to the main government “white house” across from an imposing statue of Lenin. Osh’s face seems to be that of “two cities” which “captures the coexisting presence of divergent orientations and aspirations within the city.” But even a two-sided characterization dividing this urban space into a Soviet sector and a “traditional Central Asian” sector is too simple (Liu 2007, 66). Seemingly contradictory beliefs, habits, and identities overlap and spill out of bounded domains, manifesting themselves in various aspects of everyday life in Osh, whether in terms of religious identification, ethnic connections, national sentiments, linguistic choice, or merely commonplace practices.

In these interviews, language reveals itself to be a window onto the forces that shape cultural and national identity: it illuminates dialogues of power within a society and shapes politics, builds national solidarity, and underscores group divides. It is malleable, and

Grace Zhou, a senior at Columbia University, is majoring in Linguistics and concentrating in Anthropology. She is interested in post-Soviet regions and greater Eurasia. This summer she will be journeying back to Osh, Kyrgyzstan, to learn Uzbek.
yet powerful in its malleability. To understand the variety of existing attitudes and sentiments towards the different languages spoken in Osh and, on a more general level, in Central Asia, we must examine the historical trajectory of nationality policy. The legacies of Tsarist administration and Soviet nationality policy in attempting to consolidate language and identity in Central Asian republics may have directed identity development in a particular fashion and shaped the official methodology of nation-building after the collapse of the Soviet Union, but they did not undermine the region’s diversity; in fact, they may have highlighted it.

In my attempt to trace the complexities of identity in Central Asia, I first discuss the theoretical background of nation-building and language planning, then the historical essentializing processes of nationality during Tsarist and Soviet times, both of official policy and native involvement. I then bring the discussion into post-Soviet times by examining how essentialist Soviet legacies are maintained in independent Kyrgyzstan. Finally, I look at a case study from my field research in Osh. Osh, as a reflective microcosm of Central Asian society, illustrates a linguistically dynamic conception of identity. While explicit, verbal identification is significant to notions of ethnicity, the coexistence of multiple languages within the same sphere of usage has produced an interesting phenomenon, that of habitual code-switching. In the last section of this paper, I address code-switching, along with other everyday sentiments and attitudes, in an attempt to dig under official discourse and break the essentialist model.

THEORIES OF NATION-BUILDING AND CREATION OF NATIONAL LANGUAGES

The point must first be made that idea of the nation-state—the right of each unified, distinct national entity to sovereignty—is a fairly modern concept emerging out of a specifically Western historic and cultural context in the nineteenth century. Moreover, theorists, most prominently Benedict Anderson (1983), argue that nations themselves are “imagined communities.” If unifying con-
nections are to be found on as wide a scale as the national level, then those linking threads must be forged through consciously standardized narratives of shared history and authenticated heritage. This paper is concerned with the process of nation-building in these terms, i.e., creating and promulgating an official discourse of national legitimacy and collective identity rather than the building up of state capacity and infrastructure. Language plays a key role in this cultural imagining of community. While formal Saussurian linguistics make a strong distinction between the primacy of speech over writing, when it comes to examining the political power of language, writing plays an equally, or perhaps even more, crucial role as spoken language. In a world of over six thousand spoken tongues, there are only two hundred written languages. Writing provides fixity, slows down language change, and is thus conducive to standardization and the spread of literacy. In instances of language planning and consolidations of national identity, it is writing that exerts force and fuels these processes. Anderson argues that what really swept this movement into force was the standardization of written tradition with new print technologies which widely disseminated a uniform system of written language and literature. Because all languages have degrees of dialectical variation, at some point in the history of any modern nation a moment must occur when a centralized government seeking to unify regionalized territories attempts to create a national standard language, and, through this, a unified national identity.

There are several different ways to go about establishing a standard language, and all are aided by the creation of a strong writing system, which gives fixity to amorphous orality. One method is to base the standard on a classical historical form. This normally occurs with languages with rich written traditions. For example, Modern Standard Arabic is based on the seventh to ninth century classical Arabic of the Quran and other Islamic literatures. Sometimes the creation of a standard can be completely devoid of vernacular influences. Standard German until about the nineteenth century existed almost completely only in writing; its existence comes from the molding of written language over centuries by writ-
ers who sought to write in a way most comprehensible to speakers of the different German dialects. A standard language can also be constructed by creating a more egalitarian blend of its different dialects; or an attempt may be made to trace back to an ancient “proto-dialect”. One frequent method of standardization is the emergence of one dialect’s dominance over other rivals. Sometimes this happens through more natural processes—often under the influence of writing. The Italian standard language, for instance, is based primarily on the Tuscan dialect as a result of the influence of Dante’s writings.

Sometimes the dominance of one dialect over the others is enforced by centralized authority. China, despite its projected identity as a solid and unified ancient culture, has undergone various attempts at unification, and thus many movements for language standardization. The Beijing dialect became the standard in the early twentieth century through concerted efforts by the Commission on the Unification of Pronunciation. The spread of this standard was the result of its mandatory instruction in all schools across the country. In the case of Kyrgyz, Soviet planners methodically sought to distinguish it from other Turkic languages, its linguistic neighbors. Robert Lowe writes that “language was seized upon as a key element of identity and the underdeveloped Kyrgyz tongue was given a script (Arabic, then Latin, later Cyrillic), an expanded vocabulary, grammar, dictionaries, literature, and other elements necessary to declare it formally as a separate Turkic language” (Lowe 2003, 109–110). Creating national identity through such means as standardizing language basically essentializes identity. Essentialism takes the perspective that “those who occupy an identity category are both fundamentally similar to one another and fundamentally different from members of other groups.”

While the theoretical position of essentialism is deconstructed and criticized in academic discourse as it does not take into consideration the overlapping and constantly changing nature of identities, a mentality of essentialism nevertheless guides the goals and methods of nation-building. A collective identity is solidified in opposition to a foil of alterity, an Other to heighten the sense of
sameness within a group. Heterogeneous environments are therefore conducive to strong group formations due to repeated contacts with differentiated Others.

HISTORY OF NATION-BUILDING AND LANGUAGE PLANNING IN CENTRAL ASIA

The impulse linking standard language with standard literature as the soul of national consciousness, discussed in early nineteenth century texts such as Fredrich von Schlegel’s “Literature and National Character,” spread outwards from its Western roots with European colonization, reaching Central Asia through Russian expansion. The Romantic essentializing of identity through linguistic standardization and the establishment of a polar “us” versus the Other mentality is apparent in the sphere of Central Asia throughout Tsarist and Soviet rule.

Before Russian expansion into Central Asia, the two reigning linguistic trends were acceptance and assimilation. This region, with its diverse collection of cultures, languages, and religions, was open to the influence of the languages crossing its social space. Bilingualism or multilingualism was the rule and language was seen as a tool for communication; different languages commanded different niches depending on situation and addressee. Russian movement into the steppes and Turkestan brought with it the Western Romantic idea that language is not only functional, but also a marker of shared consciousness and identity (Segars 2003, 92–94). The fixation on language as the defining feature in categorizing groups is revealed by the legal category of inorodtsy in Imperial Russia, which was meant to distinguish aliens from ethnic Russians. The evolution of the meaning of inorodtsy over the years reveals the difficulty in finding a satisfactory marker of that identity. The classification of the peoples of Central Asia caused great confusion because ethnicity in the region historically had not consisted of neat categories dividing each group clearly from others. People identified with multiple groups variably based on language, kinship, religion, and lifestyle; identity was often contradictory, but it was an accepted
norm. The Western and particularly Enlightenment sensibility compelled the classification and ordering of this space into a clearly delineated grid. The term *inorodtsy* shifted from being anchored in race to religion, finding none of these fully adequate, and finally settling on language as the most appropriate fit.

With that essentializing model of political rhetoric hanging in the air from Tsarist legacy, new Soviet rule unhesitatingly exploited this mechanism, establishing concerted language planning efforts. Discussion of the Soviet empire and its role in building the nationalities of Central Asia raised the question of whether this new empire fit a more “classical” model of colonial empire or whether it conceived itself as something else altogether. Scholars have written that it is too simplistic to see the Soviet regime as merely another colonial power which, when carving territory for the national Autonomous Soviet Socialist Republics, was really doing so as a “divide and rule” tactic, though admittedly delimitation did make the region easier to administer.

In reality, Hirsch argues, according to Soviet ideals, “nation” was really just a “transitional stage on the evolutionary timeline,” and they imagined the future of the Soviet Union as one devoid of nations and nationalism, an integrated socialist union (Hirsch 2000). During the Soviet period, the ultimate goal was the formation of a large socialist polity under which each of its citizens first and foremost considered himself to be a Soviet citizen. The interesting contradiction, however, is that toward the process of creating a whole, the parts first had to be separated and “strengthened,” creating a bifurcated identity of *dva potoka*. The five Central Asian Soviet Socialist Republics were carefully delimitated and their constituent nationalities defined and cultivated, with the official label that this was a necessary but passing stage of modernizing the native populations of the region. Arne Haugen (2003) writes that the Soviet regime strove to reproduce the nation as a modern entity—with a centralized bureaucracy, administration, education organization, and so on—in Central Asia. It seems that this school of thought long had precedence with Russian rule in Central Asia, from Catherinian notions of cultivating Islam on the steppes as a
modernizing force among the backward nomads to Ilminsky’s native language policy of first teaching schoolchildren to read with Cyrillic orthographies of their native tongue before advancing to the Russian language in hopes of improving literacy. It is not difficult to believe that nation-building was undertaken to modernize, rather than solely to weaken, Central Asia, but at the same time this goal of modernization is also intrinsically essentialist. The method of carving the region into separate republics was based on “three main principles: national-ethnographic, economic, and finally the principle of administrative order” (Haugen 2003, 181). The latter two principles often overrode the first, because Soviet administration viewed pure ethnic homogeneity as of no inherent value. To establish truly modern nations, other principles were more important—delineating areas with viable economies and ensuring that each republic be equipped with a major urban center to maintain administrative order. Language policies of this period were also motivated by the goal of modernization.

Isabelle Kreindler divides Soviet language planning into three periods: from 1917 to 1930, 1930 to 1958, and 1958 to 1985. The first period, based on Lenin’s attitudes and the slogan “national in form, socialist in content,” advocated the active promotion of native languages. The first stage of the indigenization, or koronizatsia, campaign, adopted in 1923 until 1927, stated that “administration of national republics should be made up of predominantly local people, knowing the language, way of life, morals and customs of the corresponding peoples” and that all official positions must be occupied by native representatives using the native tongue. Orthographies of Central Asia’s spoken languages were Latinized in 1926 in order to eradicate mass illiteracy, which was cast as a relic of backwardness and exploitation by the bourgeois elite. During the second phase of Soviet planning period, however, language policy took less active initiative toward native language promotion, perhaps due to an unexpected backlash and the complications that emerged out of implementation during the first phase. Qualified native officials had been scarce due to low literacy levels among the largely rural populace, and even when they were available, many
were loath to stay in positions that immersed them in a Russian-dominated environment where they often encountered bitter sentiments from their “elder brothers.”

During the second period in 1938, Russian instruction was made mandatory and the Latin writing systems put into usage in 1926 were replaced by the Cyrillic alphabet. Interestingly, Russian was pushed forth as the dominant and common language of the Soviet Union, crucial for the Sovietization process of non-Russian populations, but “on the other hand, the regime was unwilling to retreat on native-language education sufficiently to ensure that young people would learn Russian first and foremost” (Blitstein 2001, 267). Blitstein describes the realities of this program further, saying that “native-language schools in most union republics did see an increase in the time devoted to the Russian language, but the decree actually led to a decrease in the number of hours devoted to Russian in the non-Russian schools of the RSFSR, compared to schedules for years past” (Blitstein 2001, 258).

In 1958, Soviet language policy entered its third period where mandatory native-language education was eliminated and non-Russian parents were provided a choice of sending their children to Russian language schools if they so desired. The true and actuated advance of the Russian language into the space of native tongues was effected during this period through the introduction of a new type of non-Russian school. In these schools, native language and literature remained as curricular subjects, but the general medium of instruction became the Russian language. The methods of this last period persisted until perestroika in 1985.

NATIVE INVOLVEMENT IN DEFINING LANGUAGE AND IDENTITY

In order to put official policies in perspective, it must be noted that the idea that the nationalities of Central Asia was created by an external, essentializing hand, is too simplistic a model. The native subjects themselves were not completely acquiescing puppets controlled by forces over which they had no say. During both the
Tsarist and the Soviet periods, the indigenous populations of Central Asia were often actively involved in the process of defining their group identities and national languages. The strongest example of this during the Tsarist era was the Jadid movement, which called for the modernization of Islam and promoted the creation of a Pan-Turkic language and nation. The Jadid movement was interesting because at the same time as it drew influence from Russia’s embodiment (to the Jadids) of Western progressive ideologies, it sought to distinguish itself as a people separate from the Russians. This led the Jadids to criticize the traditional Muslim education of the madrasa system as backwards and to emphasize Russified or Westernized forms of modernity and education, which entailed reformed Islam, an understanding of history, and improved literacy. In Jadid literature, such as Tales of an Indian Traveler by Bukharan Jadid Abdurrauf Fitrat, Russian personas often represented education and intellectuality, “models to be emulated.” Adeeb Khalid writes that “the Jadids had internalized several categories of colonial knowledge, yet their very assertion of the universality of progress subverted these categories” (Khalid 1997, 191–200). This same inspiration gotten from the Russian “elder brother,” inspired the Jadids to form a more definitive and distinct pan-Turkic identity. In hopes of unifying the Turkic peoples of Central Asia, Ismail Bey Gasprinsky, ideological predecessor of the Jadids, pursued several principles of unity, one of them being the unity of language. He proposed the usage of Ottoman Turkish as the common Turkic literary language, but ultimately this failed to flower due to nontrivial differences between the Turkic tongues spoken across such a wide area as from Crimea to the Volga to the Kazakh steppe and deep into Turkestan.

While Soviet policies in later years sought to mold its territories and the people within them, creating an environment for nationhood by giving ethnic identity real political and economic meaning, titular nationalities under the Soviet Union were not merely passive sheep in a pen. Adrienne Edgar gives examples of this in her analysis of nation-building in Soviet Turkmenistan. The Turkmen played important roles in shaping the discourse on na-
tionality, rejecting the idea of a pan-Turkic language “as linguistic imperialism,” instead “preferring to emphasize the distinctiveness of the Turkmen vernacular rather than its commonalities with other Turkic dialects” (Edgar 2004, 9). But although Turkmen intellectuals had adopted Soviet ideas on language as an essential component to shaping nationality, they were unable to divorce their conceptions of language from their traditional Turkmen genealogical and tribal basis of identity formation. Ultimately, through an essentializing process, these internal differences were covered up and the Teke dialect, which was the dialect of the largest tribe and constituted the majority of Turkmen intelligentsia emerged as dominant, and as representing the whole of a united Turkmen identity. This procedure diluted linguistic differences and succeeded because a greater goal was at hand, which was that this unity was necessary to distinguish the Turkmen as a legitimate nationality in contrast to other groups in the region. As the new Turkmen elites were drawn together from across the republic to Ashgabat, “the subdivisions within the Turkmen population came to seem less important than the huge cultural and linguistic gap between Turkmen and Europeans” (Edgar 2004, 94). This example of native involvement in identity building in Turkmenistan is merely one specific example of native groups participating in the process of nationality creation across the Central Asian space.

For example, during the period of geographic and national delineation of the Ferghana Valley, the Kyrgyz and Uzbeks adopted Soviet ideas and phraseology for ethnic categorization and administrative functionality to articulate their desires for territory. They used the standard of “psychological make-up” from Stalin’s ideology of a nation to argue for their own ideas of national boundaries. Many groups of people wrote letters to the government asking to be incorporated into a certain republic claiming that nationality should not be seen along along Tsarist- and Soviet-influenced divisions of socio-economic lifestyles, but should rather be conceived along how individuals personally believe them to be.

The people of these regions, using whatever standards and measures they have adopted, self-identify with one group or an-
other, and thus their “psychological make-up,” according to Stalinist rhetoric, determines which nation they feel they best fit into. At the same time, it is a stretch to argue that these native groups are purposefully using official frameworks in which they themselves have no faith, to subvert established order. It is apparent that they also followed these categories and definitions—as evidenced by later independent nations corresponding to earlier administrative units, and the lasting presence of the five main nationalities that emerged during the Soviet era in post-Soviet political space. Turning back to the specific case in Kyrgyzstan, we can examine how nation-building was carried over from the Soviet period into independence.

NATION-BUILDING AND LANGUAGE PLANNING IN POST-SOVIET KYRGYZSTAN

When the Soviet Union collapsed in 1991, the independent republics of Central Asia were faced with the task of maintaining national rule and unity on their own. Though independence had come undesired to Kyrgyzstan, the emerging nation was now nevertheless confronted with the need to exist as a truly independent and coherent nation-state. Until the Soviet period implemented its policies to solidify national identities, Kyrgyz society was based on kinship and lineage ties rather than national sentiments.

Soviet scholar S. M. Abramzon admits that “the question of the origins of the Kyrgyz nation is among the most complex and controversial aspects of the ethnic history of Central Asia” (Huskey 1993, 412). Adopting Soviet-styled policies and essentialist rhetoric, the independent government former Kyrgyz president Akayev created and promulgated a national ideology that sought to legitimize the nation’s existence. The Kyrgyz SSR had been one of the least “nationally conscious” republics in the Soviet Union and the challenge post-independence was that of filling the “ideological vacuum left by the discrediting of the communist system” (Lowe 2003, 114). Although a significant number of diverse ethnic minorities lived in Kyrgyzstan, the path toward building a strong Kyrgyz nation was one which muted diversity and elevated symbols of
the dominant culture and its traditions.

Although in 1989 there was an ethnically heterogenous population in Kyrgyzstan, 1999, the census reported the number of individuals identifying themselves as Kyrgyz as increased by twelve percent. The mass migration of Russians and other European groups, most of whom had higher levels of education and occupied positions in the academic, governmental, or white-collar sectors, was termed the “brain drain.” While a recent survey showed that only 12 percent of Russians had emigrated due to “linguistic reasons” (most cited “greater economic opportunities”), the government of Kyrgyzstan attempted to promote the nation as a “multi-ethnic, civic state” (Korth 2005, 119). The government sought to reassure minority groups of their welcomed status in the new nation—as exemplified by Akaev’s slogan “Civil Consensus and Nation Unity—Yes; Chauvinism, Nationalism and Extremism—No.” Nevertheless, the major steps taken by official policy were those which sought to foster a shared sense of Kyrgyz national pride and to ground Kyrgyz national identity in “authentic” origins, as defined by standards of the same thread of Western Romantic thought that was imported to Central Asia through Russian and Soviet rule. The contradiction in essentialist policies becomes clear here as “national consolidation is essential to avoid the kind of fragmentation and political schism arising from such division that was witnessed in Tajikistan. Yet nation-builders in their homogenizing zeal fail to harness the cultural and other diversities within the nation” (Patnaik 2003, 144).

The process of strengthening the sense of nation included an emphasis of ethnosymbols, such as the symbolism of flag design. Like the banner flown by the legendary Kyrgyz warrior Manas, the bold red flag is emblazoned with a forty-rayed sun representing the forty historical Kyrgyz tribes, and also with a tunduk, the roof of a yurt, symbolizing the Kyrgyz’s nomadic heritage. Ethnosymbols also sought to link the people and nation with the specific plot of land they now occupied. The state seal of Kyrgyzstan, for example, depicts the sun rising above the striking peaks of the Ala-Too mountains on the shores of Lake Issyk-Kul. History is also
used in the legitimizing process as authorities search for “original authenticity” in order for past unity to justify the existence of a collective modern nation. Having been a non-literate culture for centuries, the oral Manas epic has seen intense and mass revival as both mythology and ancient history. The epic is not only recited, it is also reenacted at elaborate festivals on national holidays, and has inspired books, films, TV shows, operas, even comics (Lowe 2003, 117). Islam, embodied as national religious custom, is another marker of national identity, though on the official level, this appears in a glossed, superficial form. Each of these essentializing strategies for identity building could be analyzed in much greater depth, but in this paper, I will select to look at linguistic strategies in more detail.

During the last years of Soviet rule, “it was clear that Kyrgyz was one of the weakest titular languages in the Soviet Union” and, in fact, Russian had taken the position as “the first language of the urban Kyrgyz elite” (Lowe 2003, 118). Post-independence, therefore, people of the former Kyrgyz SSR found themselves liberated from policies of former “colonizers,” but habits of that earlier era still remained. And so the Russian language still found itself being used in education, administration, and even for private matters, mostly among the urban and educated population. Efforts were made to increase the usage and prestige of Kyrgyz in society, yet even these attempts contained within them the remnants of a colonial mentality and policy.

Just as the Tsarist and Soviet administrations had held on to a tenuous link between nation and language (perhaps more clearly seen during the last period of Soviet language policy, when Russian became inarguably the dominant language throughout the multiethnic Soviet Union), “many Central Asian governments have turned to the model of the modern nation-state in which the titular nation defines the state language, although this model of a mono-ethnic and monolingual state does not correspond to their social reality” (Korth 2005, 116). Even before the complete collapse of the Soviet Union, the Kyrgyz parliament passed the Law on State Language, for perhaps more symbolic reasons of asserting national
pride than anything else, which made Kyrgyz the state language and Russian the “language of inter-ethnic communication” (Lowe 2003, 118). This corresponds to the current status of official languages in the independent Kyrgyz Republic, which maintains Kyrgyz as the “state language” and Russian as the “official language.” Most schools are double-tracked, meaning that parents can choose to send their children to school on either a Russian track, where the language of instruction is Russian, or a Kyrgyz track, where the language of instruction is Kyrgyz. At least in urban areas, students on the Kyrgyz track still take a Russian language subject class, and vice versa. This is reminiscent of the final period of language planning during the Soviet period, when non-Russian families could choose to send their children to an all-Russian instruction school. Nevertheless, Kyrgyz is gaining ground with purposeful planning efforts—place names have been stripped of their Soviet attributions and given Kyrgyz ones instead, and the National Commission on State Language has announced requirements that mandate the use of Kyrgyz in all official documentation. Most scholars, including Lowe and Korth, however, predict that despite the public’s general acceptance of official rhetoric enhancing the prestige of Kyrgyz, Russian will still be planted firmly as the lingua franca among the majority of people due to high ethnic diversity which most Kyrgyz-language campaigns do not specifically accommodate.

In response, other minorities employ what could be termed “strategic essentialism” in order to emphasize their voice as cohesive and legitimate within Kyrgyzstan, where they see the rise of Kyrgyz nationalism as a threat. Strategic essentialism may be used in the formation of any social identity where complexities are oversimplified for a crafted subversive purpose. For example, faced with official characterizations of Kyrgyz national identity, Uzbeks living in Osh assert their own collective identity as a firm and fixed unified whole. In my conversations with Uzbek individuals living in Osh, they asserted that historically, the Ferghana Valley has been the dominant home of the settled Uzbek population, and that the original home of the nomadic Kyrgyz was further north in the mountains around the provinces of Chuy and Issyk-Kul. This ig-
nores the complex history of settlement in the Ferghana Valley as well as the initially vague separation of Ferghana’s population into distinct and separate ethnic groups, but this is a reassertion of the legitimate belonging of the Uzbek population in a place they view as their historic homeland. They connect the language they speak to the Uzbek language spoken by over 30 million people worldwide—many more than those who speak Kyrgyz—even though the Tashkent dialect of Uzbek is different from the dialect spoken in the Ferghana Valley. In a region of the world where, only a few centuries earlier, languages were seen more as distinct but not wholly unconnected dialectical shades, this characterization nevertheless imbues the Uzbek language with a sense of power and authority in a Kyrgyz nation where it is neglected on the official level.

**BREAKING DOWN ESSENTIALIZED IDENTITY**

Ultimately, however, essentialist models are analyzed in order to be broken down. Bucholtz elaborates, “[I]dentity inheres in actions, not in people. As the product of situated social action, identities may shift and recombine to meet new circumstances” (Bucholtz and Hall, 2003, 376). And so those same Uzbeks described above, who in one situation emphasize the important distinctions between themselves and the Kyrgyz, may in another prioritize the distinction between their urban Russified status in contrast to rural and traditional populations. They may, while in conversation with a Bishkek native, characterize themselves as “Osh-ski,” a person of Osh; or in dealings with an Uzbek from Tashkent, actually prefer to name themselves as citizens of Kyrgyzstan. Paitnak writes:

Post-structuralists suggest that ethnic identity is a shifting, contested and negotiated category. Since it is always in the process of being constructed and reconstructed, the collapse of the Soviet Union put strains on those structures that defined ethnic identity in Central Asia, notwithstanding the attempts to create a ‘simple, given, bounded’ national entity. The nationalist project of fur-
ther fortifying those structures have complicated inter and intra-ethnic relations (Paitnak 2003, 154).

Montgomery shows that the diversity of religious beliefs and practices including common, cultural “religion on the streets” in Kyrgyzstan and, by relation, in the rest of Central Asia is part of a process of socialization (Montgomery 2007, 366–367). The shifting identities in community customs of Islam is, from my experience, a good analogy for the performance of language as a part of this process of socialization as well. Turning back to the example of Suleiman-Too set up in the introduction, upon closer examination, it is clear that in addition to the sacred usage of Suleiman-Too Mountain as a place of religious worship, the mountain is a setting for the constantly changing values of space and community, and has come to symbolize meaning outside of the religious as well. Suleiman-Too creates a social space for the diverse population of Osh and its surrounding regions; older religious traditions and even those cultural meanings have been overlaid with a newer, socially significant purpose. Individuals, friends, families, couples, regularly climb the mountain not necessarily with religious pilgrimage in mind (though some certainly do), and for young lovers, it can serve as a venue for more amorous activities. The cemetery at the base of the mountain next to Rabat Abdullah Khan Mosque also becomes a meeting spot for Osh’s drug users. The place is seen by them as a sort of safe haven for their activities. Thus, the “multiple meanings” of this sacred space are not stagnant, but continue to accumulate through continued usage and continued acceptance of its religious, cultural, and social importance in the community over time.

Language usage and linguistic meanings are, similarly, neither stagnant nor static. Meaning in language accumulates and changes as part of the process of socialization. In the analysis of language in this final section, we shift away from essentializations of linguistic identity to the elusive moments, normalized contradictions, and continuous shifts of daily communication and linguistic attitudes in Central Asia—specifically turning back to the scope of youth in Osh, Kyrgyzstan. In order to deconstruct essentialist perspectives,
it is sometimes necessary to step away from established academic analyses and methods of cohesive interpretations, and to do so I will be incorporating my sociolinguistic field research and personal encounters with everyday life in Osh.

The general linguistic divide in Osh could be simplified into Russian versus Kyrgyz versus Uzbek, but this picture is complicated by the realization that there is often little correspondence between nationality and national language. And so both urban Kyrgyz and Uzbeks will distinguish themselves from rural populations using the Russian language—urban-educated groups speak Russian, while the rural—even those who have in recent years moved into the city—do not. A similar distinction is made in contrast to those who lead a traditional lifestyle by those Kyrgyz and Uzbeks who see themselves as modern and more “Western”—they read Russian-language literature and publications, watch the Russian “First Channel” on television—because they or their parents were educated in Russia during the Soviet period.

To further complicate matters, we can once again bring into conversation the young Uzbek-Kyrgyz man who identified as Russian because of a personal connection with the beauty of the Russian language or engage in dialogue several Russian teenagers I interviewed whose families still remained in Osh after the massive deflux of the Russian population and who told me that they fluently speak both Uzbek and Kyrgyz with their friends, having learned these languages not in schools but as children playing outside together in the dvor. We also have the patriotic Kyrgyz youth who insist that all must speak Kyrgyz and who, even when they speak to Uzbeks, always address them in either Russian or Kyrgyz, resulting in whole conversations carried out with each individual speaking his own tongue, despite the fact that they may have basic fluency in the other’s language. There are also Uzbeks who say that maintaining the Kyrgyz language is important since, after all, they are citizens of the Kyrgyz Republic, and there are Kyrgyz who likewise believe in the importance of knowing enough Uzbek to communicate with a large sector of the population of Osh and its surroundings. Uzbek singers and Uzbek language television serials
are popular among Uzbeks and Kyrgyz alike, and a young man visiting family in Osh from Tashkent told me that these Uzbek stars are in fact more popular in Osh than they are in Tashkent, where most prefer Western pop culture.

In full recognition that Central Asia does not exist in isolation and that linguistic attitudes and identities also change in respect to new external influences, the demand for and popular status of the English language must be acknowledged. Most schools incorporate English as a subject from the fourth grade until graduation from secondary education, and many of the young people interviewed stressed the importance of English as the current language to learn and speak. Not only are schools doubled-tracked in Russian and Kyrgyz, as described earlier, but new Turkish lyceums are opening up all around the city, where classes are attractively conducted with English as the language of instruction, and Turkish—Central Asia’s perceived connection to the rest of Europe nowadays—being a language subject.

CONCLUSION: CODE-SWITCHING & THE FUTURE OF KYRGYZ IDENTITY

Expressed attitudes aside, perhaps most interesting and telling are the actual behaviors and habits of language choice and usage. The phenomenon of particular note here is the manner of habitual code-switching between Kyrgyz and Russian, Uzbek and Russian, or all three that has become the norm for communication at home and, more often, with friends, especially among the youth of Osh. Code-switching is a common linguistic phenomenon in situations of extended language contact. Formal analyses of code-switching often define it through systematic classification of its structures; for example, differentiating it from other phenomena of linguistic contact as intrasentential embedding of phrases and words of one language into the syntactic structure of another “matrix” language. Here, I am less concerned with the strict forms of code-switching, and more interested in the social performance and usage of “switching” as a broader category. Contemporary studies of the so-
cial importance of code-switching have been largely preoccupied with power and prestige, analyzing specific motions of switching to convey authority or deference. Certainly these instances of code-switching between Russian, Kyrgyz, and/or Uzbek can also be examined in this fashion and certainly individuals do consciously switch in certain situations to strategically communicate something about social relationships. But what is particularly telling about the multiplicities of identity in the situation I encountered in Osh, and which is applicable across most parts of Central Asia, is the way “code-switching” has become the norm in communicating, particularly among the younger generations. Many young Uzbek and Kyrgyz describe it as being the most “natural” and “comfortable” way of communicating for them—easier and more familiar than speaking in pure Russian, Kyrgyz, or Uzbek. Carol Meyer-Scotton writes that “codeswitching can index a speaker’s self-perception, as a multidimensional person, whether as a member of a specific group, or as a member simultaneously of several groups. As a tool, codeswitching can be used in an ongoing conversation to step in—or out—of a presumed or expected identity” (Myers-Scotton 2006, 73). Habitual code-switching is code-switching that has become so normalized that jumps from one identity to another are no longer taken purposefully or even consciously. It reflects the simultaneous layering of multiple identities, and often suggests that these identities exist on the same level and with the same degree of comfort and sense of belonging. It socializes the youth of Osh into a diverse community that does not define itself statically as purely Kyrgyz, Uzbek, or Russian in everyday life, but where identity is constantly being negotiated.

Thus, linguistic identity and language in everyday usage reflects a rich mix of all the historical legacies, religious influences, cultural contacts, and ethnic interplays that have touched the region, from pre-Tsarist times through the Tsarist and Soviet periods. It shows that despite contemporary official policies to assert the dominance of one consolidated nationality over the others, identity in Central Asia will most likely remain fluid, multifaceted, and versatile.
Works Cited


THE BOUNDARIES OF DIGITAL ACTIVISM: DISSIDENT USE OF SOCIAL MEDIA IN IRAN

Nishi Kumar

This paper examines the various ways Internet social media networks and new media were used in the June 2009 Iranian demonstrations following the contentious presidential election, the effectiveness of these media in mobilizing and informing domestic and international audiences, and the implications for freedom of speech, democratic norms, and human rights in Iran. While social media played an important role in providing news and images to the global audience, its domestic use was limited due to widespread distrust and government censorship. Although new technology and digital activism enabled a flow of information that would not traditionally exist in a closed society, the contributions to Iranian democracy and human rights were negligible and possibly even harmful.

WHAT IS SOCIAL MEDIA?

Social media is distinct from traditional media such as newspapers, broadcasting, cable television, radio, and film. Whereas traditional mass media can be characterized as “one to many,” social media transforms the conversation into a “many to many” dialogue by facilitating social interaction via relatively new technologies, such as the Internet. Social media’s accessible and interactive publishing techniques mean that content is easily produced and information is widely available. Through social media networks and new media technology, individuals, formerly content consumers, become content producers (Benkler 2006). Social media builds on the capacities of Web 2.0 tools that facilitate user-centered design and

Nishi Kumar is a senior at Columbia University majoring in Political Science-Economics. Upon graduation, she plans to teach high-school mathematics in New Orleans with Teach for America.
content production through web-based communities, forums, and applications, video and image sharing cites, wikis, and blogs (Clark 2009).

Building on the capacities of Web 2.0 tools, social media is distinct from traditional media in five important ways: it is global, instantaneous, highly accessible, interactive, and alterable (Benkler 2006). Social media has a global reach and can relay news and information almost instantaneously. Unlike mass media, social media is relatively inexpensive, requires little capital investment or resources to publish information (e.g., a printing press), and is accessible to an increasingly large number of the global population. Social media facilitates interaction between users and readers by enabling discussion, commenting, and editing. It is more participator than, and lacks the permanence of, older media.

“Citizen media” is a term closely tied to social media and new media technologies. It is produced by private individuals who have no official political, social, or corporate affiliation and emphasizes public issues that are often absent from mainstream mass media, and often has a partisan bent. Citizen media empowers individuals to produce their own politically- and socially-motivated content. Andrew Shapiro argues that the emergence of new digital technologies signals “a potentially radical shift of who is in control of information, experience, and resources” (Croteau, 2003, 322)

New media have significant connections to and implications for globalization: they “radically break the connection between physical place and social place, making physical location much less significant for our social relationships” (Croteau 2003, 311). Global movements using these new participative technologies allow diverse networks of participants to collaborate, problem-solve, and discuss regardless of their geographic location.

FREEDOM OF SPEECH AND EXPRESSION IN IRAN

Article Twenty-Four of the Iranian constitution states that while “publications and the press have freedom of expression,” it is unlawful to express views that are “detrimental to the fundamental
principles of Islam or the rights of the public.” The Iranian Press Law of 1986 further clarifies unlawful media activity, which includes “promoting subjects which might damage the foundation of the Islamic Republic” and “offending the Leader of the Revolution and recognized religious authorities” (Open Net Initiative 2009). Violators face criminal prosecution, months of incarceration, torture, and even the death penalty. Ethnic-minority journalists are specifically targeted for harsh punishment: in 2007, two Iranian Kurdish journalists were given the death sentence after being declared “enemies of God” and suspected of endangering national security (Freedom House 2009). Journalists and bloggers who support issues deemed “controversial” by the regime were also targeted, as were individuals with actual or suspected connections abroad.

Open Net Initiative, an academic project that studies Internet filtering, finds the Iranian press legislation to be unusual “in that it not only describes restricted speech but also lays out normative objectives for the press, who are required to ‘propagate and promote genuine Islamic culture and sound ethical principles.’” In May 2009 Freedom House ranked Iran 181 out of 195 countries for media openness—tied with China and Rwanda, and below Syria, Sudan, and Somalia. The strict laws regulating the freedom to express political and moral ideas have led to complete state control over domestic newspapers, press, radio, and television. The Ministry of Culture must approve the domestic publication of all books, and inspects foreign books before their distribution. The Press Court has discretion to prosecute journalists, editors, and publishers based on its interpretation of constitutional stipulations such as “insulting Islam” and “damaging the foundations of the Islamic Republic.” In just the first few months of 2008, the Court banned seventeen newspapers (Freedom House 2009).

The government also frequently issues gag orders that limit coverage of specific topics and events. Before the 2009 presidential election, foreign correspondents were expelled from Iran and domestic journalists were strictly censored and even jailed. One month after the elections and protests, nearly forty journalists still remained in Iranian prisons (Reporters Without Borders 2009).
The Ahmadinejad government has made it clear that the media’s duty is to support government policy and not to provide commentary or hold officials accountable for their actions. Pervasive fear among journalists has created an environment of self-censorship, inhibiting public criticism of the president and regime (Milani 2009). In 2007, the Iranian Journalists’ Association (IRNA) reported that the quality of domestic journalism suffered due to the government’s crackdown on independent newspapers. The government has systematically and persistently harassed the IRNA, pressuring it to include pro-government journalists (Freedom House 2009).

The number of Internet users in Iran increased from 7.5 million users in 2005 to twenty-three million in 2008—almost twenty-five percent of the population (Open Net Initiative 2009). Originally, the government censored Internet content by forcing service providers to block access to a growing number of sites deemed immoral or politically controversial. In 2006, the authorities created a central filtering system to block these websites, identify domestic Internet users, and keep a record of visited sites. As of January 2007, Iranian bloggers were also required to register with the government—a formidable feat given the estimated 60,000 active writers in the Persian blogosphere (Open Net Initiative 2009). In July 2008, a bill legalizing the death penalty for “the creation of webslogs and websites promoting corruption, prostitution and apostasy” was passed by the parliament. Under this law, Iranian-Canadian blogger Hussein Derakhshan was arrested in Tehran during a visit in November 2008 for “insulting religious figures” (Freedom House 2009).

The government has implemented its filtering system by routing all public Internet traffic through proxy servers. This allows filtering software to target specific webpages and to block keywords. The government openly admits that it blocks these websites in order to demonstrate its power over information. Users that attempt to access a blocked site reach a page warning them that they are not permitted access (Open Net Initiative 2009). In 2008 and 2009, ONI tested five Internet service providers in Iran: ITC, Gos-
The results confirmed that the extent of Iran’s Internet filtering is second only to China. Iran filters a broad range of websites deemed offensive to the moral standards of Iran; for instance, those in support of women’s rights groups, minority rights groups, government critics, human rights organizations, and political opposition and reformist parties.

The government has also limited Internet speeds for public organizations and private households. Before this policy was implemented, fiber-optic networks had been expanding rapidly in Iran, more than doubling from 2005 to 2007. Mohammad Soleimani, the Minister of Information and Communications, publicly defended the limited access speeds, stating that slower speeds were adequate for the country’s needs and that there was no demand for higher speeds (Tait 2006). Iran remains the only country in the world to implement a limit on access speeds for private households.

The government often threatens legal and violent action against Internet dissenters, and increasingly it follows through on these threats. Although the Iranian constitution prohibits arbitrary arrest and detention, they are relatively routine practices. Government officials use arbitrary detention and intimidation to encourage self-censorship by professional and citizen journalists. Use of torture against dissidents is suspected in both the unofficial detention centers and the notorious Evin prison (Penketh 2009). A 2004 law outlawed the use of torture in interrogations, but reports from 2008 have shown the practice has persisted. Political prisoners are subjected to inhumane conditions and brutal treatment, and those who advocate for their rights are also subject to persecution (Freedom House 2009).

At the same time as it limits public access, the government has also started to use the Internet to promote its own ideology and extend its authority. The Revolutionary Guard recently declared a campaign to launch ten thousand blogs written by members of the Basij, a volunteer Iranian paramilitary force (Bruno 2009). The government has also actively spread pro-regime propaganda and misinformation about dissidents over the Internet. This suggests
a new policy directed at shaping an Internet presence favorable to the regime: an active, government-backed Internet campaign specifically targeting dissidents, opponents, and critics.

JUNE 2009 PRESIDENTIAL ELECTIONS

An unprecedented number of voters turned out for the June 2009 Presidential elections. The results were rejected by all three opposition candidates. Following the June 13 declaration of incumbent Mahmoud Ahmadinejad as the winner with 63 percent of the vote, widespread protests took place across Iran, particularly in Tehran, and in major cities around the world (Worth 2009). The police reacted violently to the protesters, many of whom were students, women, young activists, and supporters of the opposition candidate Mir Hossein Mousavi. Whether or not the election results were in fact fraudulent, the demonstrators’ use of new media reveals valuable information about how information-sharing technologies, social networking, and new media can support mass mobilization and political demonstrations in the face of state censorship and suppression.

Ayatollah Ali Khamenei declared the historic voter turnout a “divine assessment” of Ahmadinejad and urged the nation to unite in support of the results, but was later forced to investigate the claims of fraud in the face of domestic and international pressure. Ahmadinejad called the election “completely free” and the outcome “a great victory” for Iran (BBC 2009).

The police and the paramilitary Basij suppressed both peaceful demonstrations and the increasingly violent riots, using a variety of weapons and firearms. The Iranian government has confirmed the deaths of thirty-six people during the protests, while supporters of opposition candidate Mousavi allege that there were seventy-two deaths in the three months following the disputed election. Iranian authorities further suppressed the movements by closing universities in Tehran, blocking web sites, intercepting cell phone transmissions and text messaging, and banning political rallies. As of August 2009, the Iranian government confirmed that over 4,000 people were
temporarily detained during the protests (AFP 2009).

According to the Telegraph on June 14, “Iran’s regime was doing its utmost to choke off the flow of news” to other areas of the country and the world (Blair 2009). Al Jazeera English accused the government of direct censorship, stating that “some of the newspapers have been given notices to change their editorials or their main headlines” (Blair 2009). The al-Arabiya offices in Tehran were closed on June 14 for a week, and NBC News offices in Tehran were raided and equipment confiscated. Peter Horrocks, director of BBC Worldwide, said that audiences across the Middle East and Europe were affected by an electronic block on the satellites used to broadcast the BBC Persian Television signal, which, Horrocks said, “seems to be part of a pattern of behavior by the Iranian authorities to limit the reporting of the aftermath of the disputed election” (Horrocks 2009). Further corroborating this statement, the Iranian Ministry of Culture banned all foreign media journalists from leaving their offices on June 16. While international news agencies could still report on the elections and protests, they were not allowed to directly witness the demonstrations. On June 20, the Ministry of Culture further banned the international media from issuing reports on what was occurring without explicit permission from the Iranian government (Reuters 2009).

Nearly 35 percent of the Iranian population are active web-users, the Internet and new media played a role in both the election process and the resulting demonstrations, acting as primary source of information and news for domestic and international audiences (Open Net Initiative 2009). Websites such as Facebook, Twitter, Flickr, and YouTube helped the opposition movement share the government’s brutal response with a global audience.

As the Iranian government limited traditional news networks’ ability to report from the field, citizen journalists on the streets used new technology, such as cell phones, to share news, images, and videos of the protests via the Internet. Iranian officials have attempted to deflect blame onto these new media, and accuse the West and foreign media of backing the political opposition in order to undermine the Islamic government.
CAUSAL MECHANISM OR PROTEST TOOL?

Months after the election, the causal role of social media in organizing protests is still disputed. The “Twitter revolution” argument holds that social media was remarkable and innovative, and responsible for bringing protesters together (Fisk 2009). Another point of view argues that the use of the Internet as rather a tool of citizen journalists to share news and images—that would otherwise be suppressed both domestically and internationally. A third interpretation is that the use of social media to increase free speech actually hindered democratic movement and increased government repression and human rights violations.

In order to evaluate the role of social media in response to the Iranian elections, the mechanisms through which it contributed to the protests need to be understood and its effectiveness evaluated. Only then will the implications for freedom of speech, democratic norms, and respect for human rights become apparent.

Communication and Mobilization of Protestors

Perhaps nothing is more illustrative of Twitter’s perceived value as a communicative tool between Iranians as when the U.S. State Department asked Twitter to postpone scheduled maintenance so Iranians would have access to the service during the street protests (Labott 2009). Texts and images calling for “death to the dictator” were widely distributed through email and posted on YouTube, and Facebook, and Twitter, creating the illusion of a collective resistance of media-savvy young people at the forefront of a political uprising similar to 1979’s. The foreign press promoted this image of a “Twitter revolution” to the international audience (Sullivan 2009). In reality, however, extensive government censorship, infiltration, and blocking limited the role social media played in mobilizing protestors within Iran (Schectman 2009). While protest organizers were sometimes able to temporarily bypass government censors, the larger population, wary of the new technology and of potential government infiltration, often distrusted mobilization at-
Sysomos, a blog, analyzed the use of Twitter in Iran during this period. After the election, there were 19,235 Twitter users in Iran, compared to the 8,654 in mid-May 2009. Sysomos determined the number of users by re-indexing over thirteen million Twitter accounts. They used information provided in user profiles as disclosed in May (for users who joined before June) to determine the location of users, in order to avoid counting those who changed it later to Tehran to protect and support protestors. Sysomos also looked at when Twitter accounts were created over the past 15 months. The number of Twitter users in Iran grew in 2009, with March and June the most active months, when 9.81 percent and 9.93 percent of all Twitter accounts were created, respectively (Sysomos 2009). However, these numbers only show that more people in Iran were using Twitter during this period, and not that they were using it specifically for mobilization or communication connected to the protests.

In fact, Iran experts and social networking activists say that while Iranian protesters have used social media tools, no one technology has been instrumental to organizers’ ability to mobilize. Most of the protest organization was done through SMS text messages and word of mouth. “Social media is not at all a prime mover of what is happening on the ground,” said Ethan Zuckerman, a senior researcher at Harvard University’s Berkman Center for Internet & Society. “Social media are helpful in exposing what’s happening to the outside world, but it’s a mistake to think that these protests [in Iran] are because of social media. It’s more conventional things like word-of-mouth and phone calls that really bring massive numbers of people into the streets” (Schectman 2009).

Additionally, one of the distinctive characteristics of social media, the relative anonymity of tweeters and bloggers, was actually detrimental to protesters’ communication and mobilization. For example, a Twitter user claiming to be an Iranian student posted misinformation—including a report that demonstrations had been canceled—and listed a phone number of a supposed safe house for protestors that was a trap. As rumors spread that government
officials and secret police were using false Twitter feeds to spread information and trap insurgents, many citizens began to distrust Internet messages and to rely instead on more reliable sources (Kucera 2009).

Increased Iranian censorship and website shutdowns during this period reinforced the public perception that the Internet was an unreliable source of information and communication (Open Net Initiative 2009). In this context, only a limited amount of communication and mobilization among protestors actually occurred via social media. As their policies regulating speech and press have shown, the Iranian regime refuses to tolerate open communication channels within its borders and has been successful in limiting the flow of information through a centralized system. This is supposedly intended to prevent “Western media infiltration,” but is also designed to minimize communication between its own citizens (Djavadi 2009).

Information-Sharing

With Iran’s strict control over domestic broadcasting and the ban on all foreign journalists after the election, social media functioned to facilitate citizen journalism within Iran and throughout the world via stories, photos, and videos.

Domestically, this mechanism was muted because of the widespread distrust of Internet sources and the strict government censorship and filtering software discussed above. Information online was intensely partisan and sometimes factually untrue, leaving Iranians to rely more on state-controlled media or word-of-mouth for details about the protests. For example, two false rumors that three million protestors were on the streets and that Mousavi was under house arrest spread quickly through Twitter, Facebook, and blogs without undergoing the fact- and source-checking that would have been by professional journalists (Kucera 2009).

Social media played its most prominent role in relaying information, though not always accurately, to a global audience. The Iranian government worked “on many fronts to shield the outside
world’s view of the unrest, banning coverage of the demonstrations, arresting journalists, threatening bloggers, and trying to block Web sites like Facebook and Twitter.” However, social media allowed news, images, and video to spread throughout the world (Heacock 2009).

As the Iranian government restricted professional journalists’ access to events, the protesters were able to use Twitter’s innovative communication system to direct the public and journalists to video, photographs, and written material related to the protests, even when the Twitter site itself was blocked domestically (Cohen 2009). User-generated content and reporting from Iranians on the ground allowed foreign news networks and bloggers to share information worldwide, although in many instances confirmation of the facts (when such was possible) was delayed. “We’ve been struck by the amount of video and eyewitness testimony…[Sometimes] up to five videos a minute from amateur sources,” confirmed Jon Williams, BBC world news editor. “The days when regimes can control the flow of information are over” (BBC 2009).

U.S. officials said the Internet, specifically social networking sites like Twitter and Facebook, provided the United States with critical information in the face of an Iranian ban on Western journalists covering political rallies. Because the United States has no official relations with Iran and does not have an embassy in the country, it relied on information from media reports and the State Department’s Iran Watch Offices located in embassies around the world. Although U.S. officials would not say whether they were communicating with Iranians directly, one noted that the United States learned that individuals had been identified and detained based on Twitter posts. “It is a very good example of where technology is helping,” the official said (Labott 2009).

The situation in Iran is an example of the State Department’s efforts to increase its use of technology in diplomacy, including social networking sites and Web video, to reach large numbers of people who would otherwise be difficult to reach (Pleming 2009). The Internet and the new media outlets gave international audiences an unprecedented look into the workings of both state violence and
massive unrest in Iran that could not be limited by traditional state forces of suppression.

Public Forum

Social media also contributed to the protests by creating an online public forum through which protestors, citizens, and observers could discuss, share, collaborate, and air grievances and ideas that the government would traditionally repress. Iranians were also able to instantaneously and continuously communicate with Iranians living outside Iran as well as with foreign observers and supporters—social media facilitated relationships between Westerners and Iranians that are blocked on the formal international level. A physical manifestation of this was the spread of the election protests to cities around the world through the use of similar slogans and symbols (e.g., “Sea of Green” in reference to Mousavi’s supporters), and the letters of support issued by Iranians living in the West (Dabashi 2009).

A study by Mike Edwards, a social network researcher at The New School for Design, examined 79,000 tweets related to the Iran protests and found that one-third were repostings of other tweets (one-twentieth of all posts on Twitter, on average, are reposts) (Scola 2009). This suggests that many Twitter users in Iran did not use the network to communicate new information, but to express symbolic solidarity and unity in the form of “retweets.”

Individual photos and images, and stories of fraud, repression, and brutality, helped harness the emotional aspect of the protest and allowed citizens a public space to discuss topics and alternative political ideas prohibited under Iranian law. State power no longer has a stranglehold on information and association, at least not in the way it did before the emergence of new media and its creation of cyber-areas for political discussion (Djavadi 2009). These new technologies were used to bypass government censorship and geographical distance, and they created forums in which citizens and international observers could interact by sharing images, news, and ideas. Global opposition and transactional advocacy networks
emerged and created wider spheres of exchange, dialogue, and resistance. For example, protesters around the world set up proxy servers, making their own computers available to Iranians (Gross, 2009). Mobile phone footage of the shooting of Neda Soltani, a young Iranian woman killed during a protest, was posted on YouTube and other websites almost instantaneously and became a defining image of the Iranian crisis.

The Iranian uprising utilized the power of violent images disseminated throughout cyberspace. The distribution of images was a key strategy that attracted worldwide media coverage. The protests’ success in that area showed that new media could create alternative public spheres in the face of government oppression. Within these virtual public spaces political activists acted as a unified force against the regime and promoted positive intervention and change (Lucas 2009). Unlike the more temporary mobilization and information-sharing uses of other social media, these public spaces and the strengthening of international networks are lasting results of the protests that have the ability to continue and grow with future instances of collective discontent.

Government Backlash

ONI confirmed the expansion of targeted website blocking during the protests, which made access to reporting of events as well as political organizing far more difficult for Iranians. YouTube, Twitter, Facebook, the English and Persian versions of BBC, and websites of the major opposition candidates were all blocked. The government also blocked the blog host, blogfa.com, preventing many Iranian bloggers from updating their blogs. Although it stopped short of shutting down the Internet altogether, the government limited bandwidth and access speeds across the country. However, after years of Internet filtering, Iranian hackers and online activists had developed tools to circumvent government restrictions—an unintended consequence of government policies.

The government response to traditional and new media journalism was similar to their violent response to street protesters,
and continued through 2009. “At least a hundred journalists and cyber-dissidents have been arrested since the June 12th elections, and twenty-three of them are still being held” (Reporters Without Borders 2009). The Committee for the Protection of Journalists called Iran the “world’s leading jailer of journalists.” The government uses fear of violent action and criminal prosecution to intimidate journalists and further limit freedom of speech, and reports from Tehran indicated a government strategy of targeting citizen journalists during the protests (Committee to Protect Journalists 2009).

Among these three categories—information-sharing, creating a public space, and mobilizing protesters—social media has primarily been successful as a tool to share news and images with a global audience to gain the attention of the international media networks, and to create an open space for discussion and symbolic solidarity. It was not, however, a causal or practical mechanism that fueled or enabled the weeks-long protest. Government suppression, infiltration, and censorship limited the ability of protestors to use social media to mobilize, communicate, or share information with domestic audiences.

EXPANDING FREEDOM OF SPEECH

There are many different ways to understand a “free” media environment: constitutional provisions protecting press freedom and freedom of information; enforcement of the provisions; an absence of laws restricting reporting; freedom of media outlets to determine content; free access to sources; a lack of official censorship and journalist self-censorship; freedom of media outlets from intimidation and violence; freedom from economic control by the government; freedom from economic manipulations and bribery; and a transparency of ownership which allows “consumers to judge the impartiality of the news” (Whitten-Woodring 2009, 598). Especially in the context of reporting the elections and protests, the traditional and new forms of media in Iran arguably fail to meet almost every one of these requirements. The Iranian regime has
systematically and purposefully limited, if not eliminated, the mainstream media’s ability to act as a “watchdog” of government by denying it these freedoms.

The role social media played in creating a public space for discussion and solidarity, as well as distributing Iranian news and images internationally, calls into question some of the more stringent requirements for a media to be classified as “free.” A narrower definition of media freedom is the “ability to perform a watchdog role, monitoring and criticizing government behavior, because when the media performs this function it is able to act as a forum for political debate regardless of other limitations on its freedom” (Whitten-Woodring 2009, 601). Using this narrower definition of the media, new and social media forms were able to expand freedom of expression to the point where they were able to fulfill the “watchdog” role traditionally held by traditional media outlets in countries without pervasive government control and censorship of the press. Although the Iranian regime limited the impact of social media, especially with regards to domestic mobilization, communication, and information-sharing through censorship, blocking, and infiltration, it stopped short of shutting down the internet entirely, allowing digital “holes” to open up in an otherwise closed society.

New media provided a public space for debate and disseminated news and images about the protests, which questioned the legitimacy that the regime hoped to claim in holding elections. Information regarding police brutality, beatings, and arrests was leaked to the world through vivid photos, stories, and videos, despite government attempts to repress its release through stringent press policies. Although new media cannot fill the role of a traditional free press and constitutional protection of free speech, the mechanisms of information sharing and creating a public space allow it to partially take on the “watchdog” role of exposing government abuses and the manipulation of the elections.
EFFECT ON DEMOCRATIC NORMS AND RESPECT FOR HUMAN RIGHTS

Scholars and proponents of free speech have argued that freedom of speech is more important to democracy than the right to vote, relying on the belief that “if citizens have the right to complain, to petition, to organize, to protest, to demonstrate, to strike, to threaten to emigrate, to publish,” government will be more responsible and more responsive to its citizens (Mueller 1992, 984). According to this argument, the use of new and social media in expanding freedom of speech in Iran should have limited the Iranian regime’s human rights abuses.

However, during the months since Ahmadinejad’s election, Iran’s human rights situation has worsened, with mass arrests and detentions as well as brutal repression of demonstrations. According to Tehran’s prosecutor-general, Qorban-Ali Dorri Najafabadi, as of August 2009 more than 4,000 had been arrested in the capital alone. More than 240 prominent citizens were arrested by “unidentified agents and taken to undisclosed locations...[They are] held in incommunicado detention and have no access to legal counsel” (International Campaign for Human Rights 2009). These pervasive actions are explicit violations of the basic rights guaranteed by Iran’s legal codes and by the international treaties it has signed.

This repressive and abusive government response supports Jenifer Whitten-Woodring’s conception of the relationship between free expression and democracy: she argues that free press is a characteristic of democratic societies, and that in authoritarian states which lack democratic accountability, expansion of free speech actually begins a cycle of protest and government repression. Unlike in developed democracies, free speech under authoritarian regimes has a negative effect on the development of democratic norms and respect for human rights. In the case of Iran, which is by no classification a real electoral democracy, expansion of freedom of speech has brought with it increased arbitrary arrest and detention of journalists and dissidents, and distortion of the rule of law for protestors who have been incarcerated and criminally prosecuted.
Both on the streets and in the prisons, government repression has decreased respect for a fundamental human right to personal integrity. The right to the integrity of one’s own person, defined as the “right to life and the inviolability of the human person...[protection from] prolonged and arbitrary detention, extrajudicial killing, torture, genocide, and other severe violations of bodily integrity,” is considered indispensable to human dignity (Cingranelli 1999). The government response towards the protestors and dissidents has been swift and brutal, using systematic state violence to maintain political power and dissuade further uprisings.

A more complex question concerning social media’s role is whether the international dissemination of news, images, and videos of government human rights violations has helped mitigate the violent aftermath of the protests—limiting what could have been a more severe crackdown on political agitators and protestors without the worldwide attention created by new media. Google chief executive Eric Schmidt said he hoped that the many clips of violent protest scenes posted on YouTube has helped to “moderate an over-reaction by the government.” Schmidt refers to the theorized “YouTube effect” whereby “video clips, often produced by individuals acting on their own, are rapidly disseminated throughout the world thanks to video sharing websites...[Every] month, YouTube receives 20 million visitors who watch 100 million videos each day” (Naim 2007, 102)

In the “double-echo chamber”—where web-content which is re-aired by TV networks and is made permanent through web distribution by bloggers—these videos are expected to bring greater transparency and accountability to governments in traditionally closed and undemocratic societies. Building on the CNN effect of the 1990s, scholars hypothesize that wide international dissemination of these videos will help to determine foreign policy and humanitarian interventions by enhancing public knowledge of government atrocities. They expect the dissemination effect to be even greater than that of mainstream news networks: “International news operations may have thousands of professional journalists, but they will never be as omnipresent as millions of people carrying
phones that record video” (Naim 2007, 103).

Iranian government leaders, while outwardly blaming the foreign media and Western government officials for fueling the protests and meddling in their affairs, did feel the need to justify the election and protests to the world. The current regime claims its legitimate rule is derived from two forces: divine favor and popular support stemming from the revolutionary heritage of 1979. In this sense, maintaining at least a façade of democratic participation through popular elections is important to the its political power. In an attempt to re-establish legitimacy after the protests, Khamenei publicly announced an investigation regarding the accusations of voter fraud a few weeks after the election, and eventually declared Ahmadinejad the winner (BBC 2009). The digital movement has eroded the legitimacy created through the Iranian electoral process and limited what may have been even greater human rights violations against the protestors. Despite this, the government is still not being held accountable for violating the physical integrity of its citizens. Foreign governments and international organizations are widely abstaining from commenting on or involving themselves in Iran’s internal affairs. U.N. Secretary General Ban Ki-moon sent a traditional congratulatory message to Ahmadinejad after his inauguration in August of 2009, despite a hunger strike in front of the United Nations’ New York headquarters led by dissident journalist Akbar Gangi calling for the release of political prisoners arrested in the election aftermath (Penketh 2009).

PUTTING IT IN CONTEXT:A GREEN REVOLUTION?

In order to apply the lesson we learned through the Iranian protests about the role new and social media can play in demonstrations against repressive regimes and within closed societies, we must evaluate whether or not the movement can actually be categorized as a “revolution.” Jack Goldstone defines a revolution as “an effort to transform the political institutions and the justifications for political authority in society, accompanied by formal or informal mass mobilization and non-institutionalized actions that
undermine authorities” (Goldstone 2001). International audiences and mass media outlets have characterized the protests surrounding the Iranian presidential election, and the demonstrations that have continued throughout the summer and fall of 2009, as signifying a fundamental, revolutionary shift in the Iranian political structure that has not been witnessed since the Islamic Revolution of 1979 (Sullivan 2009). While it is true that the protests have harnessed much of the discontent and unrest within Iran, especially among younger activists and women who supported opponent candidate Mousavi, there has not been a widespread effort to supplant the regime that was put into power in the late 1970s by mass demonstrations and student movements. The protestors were mobilized by what they believed were the fraudulent results of an election that was undemocratic to begin with—the Islamic leaders vetted the potential candidates and determined who was allowed to run for the office in the first place. It has been suggested that even if Mousavi had been elected, no fundamental changes would have occurred in Iranian society or political sphere (Schectman 2009). Whether or not the continuing protests and growing civil resistance movement develop into a revolution with the larger goal of changing the structure of Iranian government cannot yet be determined. Since the June 2009 elections, protestors have used several national holidays as opportunities to express their continuing discontent with the regime, and media coverage has served to re-draw the world’s attention to the situation.

The role social media played in Iran’s “summer of discontent” signifies both the possibilities and limitations of digital resistance within traditionally closed societies. A true revolution may develop, similar to the “color revolutions” earlier this decade that emerged from contested elections in the former Soviet Union and Balkan states. However, for this to happen, the spirit of the protests and desire for fundamental political change would have to spread beyond Tehran’s activists to Iranian society as a whole, including the lower classes, rural inhabitants, and the older population. Until this occurs, the regime may still claim legitimacy through popular support and blame Western influences and media for the uprisings
against the government. The Internet and social media are limited in their ability to engage and mobilize these other sectors of society who are not technology-savvy enough to surpass pervasive government filtering and censorship, who distrust the information provided by new media forms, or who may not have access to the technology or tools necessary (Fairbanks 2009). The barriers of digitalized resistance help explain why participation in the demonstrations that have continued throughout the fall has remained limited to the original June protestors—young activists in Tehran. The continuing protests have been met with government repression both on the streets and online, and the regime has shown its ability to adapt its censorship techniques to new technological developments. Reporters Without Borders found that the government further limited access speed and increased webpage blocking days before the anticipated protests on November 4 and December 7 to prevent online organizing and mobilization.

Examination of the use of new media and online social networking has shown a remarkable ability, in the context of resistance movements, of protesters to harness the new technology available to provide almost instantaneous information and images to a global audience, as well as to create a global cyber-civic society in which individuals and organizations may discuss and collaborate around the movements (Dabashi 2009). The worldwide attention, networks of solidarity, and media coverage of the Iranian demonstrations have prompted the street protests to continue, despite limited gains and costly violent suppression and abuse. Without these tools, the dissidents would be protesting in a traditional closed society with no ability to communicate or share information with audiences abroad.

Using social media tools, Iranian protesters were able to greatly expand their freedom of speech and information-sharing abilities, but often at the cost of increased human rights abuses, arbitrary arrest, and torture. Beginning in August 2009, the government began to legally prosecute prominent protestors, journalists, and dissidents. The five death sentences handed down thus far serve as a constant reminder to potential protestors that the regime still
wields substantial power, and accusations of rape, torture, murder, and forced confessions re-emphasize the cruel acts through which the regime’s authority is be manifested. However, the vast and continuing international media coverage of the so-called “Twitter Revolution” as well as the regime’s response has also shown that the government will never be able to fully silence the Iranian electorate as long as digital protestors continue to tweet, upload pictures and videos, and blog for a global audience. Within this framework, the world—including foreign governments and international organizations—will continue watching Twitter feeds and You Tube videos as events in Iran unfold.

Works Cited


Pleming, Sue. “State Department speaks to Twitter over Iran.” *Reuters*, June 16, 2009.


The sovereign state is a basic and familiar unit of political organization in the international system. It is often considered a stage in the evolution of territorial states in which leaders are accountable for the security and well-being of their citizens (Weiss 2004, 44). In its essence, sovereignty's function is to secure power and order in the state. That power includes the ability and the authority to exercise force. To understand why sovereign actors choose to use force rather than other tactics to meet their goals, Robert Art explains that although military power is more useful for realizing certain goals over others, it “is generally considered of some use by most states for many of their goals” (Art and Jervis 2005, 141).

The use of force, “a continuation of policy through militaristic means,” reflects different perceptions of power among various political actors (von Clausewitz 1976, 27). Consequently, when a state acts, a balance of power schema emerges in which each actor hedges its own strength against all other actors within a given level sphere of activity, be it domestic, regional or global. As Inis Claude explains, the balance of power is designed to persuade actors to behave in certain ways in order to limit conflict. While Claude’s analysis helps predict the timing of conflicts based on managing and elucidating power relationships (Claude 1962), Art suggests that a lack of transparency among potential combatants is the cause of different uses of force.

Even the keenest understanding of the balance of power in a given situation would not guarantee the ability to contain every conflict. In each situation there is a variety of tactics and strategies
that actors may pursue, including unilateral, bilateral and multilateral action. Unilateralism, as understood in contemporary discourse, is a concept in which a given state engages in an act on its own as an expression of its sovereignty. Typically these actions are made without regard to other potentially affected parties, and are often against the wishes of those parties (Kane 2006, 2). Common unilateral actions include imposing economic sanctions on a belligerent nation, invading a neighboring country, or creating an imposed barrier between territories without the other party’s consent.

This analysis will discuss the unilateral withdrawal from a shared-border territory without a bilateral or multilateral agreement. In such a situation, the relatively stronger military power decides to unilaterally partition or to disengage from a given conflict area. Typically, this occurs after political-military goals have not been met and seemingly cannot be met without high risks or costs. Although withdrawal does not terminate the conflict, it can serve as a moderate development that expresses the goal of managing violence without the direct use of force (Bar-Siman Tov, 262). The mitigation of the conflict is dependent on the other parties’ favorable reception of the unilateral act.

In order to further understand the decision-making process that leads to unilateral withdrawal, this paper will focus on two recent Israeli examples: the military pullout from southern Lebanon in 2000 and the disengagement of settlers and military posts from the Gaza Strip and northern West Bank in 2005. An analysis and comparison of Israeli leadership and discourse in the two episodes will show the factors and common characteristics behind the repeated Israeli decision to unilaterally withdraw as a means to lessen conflict. It is important to note that this analysis will be largely restricted to the internal Israeli debate regarding its own policies, and that only Israeli factors will be considered, insofar as they relate to the specific concerns within decision-making processes. Further work must be conducted in order to examine the concurrent debate that unquestionably arose within Lebanese and Palestinian societies.

In the two Israeli cases, unilateral withdrawal initially allowed
for short-term gains for the state, but was ultimately responsible for negative long-term results. Though they may fulfill their short-term goals, policymakers hurt their broader ambitions of lasting peace and stability. The Israeli examples, then, must be understood to have wider regional and global relevance. A deeper understanding of its usage is particularly relevant to the Middle East peace process, in which unilateral action has been recurrent.

THE WITHDRAWAL FROM LEBANON, MAY 2000

As Avraham Sela explains, the Israeli presence in southern Lebanon and its “security zone” were products of an Israeli attempt “to prevent Palestinian guerrilla infiltrations into Israeli territory.” Only after the 1982 war between Israel and Lebanon did indigenous guerrilla groups begin their efforts to drive the Israelis from Southern Lebanon (Sela 2007, 60). To the Israeli public, the financial and military burden of the Israeli-Lebanese conflict remained an issue far removed from everyday life until 1997, when a military helicopter accident killed 73 Israeli troops in transit to the security zone (Clawson and Eisenstadt 2000, 89). Following the increase in media exposure and the development of grassroots protest organizations, Israeli public opinion began to shift toward a reevaluation of the military presence in southern Lebanon. By 1999, public opinion polls indicated that over 55 percent of the Israeli public supported the withdrawal of Israeli troops from Lebanon, and over two-thirds favored a reassessment of Israel’s military policies there (Sela 2007, 70). Additionally, lobbies and parliament members, known as Knesset Members (MK), began pressuring the government to remove troops from the security zone.

In the 1999 parliamentary elections, candidate Ehud Barak ran on a platform of “returning the boys home.” This campaign promise to withdraw troops from the southern Lebanon “security zone” marked a substantial shift in both Israeli public opinion and military procedure. To many observers the impetus for withdrawal was directly linked to the costs borne by the Israeli military and psychologically stamped on the collective Israeli conscious. Since
the formalization of the security zone in 1982, over 850 Israeli soldiers had been killed. As the Israeli government took responsibility for funding the South Lebanon Army (SLA), its financial commitment had increased by tens of millions of U.S. dollars (Clawson and Eisenstadt 2000, 89). Struck by the high costs of the Israeli occupation, many Israelis began to question the purpose of the security zone’s existence. The traditional Israeli stance emphasized the importance of a buffer zone between Palestinian and Lebanese guerillas in order to protect the northern Israeli border from attacks (Sobelman 2004, 9). However, multiple air raids and larger campaigns by the Israeli air force to combat militant rocket attacks challenged the myth of a quiet border (Sela 2007, 65).

As Israeli withdrawal became more likely, Hezbollah leaders began emphasizing their role in the departure of Israeli forces from Southern Lebanon. They pointed to their “group’s continued ability to attrite Israeli forces” as the key factor sparking Israel’s domestic dissatisfaction with its military presence in Lebanon and the subsequent removal of Israeli troops.1 In light of this rhetoric, Israeli political and military leaders engaged in a robust debate regarding a potential withdrawal. Many in the military establishment, such as Brigadier General Shlomo Brom, strongly recommended “Israel’s continued occupation of the security zone until a peace treaty with Syria is signed” (Clawson and Eisenstadt 2000, 90). Daniel Sobelman adds that much of the military brass was opposed to a unilateral pullout that would allow militant groups, mainly Hezbollah, to gain strength and mobility closer to the Israeli border. Their fear was grounded in an understanding that future tensions with armed militants would create an escalation in violence leading to cross-border conflict (Sobelman 2004, 30). Other critics feared that if the security border zone were dismantled, groups such as Hezbollah would not end their resistance until all of Israel was destroyed (Jacoby 2007, 121). While the security zone had not achieved a full cessation of violence, it successfully halted cross-border guerilla infiltrations.

Yet these opinions hardly amounted to a consensus—even the military was divided. Major General Amos Malka, head of Is-
raeli Defense Forces (IDF) Military Intelligence, argued that an Israeli withdrawal would begin an era of mutual deterrence on both sides of the border. Using the concept of “sufficient gain,” Malka emphasized that each actor was committed to preventing civilian casualties and avoiding a dangerous escalation. Leading Israeli MK Yossi Beilin’s plan calculated that an Israeli withdrawal from Lebanon would address Syrian and Lebanese hopes for stability and prosperity. Assuming that Hezbollah had limited objectives that would be satisfied by an Israeli withdrawal, Beilin believed that the group would restrain its future operations (Sobelman 2004, 36). The inevitability of an Israeli withdrawal became apparent as the debate continued. Many leaders began to recognize that if “Israel’s actions [earned] regional and international legitimacy,” there would be a strong possibility of deterring further attacks and limiting Hezbollah (Sobelman 2004, 31). An Israeli withdrawal, many reasoned, would deprive Hezbollah of its ideological motive for existence (Jacoby 2007, 120).

Prime Minister Barak used this division among the elites to advance his withdrawal plan in spite of the failed diplomatic efforts with Syria that dominated much of his first year in office. Rather than back away from his campaign promise, Barak used the failed peace initiatives as further support for a new direction in conflict resolution, making the final decision in March 2000 (Sela 2007, 71). Responding to concerns among military leaders, Barak noted that “it is incumbent on Military Intelligence to envision worst-case scenarios,” but that as a political leader he was forced to make difficult decisions that corresponded to changing political realities (Sobelman 2004, 31). The decision to unilaterally withdraw all Israeli soldiers from southern Lebanon without a formal agreement or understanding from a regional actor continued to raise debates and concerns about the consequences and aftermath of such an action. Nevertheless, on May 24th, 2000, the last Israeli soldier left Lebanon in a hastily-conducted withdrawal that shut down the border between the two countries (Blum 2007, 201). Crucially, Israeli forces only withdrew from territories it considered part of the security zone that was created as a result of the 1982 war. Another
22 square kilometers of territory captured during the 1967 campaign known as the Shebaa Farms remains under Israeli control. Unlike the security zone, the Shebaa Farms are contested by Israel, Lebanon and Syria.

DISENGAGEMENT FROM THE GAZA STRIP, AUGUST 2005

Drawing on Israel’s experience of withdrawing from southern Lebanon as well as the removal of Israeli settlers from the Sinai desert in the late 1970s, Prime Minister Ariel Sharon announced a new unilateral plan involving the Israeli-Palestinian conflict in 2003. At the Herzliya Conference on Israeli National Security in December of that year, Sharon embraced the idea of unilateral disengagement and warned the Palestinian leadership, particularly the Palestinian Authority, that a failure to cooperate and move toward peace would lead “Israel to initiate the unilateral step” of removing all settlers from the Gaza Strip and dismantling four settlements in the northern West Bank (Zelnick 2006, 35). Although an Israeli withdrawal would be a first step in the long-desired reestablishment of Palestinian sovereignty in a territory controlled by Israel, many Palestinians were wary of Sharon’s decision. To many Arab leaders, Sharon’s hawkish history of supporting settlements in occupied land and his notorious military background did not qualify him to be a partner for peace (Zelnick 2006, 65). Comments made by Sharon prior to the disengagement fueled Arab fears that the Israel prime minister’s calls for withdrawal were being used to manipulate the situation to Israel’s benefit. At an internationally-sponsored peace initiative proposal in 2003, for instance, Sharon warned that continued hesitation to cooperate would lead the Palestinian people to “receive far less than they could obtain in direct negotiations based on the Road Map” (Bar-Siman-Tov 2007, 262).

While the disengagement plan allowed Israel to retain strategic powers in the territories by preserving its military power, border control, and water rights, disengagement was also seen as a way to lessen the enormous military and financial costs of maintaining the Gaza settlement. By the time of Sharon’s announcement in De-
cember 2003, there were only eight thousand Jewish settlers residing in the Gaza Strip among over 1.3 million Palestinians. Jewish settlers, under the protection of the Israeli military, controlled over 40 percent of arable land and 50 percent of the available water resources (Baylis 2009, 151). Under these circumstances, many Israeli leaders viewed disengagement as a way for Israel to define its own borders by its own terms (Pressman 2006, 360).

Much like Barak’s action on Lebanon in 2000, Sharon’s decision to disengage drew heavy criticism. Many among the military establishment felt that leaving Gaza would strengthen the myth that terrorism can defeat national armies. Retired Israeli Major General Yaakov Amidor labeled the operation “a strategic error of historical magnitude,” arguing that Palestinian militants had witnessed the effects of Hezbollah violence in southern Lebanon and emulated their tactics to achieve the same outcome. Amidor claimed that disengagement would establish a state in which “Hamas will have freedom of action and be joined by the umbilical cord to Hezbollah,” thus increasing the security threat on multiple borders (Zelnick 2006, 35). Yet despite heavy criticism, particularly among his own party, Sharon executed his plan. By the summer of 2005, the last Jewish settler was removed from Gaza.

REASONS FOR ISRAELI UNILATERALISM

Before analyzing and comparing the motives underlying each case of unilateral action, it is necessary to first understand the general differences in their implementation. In its May 2000 withdrawal from southern Lebanon, Israel ended a longstanding military presence in a foreign, hostile land by removing its posts and installations from a security zone that had been created to buffer Israel proper from militant activity. In Gaza, on the other hand, the Israeli army undertook a much larger operation: they not only dismantled military positions that had been held for almost four decades, but were also charged with the removal of eight thousand Israeli citizens. Although differences abound, Israel acted unilaterally in both cases, and examining common variables will help us better
understand Israel’s reliance on this method. And though each situation presented unique difficulties, and, as indicated, drew unique criticism, both withdrawals achieved the similar result of formally ending the permanent Israeli presence in those areas. This analysis will highlight three factors that influenced Israeli decision-making in both 2000 and 2005: the increase in violence, the perceived lack of political openings and partnership, and the latitude to use the withdrawal to strengthen claims to other disputed territory.

In July 2000 a poll was conducted among Palestinians on the role that violence played in Israel’s decision to withdraw from southern Lebanon. Sixty-three percent of respondents believed that Palestinians should emulate Hezbollah’s violent tactics. A majority of Israelis has consistently agreed that “armed confrontations have helped them achieve national rights in ways that negotiations could not” (Shamir 2007, 37). These sentiments echoed the Israeli criticism that withdrawal could strengthen the bases of support for militant activities. Yet by 2000, over 850 IDF soldiers had been killed as a result of violence in southern Lebanon (Clawson and Eisenstadt 2000, 89). Similarly, in Gaza, the asymmetric war against guerilla-style militants continued to cause Israeli military and civilian casualties. In the second intifada, which began in 2000 and spanned six years, over one thousand Israelis were killed (Smith 2010, 498). The growing perception of the cost of war among Israelis was a potential limit to Israel’s military options. Jeremy Pressman writes that ongoing conflicts showed that Israel “could not bring about an end to the conflict short of wiping out the Palestinians, something Israel was not prepared to do” (Pressman 2006, 370).

Israeli public opinion began to reflect a new understanding of violent conflict. The Israeli public grew tired of prolonged conflicts. Many began to see the IDF presence in southern Lebanon as more of a war and less of a security operation (Blum 2007, 237). By 1999, over 55 percent of the Israeli public supported the withdrawal of Israeli troops. Similarly, 60 percent of Israelis were in favor of Gaza disengagement by 2004, as other strategies aimed at ending the conflict appeared unsuitable (Bar-Siman-Tov 2007,
In the late 1990s, grassroots organizations such as Yesh Gvul ("There is a limit"), a group of army reservists, and Four Mothers, families who had lost relatives as a result of the conflict, began criticizing the military establishment’s decision to remain in Lebanon. These groups represented growing support behind the emergent opinion to leave Lebanon (Jacoby 2007, 124). A growing number of causalities, dramatically increased by a helicopter collision that killed dozens of soldiers, kept the issue in the public spotlight and reminded Israelis that unilateral withdrawal was a means to end the violence. The IDF received an increasing number of letters regarding the occupied territories, as well as requests by soldiers who refused to serve in the West Bank and Gaza (Bar-Siman-Tov 2007, 264).

At the same time, Israel’s adversaries began to understand the connection between increased violence and political gain. Hezbollah leaders pointed to their group’s ability to “attrite Israeli forces” as the key factor behind the Israeli debate over unilateralism (Clawson and Eisenstadt 2004, 8). For Palestinians, military resistance brought benefits that decades of negotiations and diplomatic arrangements had failed to produce. The Israeli explanations of voluntary withdrawal and strategic disengagement did not seem to convince the Palestinians and Lebanese. Instead, they believed that their own actions were the leading cause of Israeli unilateral withdrawal (Efrat 2006, 190). In both cases, the costs of a continued Israeli presence, most notably Israeli casualties, had outweighed the benefits of remaining, even if not all of Israel’s strategic objectives had been met. Critics have debated the amount of political flexibility that the Israeli government had in withdrawing. Sela has emphasized in his analysis the role of public opinion and grassroots pressures, while Bar-Siman-Tov has focused on military maneuvering of a disengagement that retained Israeli preponderance at a significantly reduced risk (Bar-Siman-Tov 2007, 261).

Although the increase in violence suggests a motive and justification for withdrawal and separation, that conclusion fails to highlight the impetus for acting unilaterally. As Bar-Siman-Tov indicates, it is extremely rare to find a situation where the “initiat-
ing side prefers, from the outset, unilateral disengagement without an agreement as a conflict management strategy” (Bar-Siman-Tov 2007, 262). Therefore, it becomes necessary to replace the respective withdrawals within the larger political framework of their time. Although Barak ran for office under a platform of withdrawing the troops from Lebanon, he spent much of his first year in office seeking peace with Israel’s Arab neighbors. During that time, there was a focused attempt for an agreement with the Syrians that would involve both the return of the Golan Heights and the withdrawal of IDF troops in exchange for peaceful relations with Israel. However, negotiations halted after ten months when Syria proved unwilling to accept Israeli demands (Sela 2007, 71). Meanwhile, Israeli leadership continued to ignore Hezbollah as a suitable partner for negotiation. These events, coupled with the protracted violence, eroded public support for a prolonged occupation and forced Barak to deliver his campaign promise without a Syrian or non-state partner (Clawson and Eisenstadt 2004, 89).

In the Palestinian context, Israeli leadership looked at the breakdown of peace talks following the Camp David summit and Taba in 2000-2001 as evidence that the Palestinian Authority did not want a negotiated settlement (Pressman 2006, 369). In the eyes of the Israelis, the outbreak of the second intifada and the years of persistent violence and terrorism indicated “a lack of morality” and that there was no true Palestinian partner ready to negotiate (Bar-Siman-Tov 2007, 264). Israel chose to isolate—not work with—Palestinian Authority leader Yasser Arafat. To many Israeli leaders, Arafat represented a Palestinian effort dedicated to the destruction of Israel through both violence and demographic pressure (Zelnick 2006, 26). They pointed to his insistence on the Palestinian right of return, his refusal to disarm militant groups such as Hamas and the Palestinian Islamic Jihad, and his tampering with other elected officials such as Abu Mazen. Even after Arafat’s death in 2004, the Sharon government refused to recognize a Palestinian partner due to Mazen’s inability, or possibly his reluctance, to end media and educational propaganda against Israel (Bar-Siman-Tov 2007, 271). This denial of any potential partner is best exemplified in a letter
sent from Sharon to U.S. President George W. Bush, in which Sharon explains the rationale of disengagement because “there exists no Palestinian partner with whom to advance peacefully toward a settlement” (Bar-Siman-Tov 2007, 268).

Unilateral withdrawal represents a decision among Israeli leaders to avoid being forced into untenable political scenarios, or to wait indefinitely for partners they perceive to be genuine; it was hailed by the international community as a creative response. It is important to note that both Hezbollah and the Palestinian Authority regularly blame Israel for the increased violence, and reject the idea of a lack of partner as an Israeli refusal to restart the peace process. The Israeli government was skeptical, if not uninterested, in small confidence-building attempts by the Palestinians, such as small shifts in policy and temporary cease-fires (Pressman 2006, 372). In Lebanon, Israel overlooked discredited local parties, choosing instead to work solely with the United Nations. Yet Israel’s unilateral behavior was grounded in its implicit cooperation with other parties. The withdrawals were conducted under international guidance and with international support. And in both Lebanon and Gaza, the adversarial Arab factions drew down or completely halted attacks to allow for the removal of an Israeli presence from their lands (Makovsky 2005, 26).

The final, and perhaps most pertinent, factor was Israel’s choice to withdraw from certain areas in order to successfully defend other territory under its control. During the Israeli-Syrian peace negotiations prior to the Israeli withdrawal, it was widely believed that the removal of troops would be tied to the return of much, if not all, of the Golan Heights to Syria. Throughout the negotiations, Hezbollah, acting as Syria’s proxy, continued its armed struggle in the hopes of pressuring Israel into giving up other areas such as the Golan Heights and the Shebaa Farms (Sobelman 2004, 68). By acting unilaterally, but still having its withdrawal internationally recognized by the United Nations, Israel not only wedged a divide between Syria and Lebanon, but also shifted the talking points in the negotiations with Syria. Prior to 2000, it was expected that a peace deal with Syria (including the Golan) would
lead Syria to exert its influence over Lebanon and follow suit with an additional peace deal (Hof 2000, 3). The unilateral withdrawal changed this equation so that the fate of the Golan Heights now resided squarely with Israel (Blum 2007, 237). Should Israel choose to transfer the Golan Heights to Syria, it will likely do so on its own terms.

Similar motivations and goals seemed to have been behind the disengagement from Gaza. That move, combined with a physical security barrier under construction in the West Bank, has contributed to a growing consensus that Israel’s move in 2005 signaled a shift in the geographic focus of settlement, rather than a total abandonment (Pressman 2006, 370). Palestinian leader and Fatah member Mohammed Dahlan blasted Sharon over his West Bank intentions and argued that disengagement was nothing more than a ploy to strengthen other settlements (Zelnick 2006, 63). At the same time, President Bush responded to Sharon’s explanation of the disengagement by affirming the Israeli point of view. In his a letter he wrote: “In light of new realities on the ground, including already existing major Israeli population centers, it is unrealistic to expect that the outcome of final status negotiations” will represent a complete withdrawal to the armistice line of 1949 (Smith 2006, 530). To the Israelis, this constituted a milestone in U.S.-Israeli relations, confirming in writing that the United States understood Israel’s strategic and demographic realities.

In such a framework unilateral withdrawal allowed Israel to dictate the terms under which land was being transferred. Israel was not forced to withdraw by violence or a political dead-end. As a result, these moves relieved international pressure for a more comprehensive and expansive land transfer (Bar-Siman-Tov 2007, 265). Unilateral action was an attractive option, as it not only sought to stabilize areas of confrontation and reduce military expenditures, but also preserved valuable Israeli territory—the Golan Heights and the West Bank (Hovsepian 2008, 282). Leaving these territories with an understanding of retaining others was strategic and came at a reasonable cost to much of the Israeli public, since practically no future peace plan included the Gaza Strip or south-
ern Lebanon in Israel (Efrat 2006, 184).

CHANGING THE RULES OF THE GAME

In effect, the two withdrawals have come to be understood as “down payments” on the respective territories Israel retained. By withdrawing, Israel has delayed the political processes in both arenas for the foreseeable future (Thomas 2009, 152) and strengthened its own position domestically and internationally. As Dov Weisglas, the senior advisor to Sharon during the disengagement, explained in an interview, “With the proper management we succeeded in removing the issue of the political process from the agenda. And we educated the world to understand that there is no one to talk to” (Shavit 2004). Fundamental to this line of thought is the idea that unilateral withdrawal provides added security. In its essence, unilateral disengagement “aims more to minimize damages and losses than to maximize gains” (Bar-Siman-Tov 2007, 263). As such, it is chosen only when it makes a situation less dangerous than alternative conflict-resolution tactics.

Many critics of the Lebanon withdrawal warned that Israel would create a power vacuum, which would lead to a large-scale intensification of the conflict. Fearing “far reaching” consequences from both the Syrians and the Iranians, military and political leaders alike assailed the initiative (Sobelman 2004, 11). Similar arguments were levied against the government five years later as the disengagement was charged with strengthening Palestinian militancy and Arab solidarity against Israel. Yet in Gaza, the continued military and civilian presence had led Sharon to believe that not to disengage was actually more dangerous to Israel (Bar-Siman-Tov 2007, 273). The redeployment of troops to the Israeli side of the border in each area has both reduced sources of friction and provided fewer targets for attacks (Efrat 2006, 187). While Israel removed its troops from within the Gaza Strip, it maintained its position of ensuring security for Israeli citizens and, as such, reserved the right of future military engagements within the territory—a position markedly different from its complete withdrawal
from southern Lebanon (Hovsepian 2008, 283). That strategy was strengthened by a significant reduction of violence over a six-year period from 2000 to 2006 in southern Lebanon (Hovsepian 2008, 282). In Gaza, removing Israeli troops was also a removal of potential targets for Palestinian militants.

In choosing to act unilaterally, Israel appeared to forsake bilateral and multilateral approaches in favor of an understanding that separation creates the least-negative results. As each withdrawal approached, some analysts began touting the idea of “mutual deterrence,” since Israel’s opponents might also be nervous about upsetting the newly-defined status quo. (Sobelman 2004, 33). These analysts saw the lack of large-scale attacks following the withdrawal as affirmation of their hypothesis. Withdrawal was a calculation by Israel, a bet that the risk of having dangerous neighbors is preferable over a costly, and potentially deadly, continued occupation (Zelnick 2006, 29).

Between 2000, the year of the withdrawal from southern Lebanon, to 2006, the year following the disengagement from Gaza, there was a steady decline in Israeli public support for the notion of a Greater Israel (Shamir 2007, 34). This unilateral mentality confirms a new point of view among Israeli leadership known as the “demographic sense argument,” which justifies withdrawal based on current demographic realities and future trends (Thomas 2009, 152). The argument prioritizes an Israel that has an overwhelming majority of Jewish citizens rather than a larger geographic area that also includes non-Jews. This shifts the impetus for peace and conflict-resolution away from humanitarian concerns and international pressure toward an inward looking understanding of the conflict. By disengaging from Gaza, the Jewish state sidelined 1.3 million non-citizens, placing them under their own Palestinian governance. Similar withdrawals from parts of the West Bank may follow in order to advance this line of thought to its logical conclusion: an Israeli state entrenched in its own territory with only a limited amount of non-Jewish citizens (Brom 2007, 15). An added incentive to this schema is that it allows Israel to put pressure on Palestinian leaders to showcase their own leadership abilities and
assume more direct responsibility for their people, a feat that has been hard to achieve due to international isolation and an economic blockade on Gaza (Bar-Siman-Tov 2007, 274).

Although U.S. policy is that “no action should be taken by any party to the peace process that would prejudice the outcome of final status negotiations” (Muasher 2008, 206), the Israelis appear to be lowering the expectations of the Palestinians (Zelnick 2006, 65). Withdrawal represents an internalization of the two-state solution and a reluctance to fight a demographic battle with the Palestinians, while the process of disengagement enabled the Israelis to redefine the physical and metaphorical boundaries of negotiations. The current Israeli position has become very clear: “Defense and security preclude a complete return to the 1967 borders” (Zelnick 2006, 65).

WILL UNILATERALISM CONTINUE?

By withdrawing from Lebanon and disengaging from Gaza, the Israeli leadership was able to gain international support, reduce its financial liabilities, and dramatically reduce violence in the short term. Israel has maintained its military superiority through incursions and its economic power through its heavy blockade of Gaza. Yet two questions remain.

First, will Israeli leaders continue to implement unilateral withdrawals as a means of addressing conflicts in the future? And second, will other actors adopt similar actions when faced with a political stalemate? To answer these questions, we must first look at the longer-term effects and outcomes of the two prior examples.

Although the first six years after the withdrawal from Lebanon yielded substantial short-term security and economic gains for both Lebanon and Israel, the outbreak of large-scale violence in the summer of 2006 casted significant doubt on the prudence of unilateralism. The border with Lebanon remained hostile due to the Lebanese government’s inability to disarm militant groups such as Hezbollah, as well as the occurrence of skirmishes and Israeli air-raids. Limiting Israel’s deterrent capability, Hezbollah had free-
dom of travel throughout Lebanon and gained much closer access to Israeli population centers (Brom 2007, 9). Finally, in response to an increase in rocket attacks and the kidnapping of soldiers, Israel engaged Hezbollah in a 35-day military conflict, known as the Second Lebanon War, which resulted in the deaths of hundreds and the destruction of civilian infrastructure along the Lebanese-Israeli border. To many, the war demonstrated Israel’s vulnerability to a guerrilla movement in an area that it had evacuated without a security arrangement (Jacoby 2007, 112). Many critics directly attributed the 2006 war and Hezbollah’s growing military capabilities to the hasty May 2000 withdrawal.

A similar power vacuum arose in the Gaza Strip following the Israeli disengagement. After Israeli forces withdrew, it became increasingly clear that true power resided with Hamas and not the once-dominant Fatah party. In the year following the disengagement, Hamas succeeded in making dramatic political gains by participating in elections and, through a military putsch, consolidated its power over the entirety of the Gaza Strip. Many Israeli critics of disengagement have pointed to Israel’s perceived weakness in retreating as the main cause for the group’s successes. This reflects a larger trend among Arab media and political outlets to frame both the 2000 withdrawal and the 2005 disengagement as Arab victories over Israeli aggression. Militant figures, such as Hassan Nasrallah, head of Hezbollah, and Iranian President Mahmoud Ahmadinejad, have become famous as a result of their defiance and militancy against Israel (Indyk 2009, 388). Some have even linked Hezbollah’s success in 2000 to the outbreak of the second intifada only a few months later (Blum 2007, 241). As a result of the Hamas elections, Israel blockaded the Gaza Strip in an effort to isolate the new government. Further aggression between Hamas and Israel has resulted in a continued military presence in and around Gaza, and multiple flare-ups have inflicted extreme consequences, such as the continued loss of life.

Despite the ambiguous, if not outright negative, long-term track record of unilateral withdrawal, the Israeli public appears steadfast in its desire to end the territory dispute. Over sixty per-
cent of Israelis still prefer reconciliation, although public opinion in favor of a two-state solution with the Palestinians dropped ten percent after the Hamas elections, (Shamir 2007, 23). Many had hoped that disengagement would lead to a new dynamic for the peace process. Indeed, before violence erupted in 2006, Sharon’s successor, Ehud Olmert, proposed a broad withdrawal plan from the West Bank known as “convergence.” The stated goal of this plan was to withdraw 70,000 additional settlers from the West Bank and to dismantle a vast majority of the settlements (Jacoby 2007, 44) with the understanding that the remaining settlements would be incorporated into the final borders of the State of Israel (Zelnick 2006, 154). While some settlements would be dismantled, the vast majority of settlers would be retained. However, the ensuing events and a dramatic increase in violence ended much of the optimism and stalled further unilateral efforts (Efrat 2006, 187).

The Second Lebanon War effectively shelved all further unilateral action. While Olmert’s convergence plan shares characteristics with the two prior withdrawals—such as the refocusing on important territory and a perceived lack of responsible partnership—the increase in violence as a result of unilateral withdrawal has shown the counterproductive and dangerous nature of conflict reduction via separation (Shamir 2007, 57). Current Israeli leadership appears unwilling to subjugate itself to further criticism and potential violence without assurances or agreements from other parties (Zelnick 2006, 136). Although violence has dropped in both magnitude and frequency, periodic escalations remain a looming threat in Gaza and on Israel’s northern border. Unilateral withdrawal has failed to usher in a renewal of peace talks or even of security arrangements; “joint management” of the conflicts has not been established (Bar-Siman-Tov 2007, 278). Instead, unilateral withdrawal has indicated solely that Israel has a propensity to act on its own. These actions have not only silenced other interested parties but have led to spikes in violence as Israel sidelines other tracks of diplomacy. Israel—and other powers—should avoid future unilateral withdrawals given that they may instigate increases in violence.
Although many Israelis fear a further unilateral withdrawal from the West Bank will bear similar results, the future of this policy nevertheless remains unclear. Whereas the models set forth in southern Lebanon and Gaza do not constitute successful policy, Israeli leadership can look at the Israeli withdrawal from Egyptian territory after the Camp David Accords as a more suitable example. In that scenario, territory was transferred from one government to the other, which allowed for consistent policing and the maintenance of a lasting peace (Indyk 2009, 409). While not purely unilateral, the Egyptian withdrawal demonstrates the need for cooperation and counts regional stability among its potential results;—it shows that working with another state partner changes a state's considerations and, likely, the outcome. Despite heavy costs, Israeli leadership has repeatedly shown its ability to act alone in an effort to dictate future settlements. In doing so, it has redefined conflict resolution by prioritizing its own security needs at the cost of final-status negotiations.

In such a paradigm it appears that Israel—and perhaps other nations—will continue to act unilaterally as long as separation remains the “least bad” solution that is most likely to guarantee security. Israel’s historical reliance on unilateral action demonstrates its strategic commitment to the policy. Accordingly, one can expect Israel to unilaterally withdraw from a territory if its leaders perceive an environment that is characterized by an increase in violence, a lack of political openings and partnership, and the latitude to use the withdrawal to strengthen claims to other disputed territory. Future decisions to implement unilateral withdrawal must balance the goal of conflict resolution with its historical precedents and the threat of conflict escalation.

Works Cited


Thomas, Baylis. The dark side of Zionism: Israel’s Quest for Security through Dominance.
REVISITING NAFTA: THE GAP BETWEEN PREDICTION AND REALITY FOR MEXICO’S SMALL CORN FARMERS

Sergio Broholm

The recent history of Mexico’s small corn farmers is a sad one, told in terms of increased poverty and illustrated by countless statistics that predict the small farmer’s demise. It is also a story of trade liberalization between the United States and Latin America, particularly of its shortcomings when not handled correctly. Years of bad news stemming from the effects of NAFTA make Mexico’s corn farmers seem like a dying breed, and unfortunately the future does not look bright. Recent events, like the shocking rise in tortilla prices in 2007 and the dismantling of the last NAFTA barriers to agricultural trade in 2008, make their plight obvious. These events have led to large protests in the streets of Mexico City against the rising price of tortillas and a serious questioning of the current administration’s commitment to free-market economics (Malkin 2007). Such events may be precursors for small agriculturalists in other Latin American states like Colombia, Peru, the Dominican Republic, Guatemala, Nicaragua, El Salvador, Panama, Honduras, and Costa Rica; these are countries which have or are in the process of entering similar free trade agreements with the United States (International Trade Administration). These agreements, like NAFTA, encompass agricultural trade liberalization that will undoubtedly have a strong impact on these countries’ agricultural sectors.

Since NAFTA predates the next oldest free trade agreement between the United States and a Latin American country (Chile) by 10 years and because it is the most heavily-researched of these accords, studying the effects of NAFTA on Mexico’s most important agricultural population, corn producers, can serve as a test

Sergio Broholm is a 2010 graduate of Columbia University, where he majored in History and Economics-Political Science. He is interested in development economics and globalization, particularly as they relate to Latin America.
case for predicting the future of millions of Latin American farmers who face the prospect of unfettered agricultural trade with the United States. The relationship between NAFTA and Mexico’s corn sector is a good example of introducing free agricultural trade between economies of different sizes because corn is a crop of great economic and societal importance to Mexico, yet one in which the United States has a clear comparative advantage. The results of trade liberalization in this sector have affected areas ranging from food prices to immigration rates, making the issue of corn under NAFTA important on both sides of the border. With this in mind, this paper seeks to answer three questions: what were the effects of NAFTA on Mexico’s corn sector? How do these outcomes differ from the expected ones? And what have been the larger effects of these results?

Although several policies can be blamed for the increased hardships that face these corn farmers, many economists and political scientists trace the beginning of this decline to January 1, 1994, when NAFTA was implemented. Keeping in mind that the plight of the corn industry in Mexico cannot be exclusively related to trade liberalization, this paper will address the discrepancies between what economists and policymakers in the early 1990s publicly predicted would happen to this sizable segment of the labor force and what has actually occurred in the last fifteen years since its ratification. From there, this paper will discuss several effects that the agricultural policies of NAFTA have had on the corn sector and the implications this has on its role in Mexican society.

NAFTA’S PROponents AND PHILOSOPHY: THE 1994 NEGOTIATIONS

Beginning with a series of reforms in the mid 1980s, Mexico zealously embraced a neo-liberal development path during the presidency of Carlos Salinas de Gortari (1988-1994) that contrasted starkly with the import-substitution ideology that had dominated the country’s economic policies until the late 1970s. Under the direction of its Harvard-educated, technocratic president, the Mexi-
can government sought to integrate itself into the world economy by pursuing open-market policies, privatizing state-owned enterprises, and entering into liberal trade agreements. Most notable among these efforts was Mexico’s entrance into NAFTA in 1994.

By creating a free trade bloc between Mexico, the United States, and Canada, NAFTA integrated three very different economies and sparked heated debate in all three countries over the costs and benefits of such a sweeping trade agreement.

However, the different ways in which these debates took shape in Mexico and the United States foreshadowed the uneven way in which NAFTA affected Mexican society. On the American side, NAFTA was negotiated by two administrations (those of George H.W. Bush and Bill Clinton) and ratified by Congress, following the constitutional rules of treaty negotiation. The negotiation process was slow in the United States due to heated congressional deliberation, lobbying efforts by corporations and labor unions, and the contentious 1994 presidential election that focused, in part, on whether or not to accept or reject NAFTA. The same type of open deliberation was lacking in Mexico, where the decision to pursue and ultimately to ratify NAFTA was made by the Mexican president. Mexico’s political structure meant it was easy for the president to approve the measure while making it difficult for opposition interests to voice their concerns. The agreement had to pass the Mexican Senate, but at the time, President Salinas de Gortari’s party (PRI) held 61 of the 64 Senate seats. In 1994, Tim Golden of the New York Times observed that “President Carlos Salinas de Gortari’s overwhelming political control make approval a formality here” (Golden 1992).

The United States, Mexico, and Canada agreed on most sectors under NAFTA, but trade liberalization in the agricultural sector was divided into separate bilateral agreements due to the importance of maintaining a strong domestic agricultural sector. Frederick Mayer notes that “Canada found it so difficult to negotiate away its agricultural policies that it opted out of the [agricultural] deal altogether,” agreed upon by the United States and Mexico (Mayer 1998, 148). Separate agricultural agreements be-
tween Canada and Mexico, and Canada and the United States were established. Considering the comparatively small level of agricultural trade between Canada and Mexico under their separate agricultural agreement, Mexico’s corn policy has been shaped by a bilateral liberalization agreement between Mexico and the United States (Jannol et al 2003, 1).

This sweeping trade agreement between the United States and Mexico was an unprecedented integration of countries that were at very different levels of economic development. Nevertheless, the Mexican government “was eager to move quickly” with negotiations and proceeded to do so without the “similar consultative structure” of the United States that would have allowed potentially vulnerable sectors to voice their concerns (Mayer 1998, 115). As a result of “a Mexican political system that suppressed populist pressures and allowed for bold presidential action,” the concerns of small farmers and criticism from unions were shut out during negotiations (Mayer 1998, 339). Instead, Mexican negotiators worked closely with a few Mexican business elites who accompanied them to many conferences as the Coordinadora de organismos empresariales de comercio exterior (Committee for Foreign Trade Business Organizations, COECE), which became informally known as “el cuarto de al lado” (the room next door). This elite group was headed by Juan Gallardo, the president of a large soda bottling company, and included representatives of “agro industrial businesses” and the Mexican Council of Businessmen representing “the 50 most prominent businessmen” in Mexico. The hierarchical structure of the COECE allowed for a unified private sector interface with Mexican negotiators but, according to report from the Inter-American Development Bank, “prompted some criticism from groups not represented by COECE (trade unions, some farm groups, SMEs [small and medium enterprises] and academic analysts) that their views [were] excluded or given less weight” (Alba and Vega 2002). The policies designed to apply to small-holder corn farmers came out of this hurried consultative process, in which they were not directly represented.
WHO ARE MEXICAN CORN FARMERS?

In Mexico, agriculture is more than just an important sector of the economy; it has served as the cultural backbone of the country for centuries. Currently 15 percent of the Mexican labor force is employed in agriculture, a much higher percentage than in neighboring Canada or the United States (CIA World Factbook 2009). Because corn farming is practiced by a variety of producers across the nation it is particularly important to cultural identity in ways that other crops simply are not. Before analyzing the effects of NAFTA on Mexico’s corn sector, it is important to distinguish between different types of corn farmers because NAFTA has had heterogeneous effects on corn farmers depending on their size, market, and farming techniques. Large competitive corn producers constitute a numerical minority of Mexico’s corn farmers but were influential in the decision-making process of the agricultural agreement of NAFTA and “enjoy profit margins that enable them to withstand the pressure of price reductions and face the challenge of competition from NAFTA-related corn imports.” (Nadal 2000, 6). They are concentrated in the northwestern states and use modern farming techniques (Zahniser and Coyle 2004, 8-9). Small scale producers operate under less favorable conditions than large competitive corn producers and “can only withstand the pressure of foreign competition through exceptional efforts” (Nadal 2000, 7). Without viable alternatives, many of these producers have continued growing corn out of desperation and settle for decreasing profits. In 2000 subsistence corn farmers constituted 40 percent of all Mexican corn producers and used outdated farming techniques on marginal land (Nadal 2000, 8). In 2000, roughly 1,644,000 of subsistence farmers were ejidatarios, people who practice small-scale subsistence farming on semi-communal plots called ejidos (Zahniser and Coyle 2004, 6). Ejidos account for 62 percent of Mexico’s corn production even though a majority of these plots are smaller than 5 hectares in size (Nadal 2000, 44). The small level of output that these farmers produce is reflected in the fact that although so many Mexicans claim to be farmers, while agriculture
accounts for only 3.8 percent of the country’s GDP (CIA World Factbook 2009). Corn production is, therefore, not equivalent to market agriculture.

Of the millions of small-scale Mexican farmers, roughly three million derive their main source of livelihood from the cultivation of corn, which is indigenous to Mexico and a staple of the Mexican diet. These subsistence corn agriculturalists grow thousands of different corn varieties according to which type is best suited to their particular plot of land, which varies dramatically as one moves from the dry highlands of the Sierra Madre to the sloped tropical forests of the coastline. For these people, corn production is an intricate yet primitive process that is passed down through the generations, preventing standardized cultivation yet allowing for the optimal use of otherwise marginal land. As a result of the corn’s diversity and the farmers’ understanding of which type is best suited for the land, sixty percent of land under cultivation in Mexico in 2000 was devoted to the production of corn (Nadal 2000, 4). Small-scale farmers are diffuse throughout Mexico and as a group are powerless to organized agricultural interests able to influence treaty negotiations.

THE PROVISIONS FOR CORN UNDER NAFTA, AND ITS EFFECTS

In reaching the terms of trade under NAFTA, Mexican negotiators publicly sought a transition period in which trade liberalization in the corn market would be implemented slowly, affording the Mexican government time to help its small corn farmers adjust. On the other side of the table, American negotiators knew that the United States had a strong comparative advantage in most grain production, so they pushed for fast trade liberalization in that sector. The final agreement on corn between the two countries allowed the United States to export up to 2.5 million tons of corn duty-free in 1994, while increasing this permissible volume by three percent annually until 2008 (Rivera et al. 2009, 91). Concurrently, the Mexican government was allowed to charge heavy tariffs (initially
215 percent in 1994) on imported American corn that surpassed the duty-free quotas. These import duties had to decrease annually until reaching 0 percent in 2008, at which point the United States would have unfettered access to Mexican corn markets.

For Mexico’s corn farmers, historically among its poorest citizens, the fifteen-year transition period did little to make integration smooth and has left many even poorer, mainly because it was never really implemented. The percentage of rural Mexicans living in extreme poverty sharply increased from around 37 percent in 1994 to about 54 percent in 1998 before returning back to 1994 levels in 2002 (World Bank 2004, 171). Although 2004 data shows that this percentage has fallen below 1994 levels, this decrease might be attributed to the fact that hundreds of thousands of rural Mexicans have abandoned the countryside for the city and for the United States. In the early years after the ratification of NAFTA, corn prices fell predictably, hurting small farmers who grew enough corn to sell on the market but not enough to sell competitively. On top of this, the percentage of the rural population living in poverty reached 85 percent in 2001, encouraging farmers to leave their lands for cities or the United States. This has led to a disintegration of rural families that have experienced increased poverty and seen their members emigrate to the United States (Henriques and Patel 2004, 4). This dire situation has been compounded as the real price of tortilla, the traditional corn-based food of the poor, rose 279 percent between 1994 and 2004 (Henriques and Patel 2004, 6). This large increase does not take into account the sudden spike in tortilla prices that occurred in the first few months of 2007 (BBC News 2007). And prices have continued to rise as a result of growing demand for corn as a source for ethanol (Keleman and Rañó 2008) and the Mexican government’s 1996 decision to stop regulating tortilla prices (Nadal 2000, 38). Today, two years after the last tariffs on corn were lifted, Mexican corn farmers face higher tortilla prices as well as unfettered competition from American corn producers, who yielded in 2005 an average of 9.3

---

1 This statistic is of prices in Mexico City, where local officials take increased measures to control the price of corn.
tons of corn per hectare as opposed to Mexico’s 2.9 tons per hectare (Randall 2007).

EVALUATING CORN POLICY: THE VIEW FROM 1994 AND THE VIEW FROM 2010

A large part of the blame for the difficult situation of the corn farmers lies in agricultural trade liberalization under NAFTA. In deciding how corn policy would operate under NAFTA, it was incorrectly assumed (1) that the horticulture sector could absorb all of the losses in the corn industry, (2) that subsistence corn farmers were economically insulated from fluctuations in the price of corn, (3) that all corn-based products would fall in price, (4) and that the government would help poor farmers adjust to trade liberalization (Nadal 2000, 28). These shortcomings in vision, combined with unforeseen events like the boost in demand for corn-based ethanol, led to many discrepancies between pre-NAFTA predictions and reality.

The first factor that skewed the negotiators’ dealings in NAFTA talks was their belief in the unlimited expansion of Mexico’s horticulture and warm-weather crops market. Hoping to break into America’s citrus fruits, vegetable, and sugar markets, Mexican negotiators broke the agricultural stalemate and surprised many observers by making important concessions on corn in exchange for liberalizing horticulture markets in the United States (Mayer 1998, 137). The hypothesis was that competitive and intermediary producers would leave grain cultivation in favor of modernized production of tomatoes and other vegetables, while poorer farmers could find more productive jobs at these larger farms (Rivera et al. 2009, 90).

There were many problems with this prediction. First, Mexican exports already accounted for 60 percent of total horticultural imports to the United States, meaning there was not much room to grow (Nadal 2000, 23). Second, Mexico’s warm-weather crops have to compete with those of other Central American and Caribbean countries, many of which are naturally better suited for that kind of
agriculture. This is especially true in the banana market, where that fruit constitutes 29 percent of the fruits imported into NAFTA countries from the rest of the world (Stout et al. 2004, 48). Also, the comparative advantage of Mexico in warm-weather crops is not as lop-sided as America’s comparative advantage in grains. Warm-weather states like California, Texas, and Florida provide for a substantial amount of America’s demand for warm-weather crops. As a result, Mexican fruit exports to the U.S. only grew by 50 percent in the ten years after NAFTA’s ratification while Mexico’s corn imports from the U.S. tripled in the same time period (Carlsen 2004).\(^2\) The United States’ strength in horticulture crops can also be seen in the fact that Mexico’s horticultural exports to the United States fall in the summer months when the climate is warmer further north (Nadal 2000, 23). Finally, economists wrongly predicted that the expansion of the horticulture sector would correlate with a comparable expansion of jobs in that sector; that is, they assumed neutral technical change. The introduction of better technologies has made increases in output more related to capital growth than labor expansion (Nadal 2000, 21). As a result, jobs in the horticulture sector have not been as abundant as the government predicted. All of this shows that the horticulture and warm-weather crops sectors, which many in Salinas’s administration saw as the future of Mexican agriculture, were never going to be large enough to absorb displaced small-scale corn farmers who relied to some degree on being able to sell their corn. By 2000, Mexico had just about maximized its import market share in America’s horticulture market, which is also served by a strong domestic market (Nadal 2000, 23).

The second factor that NAFTA negotiators and economists incorrectly assumed was that subsistence corn farmers, who primarily live off their own production, would not be affected by lowered corn prices because they do not sell their produce and thus do not compete with American producers. Even today, the idea

\(^2\) Laura Carlsen, “Mexico after 10 years of NAFTA: The price of going to market” is based on presentations by the author at the Asian Regional Workshop on Bilateral Free Trade Agreements, held in Kuala Lumpur on 26-28 August 2004 and organised by the Third World Network.
of subsistence farmers as completely separated from the market is prevalent. In 2008, Marco Sifuentes, a spokesman for the agricultural department of Mexico reiterated, “our small producers are not affected by the free trade agreement... They don’t participate in the market” (McKinley 2008). This reflects the thoughts of economists in the nineties who supported NAFTA on the prediction that trade liberalization would help subsistence farmers by allowing them to take advantage of lowered prices for corn products (Rindermann et al. 1998, 32).

Reality did not match predictions because subsistence farmers are not isolated from the market and are in fact affected by price fluctuations in corn, even if most of their corn is for their own consumption. As Nadal points out, even though a rural family may be subsistent, it still needs monetary flow to buy products that it cannot grow, like medicines, and tools (Nadal 2000, 43). The growing dependence of subsistence farmers on other sources of liquidity can be seen in the fact that on average 44 percent of rural household incomes come from non-farm-wages. Meré shows that remittances received by rural Mexicans “are becoming a relevant part of household income, particularly for those who hold a smaller plot of land, where remittance may represent as much as 30 percent of the household’s income” (Meré 2007, 6). When strapped for cash, subsistence corn farmers are forced to sell petty amounts of their corn to cover their liquidity needs, forcing them to make up for the loss in food by expanding production (Nadal 2000, 36). As corn prices dropped over the last 15 years, subsistence farmers started planting more corn on their plots to make up for the increased amount of corn they had to sell to cover their liquidity needs. For example, Zahniser and Coyle (2004) say that “the average annual area harvested of corn during the NAFTA period is larger than the average for the 10 years immediately preceding the agreement (7.7 million hectares versus 7.0 million hectares)” (Zahniser and Coyle 2004, 6). Contrary to beliefs held by the government economists, subsistence corn farmers are tied to the corn market even if the majority of their crop is for consumption. This discrepancy helps account for the increase in corn production, especially among small-scale farm-
ers who do not have the capital or know-how to switch crops, cross the border, or adopt competitive farming techniques. For many subsistence corn farmers who cannot resort to these other measures, an increase in the production of their corn is the only way to independently meet their liquidity needs.

The third dimension that was overlooked during the negotiation of the trade agreement was that possibility that some corn-based product prices would drop with trade liberalization. Direct food use of corn accounts for about one-third of Mexico’s per capita calorie supply and is principally consumed in the form of tortillas (Zahniser and Coyle 2004, 9). This statistic does not highlight the fact that tortillas are primarily the carbohydrate of the poor, because middle and upper class Mexicans can afford to eat bread. The centuries-old Mexican flatbread was subsidized in the 1930s in order to keep the cost of this staple low and support many small-scale producers (Ross 1999). As late as 2007, street protests over increasing corn prices have demonstrated the centrality of corn in the Mexican diet and highlighted the importance of corn policy in Mexico (McKinley 2008). During NAFTA negotiations, it was thought that the losses in the corn sector would be cancelled by lower corn-stuff prices as Mexico and the United States specialized in their comparatively advantaged sectors. Assuming that tortilla prices would drop, the government phased out its roughly fifty year-old system of regulating tortilla prices called La Compañía Nacional de Subsistencias Populares (CONASUPO) (Ross 1999). Although corn prices did fall, the price of tortillas rose, leaving many to ask why.

The prediction that tortilla prices would fall with corn prices was based on assumptions that turned out to be wrong. Most importantly, it was assumed that tortilla prices were dependent only on corn prices. Officials did not take into account that the tortilla is segmented into various imperfect markets (dough, flour, machinery), of which corn serves only as one (though central) input. Although the processing of tortillas is fragmented into about 45,000 different tortilla producers, the tortilla flour market that supplies them is an oligopoly where the two largest companies—GIMSA
and MINSA—account for 70 percent and 27 percent of the market respectively (Henriques and Patel 2004, 6). NAFTA allowed these two companies to import American corn at a fraction of what it cost them to buy Mexican corn while the 1996 deregulation allowed them to artificially raise prices over the years. The fact that these large companies benefited from NAFTA’s agricultural policies is no surprise, considering the close connection between their CFOs and the Mexican government. Many of its leaders are members of the same organizations, like the Consejo Mexicano de Hombres de Negocios (Mexican Council of Businessmen) that comprised the cuarto de al lado of the Mexican NAFTA negotiating team (GRUMA). In an extreme example, the “king of the tortilla” and long-time owner of the world’s largest corn tortilla manufacturer, Roberto Gonzalez Barrera, has been a long-time friend of President Salinas de Gortari. With the dissolution of CONASUPO, he received all of the subsidized corn sales of the government and even furnished the ex-president with a private jet (Ross 1999). In keeping with neo-liberal policies, Mexican presidents since 1996 have rejected reinstating price controls to fight this speculation. Facing widespread protest in 2007, President Calderón held firm in his belief in free-market economics, saying “the government cannot decide the price of these products” while promising to “to protect those who have the least” by persecuting speculators and supplying low cost corn to the poorest of Mexican society (Wall 2007). The situation was aggravated in the mid-2000s as a real shock in corn prices caused by the new demand for ethanol raised tortilla prices even further. Although the problem has its roots in NAFTA, the unpredicted rise of tortilla prices is largely a product of faulty domestic policy in Mexico. Prices could be stabilized if the government broke the tortilla flour monopoly and were more effective in combating hoarding and speculation.

The fourth failure between pre-NAFTA rhetoric and outcomes relates to the supposed assistance that the Mexican government would provide to those farmers who would be most adversely affected by trade liberalization. In negotiating the corn provisions of NAFTA, the Mexican government reserved the right to charge
high tariffs on imported corn that surpassed a certain duty-free volume. As originally planned, this tariff would slowly reduce itself to zero by 2008, making the transition to free trade smooth. In combination with this plan, the Mexican government devised two policy instruments to assist producers during the transition period. The first, the Program for Direct Assistance in Agriculture (PROCAMPO), is a direct income subsidy created in 1994 that pays farmers the difference between a guaranteed floor price and the government’s target rate (Nadal 2000, 28). The other, Alianza para el campo (alliance for the countryside), is a “matching funds” subsidy that helps farmers (generally those most able to switch to other crops) invest in advanced irrigation systems, high yield crops, and farming tools, the goal being to increase the competitiveness of Mexican farmers (Suvedi 2000).

PROGRAM FAILURE: THE RESULTS AND RESPONSE TO NAFTA

The Mexican government’s plans to ease transition pains from NAFTA, which sound good on paper, failed in many ways, largely due to budgetary problems. The first major failing was that the fifteen-year transition period was truncated into a span of thirty months. Although Mexico was importing American corn beyond the duty-free amount since the agreement was put into place, Mexico never imposed the transitional tariffs that were supposed to start at 215 percent. This was largely due to an already insufficient domestic supply of corn and the 1994 Mexican peso crisis, which saw a sudden devaluation of the Mexican peso and required additional corn to maintain corn-based food prices from spiraling too high. Furthermore, instead of reducing the tariffs over the fifteen-year period, the government reduced them in thirty months due to inflation worries, pressure from powerful corn processors, and disorganized control mechanisms at the border. As a result of this bad implementation, corn prices plummeted 48 percent between 1994 and 1996 while the government failed to collect $2 billion in tariffs that it had a right to claim (Henriques and Patel 2004, 6).
These $2 billion of forgone revenue would have been useful in implementing Mexico’s new farming policies. PROCAMPO payments to farmers, which initially stood at around $100 per hectare, did not keep up with inflation and as of 2000 provided less than $62 per hectare in real terms. Since small farmers have become dependant on PROCAMPO to cover for the increasing price of farming inputs like fertilizer, pesticides, and seeds, many PROCAMPO payments are made directly to the suppliers of these inputs who in turn supply the farmers with fewer and fewer supplies (Nadal 2000, 29). The growing difference between PROCAMPO payments and input prices has essentially made these payments welfare checks for poor farmers instead of starting capital for creating more advanced farming techniques. Alianza para el campo met a similar fate in that it started with a measly operating budget of about 200 million U.S. dollars that has been declining in real terms ever since the program began (Nadal 2000, 30). Suvedi has also seen that this program tends to subsidize those producers most able to move to more intensive production, instead of those in greatest need of assistance. Because of this, only about a fourth of those agriculturists receiving Alianza para el campo subsidies are subsistence farmers (Suvedi 2000, 15). These programs, which the administration hoped would create a smooth and fair transition to free trade, were not properly implemented and did little to achieve their purposes.

AN INDEPENDENT CAUSE? NAFTA, CORN POLICY, AND SOCIAL OUTCOMES

NAFTA’s effect on small corn farmers cannot be evaluated in a vacuum. Such a project requires an understanding of other factors that affected this sector, such as political structure, geography, technological advances, and relations with the outside world. Corn producers—both subsistence farmers and those producing for the market—saw their lives changed by NAFTA policies. These changes within the corn sector also affected how the corn sector related to Mexico. This broader question requires an understanding
of NAFTA’s exclusive role in changing the corn market. The paper now examines how the effects of NAFTA on the corn sector have made themselves felt in other aspects of Mexican society, namely immigration, the environment, and domestic spending.

Many of the adverse effects of NAFTA on Mexico’s corn industry correlate with certain domestic policies, changes in demand for corn, and competitiveness of other countries. In many cases, the discrepancies created by NAFTA would have turned out differently if it were not for these extra conditions. At the same time, certain Mexican policies of the 1990s were planned specifically to accompany NAFTA, and cannot be separated from the effects of the agreement. As Zepeda, Wise, and Gallagher argue, it is incorrect to assume the “no-NAFTA scenario would entail Mexico following the very same policies it has since the 1980s, only without NAFTA” (Zepeda et al. 2009. 3). NAFTA is intrinsically entwined with such policies as tighter monetary policy and deregulation of the corn industry. Thus it is fruitless to analyze narrowly the effects of the treaty without considering NAFTA as part of Mexico’s liberalization attempt of the mid 1980s and onwards. Keeping this in mind, when one controls for additional factors, for example, the devaluation crisis of 1994, the growing competitiveness of China, and rise of ethanol, NAFTA still contributed to the demise of the Mexican corn sector by setting in motion a series of policies that ignored the public sector’s role in promoting growth and defending at-risk industries.

Trade liberalization through NAFTA was a conscious choice made by the Mexican government; it was not a natural progression, nor was it the only available path. In its effort to modernize, Mexico chose to adhere strictly to the neo-liberal guidelines of the Washington Consensus, which called for financial deregulation, privatization, strict property rights, and arm’s length governmental-business relations. Although neo-liberal economists advocated this type of plan, India, China, and Brazil achieved more success by following development paths that strayed from this neo-liberal recipe and molded to the specific needs of each country. With this in mind, it makes little sense to argue that the situation would nec-
necessarily have been worse had NAFTA not been passed. Economist Dani Rodrik makes the argument that “neo-classical economic analysis is a lot more flexible than its practitioners in the policy domain have generally given it credit for” (Rodrik 2007, 15). In other words, overarching economic principles like liberalization, open markets, and property rights should guide policy but should not be translated into specific policy. In ratifying NAFTA, Mexico followed a theoretical recipe for success closely without seriously factoring in its special circumstances. Examples of strict neoliberal policies included the dissolution of tortilla price controls under CONASUPO, the reprivatization of the banks in 1985, and the forced dissolution of “illegal” worker’s unions (Peters 1998, 355-356). NAFTA’s passage, philosophically related to such policies, had direct harmful consequences for the Mexican corn industry. These developments in the corn industry may affect Mexicans outside that particular sector.

With this background in mind, this paper turns to broader effects of NAFTA’s agricultural policies, particularly on the environment and land use. The first noticeable effect that can be attributed to NAFTA has been on the environment. In his 1992 book that laid out the Mexican government’s objectives for NAFTA, Luis Téllez, the then Undersecretary of Planning at the Secretariat of Agriculture for President Salinas de Gortari, predicted that reductions in the output of corn would set aside marginal land and relieve it of over-tillage (Téllez 1992). This prediction turned out to be incorrect, a misunderstanding that, encoded in NAFTA, had clear consequences. Nadal concludes that the “extension of the agricultural frontier to marginal lands”, and the “deterioration of the social institutions that enable communities to maintain adequate soil conservation practices” have enhanced soil erosion problems in Mexico (Nadal 2000, 81). Increased corn cultivation as a result of falling prices has led to the disintegration of the soil in many traditional farming states like Oaxaca and Chiapas, where corn output increased 48 percent and 51 percent respectively between 1991 and 1997 (Nadal 2000, 47). The problem lies in the fact that most of the poorest farmers, who have increased corn production to cope
with lower returns for their crops, farm on sloped plots of land that have low-quality soil. At the same time, other factors help explain the disintegration of subsistence farming units and the over-use of marginal lands.

The privatization of the ejido system in 1992, allowed ejidatarios to sell their land to private parties and ended the government’s obligation to redistribute land according to peasants’ petitions (Cornelius and Myhre 1998, 2). A four-year research project established by the Center for U.S.-Mexican Studies on the consequences of the 1992 agricultural restructuring found that although changes are not happening as drastically as predicted, the privatization of the ejidos was creating a “gradual and regionally varied—though still far-reaching-process of rural transformation,” by consolidating ejidos in the hands of private land-holders, stimulating emigration to the United States, and favoring “commercially lucrative ejidos with the greatest potential for expanding export performance” (Cornelius and Myhre 1998, 19, 16).

The privatization of the ejido system in 1992 allowed ejidatarios to sell their land to private parties, splitting many farming communities that used to practice communal farming techniques. In addition, technological changes, namely the growing availability of chemical fertilizers, allow for more rigorous use of soil that needs to lay fallow to ease the erosion process. Corn yields in Mexico rose to 2.9 tons per hectares in 2005 as opposed to 2.0 in 2000, a testament to the effectiveness of new technologies but a bad omen for the over-tilled land of Mexico’s countryside (Randall 2007).

Even if these factors were independent of NAFTA, Nadal shows that trade liberalization under NAFTA “may enhance these negative trends… through more intensive use of soils.” The change in corn prices has led to a serious intensification of corn production among those producers who rely exclusively on corn and farm on marginal soil that is especially susceptible to irreversible erosion damage. By 1997 roughly two-thirds of Mexico’s agricultural land met that description (Nadal 2000, 81).

The land has also suffered as a result of emigration and the disintegration of the social institutions and networks within old
ejido communities. Since the early 1990s, rural people have fled the Mexican countryside to the cities and the United States at an astonishing rate. The percentage of labor devoted to agriculture shrank from 25 percent in 1994 to 15 percent in 2009 (CIA World Factbook 1994, 2009). As a result, community farming practices like grazing livestock on communal grounds and curating seed types, are becoming more rare. Animals graze on private farmland instead, uprooting plants that keep the soil healthy. In various communities, local corn growing practices are being lost now that younger generations are leaving the farms (Nadal 2000, 87). Looking beyond NAFTA as a cause, some attribute this trend of increased migration to more long-term changes like the growing familial, social, and economic ties between Mexico and the United States (Massey and Espinosa 1997, 989-990). Zúñiga and Molina argue that the high population growth since the mid 20th century, in combination with increased female labor force participation, led to the high emigration rates from Mexico that started in the 1970s (Zúñiga and Molina 2008, 3). But simple demographics do not explain the sudden increase in emigration from rural Mexico in the 1990s. In connection with a 2.3 million drop in total employment between 1990 and 2008 (largely due to agricultural employment loss), Mexicans have been fleeing the countryside (Zepeda et al., 2009, 13; Meré 2007, 12). In connection with the 2.3 million agricultural jobs that were lost between 1990 and 2008, rural migration to the United States grew 452 percent from 1980 to 2002, a large change considering that rural migration to the United States only grew 92 percent from 1980 to 1994 (Meré 2007, 3). This migration trend runs opposite the intentions of the trade agreement. During a campaign to sell NAFTA to an American audience, President Salinas de Gortari plainly stated in an NBC News interview: “we want to export goods, not people” (NBC Nightly News 1990 qtd. in Mayer 1998, 45). Trying to explain this migration, Philip Martin, an agricultural economist, blames the Mexican government’s handling of NAFTA provisions for rendering tens of thousands of Mexican farmers useless and forcing them to “export themselves” to the United States (Uchitelle 2007). Martin mainly blames the lack
of transition from corn to alternative crops for the unexpected increase in migration: “We understood that the transition from corn to strawberries would not be smooth…but we did not think there would be almost no transition” (Martin qtd. in Uchitelle 2007). This astonishing increase occurred in spite of the rising militarization of the U.S. border and plummeting birth rates, which would suggest a decline in emigration.

Another important effect on the rural sector that can be traced directly to NAFTA is the decrease in public spending on the agricultural sector. Although foreign direct investment (FDI) more than tripled between 1992 and 2006, overall investment rates in Mexico have remained remarkably stagnant at about 20 percent of GDP (Zepeda et al. 2009, 7). Additionally, less than 1 percent of that FDI went into the agricultural sector (Henriques and Patel 2004, 2). Hoping that NAFTA would generate large sums of private investment from home and abroad, the Mexican government simultaneously reduced its government spending on agriculture by reducing corn subsidies, eliminating corn price guarantees and dismantling CONASUPO. The government never provided more than meager budgets to PROCAMPO and Alianza para el campo. Proponents feared that high levels of government spending would “crowd out” private investment, but this did not turn out to be the case. Looking beyond levels of FDI in Mexico (which tripled since NAFTA) Zepeda, Wise, and Gallagher reveal that total levels of investment (foreign plus domestic) are extremely low in comparison to that of rapidly industrializing countries, like China, and even in comparison to previous levels of total investment in Mexico. This lack of domestic private investment is a result of pressure from imports and a lack of government spending that should provide infrastructure and credit to Mexican firms (Zepeda et al. 2009, 7). NAFTA did attract a lot of American private investment (conglomerates like Cargill and Archer Daniels Midland own a large portion of Mexico’s competitive agriculture), but the lack of comparable domestic government spending on necessary infrastructure like irrigation, roads, and housing has stunted economic growth (Uchitelle 2007). From this, one can see that NAFTA has
made it hard to translate increased FDI into economic growth, especially in the agricultural sector.

CONCLUSION

At the outset of negotiating NAFTA, poor farmers using outdated techniques on marginal plots of land fit badly with the Mexican administration’s vision of modern, export-oriented agricultural production, or at least with the policies that came out of that vision. What was created proved to be unworkable: The transition period was slashed, the relief programs’ budgets were never increased, and corn producers were excluded from the negotiation process.

“In spite of calls from economists and opponents to rework the provisions of the trade agreement, the Calderón administration has not taken any significant steps toward supporting those agrarians hurt by NAFTA” (Zepeda et al. 2009, 18). Calderón has refused to consider a re-negotiation of NAFTA, stating that “the free trade agreement, negotiated almost 15 years ago, has its pros and cons, but overall it has benefited the country.” Instead, he favors instituting domestic policies that will alleviate the pressure felt by these farmers (Jacinto 2008). In its latest informe de gobierno (Mexico’s equivalent of a State of the Union), the Calderón administration has countered claims of agricultural neglect by pointing to the roughly $1.5 billion increase in PROCAMPO’s budget and stating that new measures have been put in place to ensure a more even distribution of these payments to the smallest of producers (Economía Competitiva y Generadora de Empleo). Still, Calderón faces much opposition from groups representing small farmers who claim that he has abandoned the Mexican farmer (Sánchez 2009).

In light of continued resistance, displaced corn farmers and a growing percent of the general citizenry are voicing their frustration with NAFTA. The 2006 presidential hopeful, Andres Manuel López Obrador, ran on a platform of not honoring Mexico’s commitment under NAFTA to eliminated tariffs on U.S. corn and beans, and only narrowly lost the election (Stevenson 2006). Large
numbers of people have taken to the streets on numerous occasions, but the political structure of the government has prevented any significant changes to the corn provisions of the treaty or the accompanying domestic policies.

In 2010, NAFTA’s failure to improve the conditions for small-holder corn farmers points to one central question—one for which any attempt to answer is outside the scope of this paper. Why have so many of these farmers continued to produce corn rather than change to more lucrative crops, as the Mexican government foretold? In one way, this is a question specific to Mexico: the Mexican government’s mischaracterization of the corn sector has had serious internal consequences. This paper argues that the Mexican government did not move quickly to address NAFTA’s negative effects on corn farmers, and an examination of the original policy suggests that many of those effects could have been avoided. The content of Mexico’s corn policy, above all the assumptions made about subsistence farmers’ relation to the market, is instructive beyond the country’s borders. The method and philosophy of the treaty raises pressing questions for similar trade deals.

Although some might dismiss the NAFTA as an agreement of the past, the consequences of the agreement are still being felt today, especially now that the last and most important tariff barriers have been terminated. With the dramatic increase in tortilla prices, the spread of rural poverty, and the development of corn as an energy source, the effects of free agricultural trade between Mexico and the United States are current events, contemporary with new trade liberalization between countries of very different economic development, and a reminder that small agriculture can make a large statement.

Works Cited


CIA World Factbook 2009.”Mexico” https://www.cia.gov/library/publications/the-


Keleman, Ilder and Hugo García Rañó. “Biofuels and Rising Food Prices: Mexico’s 2007 Tortilla Crisis,” Food Security and Environmental Change Conference April 2-4, 2008 Oxford University, UK


Rindermann, Rita, Manuel Gómez Cruz, and Gary Williams, TLC y agricultura:¿Funciona el experimento? (Chapingo, Mexico: Centro de


