Three Essays on Regulation and Economic Growth

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

James W. Broughel
Master of Arts
Hunter College of the City University of New York, 2011
Bachelor of Arts
Hunter College of the City University of New York, 2011

Director: Tyler Cowen, Professor
Department of Economics

Spring Semester 2017
George Mason University
Fairfax, VA

Dedication

For Anna and Ellie.

Acknowledgments

I would like to thank Tyler Cowen, Garett Jones, and Jerry Ellig for their responsiveness and insightful comments on this dissertation. I could not have asked for a more attentive dissertation committee. I am grateful to my wife, Anna, for her support throughout my doctoral studies. I would not have finished so quickly were it not for her encouragement. Finally, I thank my colleagues at the Mercatus Center for their feedback and support on these articles.

Table of Contents

List of Tables

List of Figures

Figure **Page**

List of Equations

List of Acronyms

AIC…....…………………………………………………….Akaike Information Criteria
BCA………………………………………………………………Benefit-Cost Analysis
BIC……………………………………………………….Bayesian Information Criteria
CFR……………………………………………………..Code of Federal Regulations
CPI…………………………………………..……………Consumer Price Index
DF………………………………………………………………………Dickey-Fuller
FEVD…….……………………..…………...Forecast Error Variance Decomposition
FR……………………………………...…………………………Federal Register
GDP…………………………………………………………Gross Domestic Product
GPT……………………………………………………….General Purpose Technology
IRF………………………………………….……………Impulse Response Function
OECD……………………......Organization for Economic Co-operation and Development
OLS…………………………………………………………….…Ordinary Least Squares
OMB…..…………………………………………………Office of Management and Budget
PCE………………………………………….….Personal Consumption Expenditures
PPI……………………………………………………………....Producer Price Index
RCK……………………………………………………...Ramsey-Cass-Koopmans
SCC…………………………………………..…………..Social Cost of Carbon
SDR…………………………………….………………….Social Discount Rate
SVAR…...……………………………………………….Structural Vector Autoregression
TVM..………………………………………………………………Time Value of Money
VAR……………………………………………..………….Vector Autoregression

Abstract

THREE ESSAYS ON REGULATION AND ECONOMIC GROWTH

James W. Broughel, Ph.D.

George Mason University 2017

Dissertation Director: Dr. Tyler Cowen

This dissertation explores the relationship between government-issued regulations and economic growth. The first essay surveys the models of economic growth most commonly used by economists to produce a theoretical framework for understanding the growth impacts of regulations. Beginning from a technology-augmented Solow model, a system is presented for classifying economic growth impacts, and the channels by which regulations enter the economic system are described. Next, a more comprehensive set of economic growth models is surveyed, with the role of regulation highlighted in each model. The essay concludes by discussing some remaining unsolved puzzles in growth theory, including the role of institutions and population as contributors to economic growth.

The second essay explores how growth theory has influenced the economic analysis of individual regulations, arguing that growth theory has contributed to confusion in regulatory benefit-cost analysis with respect to the purpose of the social discount rate. Economists discount future benefit and cost flows for a variety of reasons,

including time preference, diminishing marginal utility of consumption, opportunity cost of capital, and risk aversion. Many of these rationales for discounting can be explained using the Ramsey equation found in neoclassical growth theory. This essay argues that Ramsey approaches to discounting are problematic for use in regulatory benefit-cost analysis (BCA) because they are inconsistent with certain foundational principles of BCA. It is argued that a more useful discounting framework is one that is based on the time value of money, where discounting is used as a way to compare investment projects to a baseline alternative investment. A social discount rate (SDR) used in this manner avoids ethics controversies that arise in Ramsey discounting approaches with respect to giving preferential treatment to the present generation over future generations, while still recognizing and accounting for the importance of economic growth.

The final essay uses time series econometric techniques to explore how federal regulation predicts changes in aggregate measures of the price level. The approach distinguishes between the long-run effects of the cumulative stock of all federal regulations and the short-run impacts of the flow of new regulations finalized each year. Cointegration estimates show a significant long-run statistical relationship between the stock of federal regulation and three measures of the price level. Similarly, output from structural vector autoregressions show that short-run shocks to the flow of federal regulation predict significant changes in the same three price indices. The essay concludes with discussion of the mechanisms by which regulation and the price level could be linked, concluding that regulations may be an important source of supply-side technology shock.

1 Regulation and Economic Growth: Applying Economic Theory to Public Policy

1.1 Introduction

This essay presents a framework for assessing the economic growth implications of public policies, paying particular attention to the effects of government regulations. The framework is intended to be a theoretical contribution to the field of regulatory economics, surveying the landscape of economic growth models and highlighting lessons from the models for regulatory policy. With a stronger theoretical foundation in place, regulators may be able to gain new insights into how regulations affect national income and, by extension, other important indicators of human well-being.

When economists talk about *economic growth*, they are referring to changes in national income. Typically, such changes are evaluated using measures of a country's GDP, defined as the market value of final goods and services produced inside a country's borders in a single year. GDP is a measure of the value of a nation's annual output, as well as national income. Economic growth is typically measured in changes in *real* GDP, where *real* reflects that adjustments are made to account for a changing price level over time.

GDP on a per capita basis (meaning GDP divided by the population of the country) is a reasonable approximation of a nation's standard of living, just as personal income is a reasonable measure of an individual's standard of living. GDP per person is correlated with many important indicators of well-being, such as life expectancy, and negatively correlated with characteristics that countries want to avoid, such as child mortality or corruption. There are some well-known limitations to using GDP as a measure of living standards. For example, GDP misses activity not traded in markets, such as unpaid housework or the value that people derive from leisure time. However, as a measure of a nation's annual income, GDP is reasonably accurate. Income is not an all-encompassing measure of human well-being, of course, but income is used to purchase the goods and services that matter most to people's health, happiness, and quality of life. Therefore, income is an important measure of well-being, even if it is not the only measure that matters.

Regulations—the other focus of this essay—are restrictions on human behavior. Restrictions may not be legal in nature. For example, professional baseball teams follow regulations that govern how their game is played. These rules are set by the Major League Baseball organization. Here, the focus is on regulations of the legal variety that are promulgated by government agencies. Regulations, as written by administrative agencies, are distinguished from laws written by legislatures, which consist of elected representatives of the people. Administrative agencies employ public officials who, more often than not, are career public servants. These officials are delegated law-making authority from legislatures. Furthermore, regulations are unique in that—unlike taxes and spending—the vast majority of their effects are not captured in government budgets. In this sense, the effects of regulation are largely invisible to the public.

Although the focus of this essay is on the effects of regulation on economic growth, there is little reason to think that regulations written by regulatory bodies are fundamentally different from laws written by legislatures or from other public policies. Therefore, this essay should be useful to regulators, students interested in the economic effects of regulation, and anyone interested in the growth implications of public policies in general.

1.2 The Fundamentals of Economic Growth

1.2.1 The Importance of Growth

Perhaps the most powerful lesson from economic growth theory is that small changes in output today can lead to enormous changes in living standards when those changes compound over time. This result led Nobel laureate Robert Lucas (1988, 5) to comment that "the consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about [growth], it is hard to think about anything else."

The choices we make today, for better or worse, can have huge implications for the welfare of future generations. If we care about the well-being of future generations, by extension, we must care about economic growth. Seemingly minor mistakes or successes in public policy can have ripple effects that compound over time and change the course of history. As a result, those who set economic policy, such as elected officials and regulators, have a duty to be informed about the responsibility that comes with their power.

As will be shown in this essay, not all causes of economic growth are known to economists. For example, some nontrivial component of growth appears to be an unintended consequence of human social interaction. Even today, important factors that contribute to economic growth are debated or remain a mystery. This ambiguousness may leave some students of economic growth frustrated, and it means that to some extent policymakers must act under a great amount of uncertainty. But such is the current situation.

All is not lost, however. Economists do know enough to provide some fairly strong general guidelines for policymakers. The guidelines can assist regulators who seek to boost or—perhaps more important—avoid stifling economic growth. Regulators should keep these guidelines in mind as they balance the political demands of the immediate moment with the long-term interests of a nation and future generations.

To understand the power of growth rate changes, Table 1 presents hypothetical growth paths for an economy. Beginning in year 0—the present—this imaginary economy has a level of income per capita of $100. After just five years, a country whose per capita income is growing by 3 percent per year will enjoy a standard of living 10 percent higher than one that begins at the same level of income per capita but grows at just 1 percent. After 25 years, living standards are more than 60 percent higher in the country whose economy grows by 3 percent annually. And after 50 years, living standards are more than two and a half times higher than the economy that grows by 1 percent.

Table 1: Real GDP Per Capita at Different Growth Rates
Source: Author's calculations.

	Growth rates			
Year	1%	3%	7%	10%
0	$100	$100	$100	$100
1	$101	$103	$107	$110
5	$105	$116	$140	$161
10	$110	$134	$197	$259
25	$128	$209	$543	$1,083
50	$164	$438	$2,946	$11,739
75	$211	$918	$15,988	$127,190
100	$270	$1,922	$86,772	$1,378,061

When the growth rate rises to 7 or 10 percent, these changes become even more astounding. A country whose economy grows by 7 percent per year in per capita terms will double incomes in just over a decade. A comparable change takes about 25 years when growing by 3 percent annually, and it takes almost 75 years when growing by 1 percent annually. Extending these rates far into the future, it can easily be seen that the implications for future generations are enormous. Speeding up the annual rate of economic growth by a single percentage point or two can change future living standards by orders of magnitude.

Table 1 presents a hypothetical example, of course, but now consider what has happened in the real world in recent decades. Table 2 presents annualized growth rates for 55 countries for the years 1950–2014 (Feenstra, Inklaar, and Timmer 2015). Levels of

GDP per capita at the beginning and end of the series are presented at purchasing power parity, as are annualized per capita growth rates over this period. In comparing income and growth rates across countries, it is critical to make comparisons at purchasing power parity. Because countries typically evaluate their GDP using domestic currency, all GDPs must first be converted to a common metric, such as 2011 US dollars. Next, because one dollar may buy more in some countries (e.g., India) than others (e.g., Switzerland), adjustments must be made for the different price levels across countries.

Table 2: Annualized Growth Rates in Real Income Per Capita, 1950–2014
Source: Author’s calculations based on Feenstra, Inklaar, and Timmer 2015.

Country	Real GDP per capita 1950 at purchasing power parity (2011 US$)	Real GDP per capita 2014 at purchasing power parity (2011 US$)	Annualized growth rate (%)
Egypt	604	9,909	4.5
Japan	2,616	35,358	4.2
Thailand	1,072	13,967	4.1
El Salvador	673	7,843	3.9
Portugal	2,727	28,476	3.7
Cyprus	2,784	28,602	3.7
Germany	4,714	45,961	3.6
Spain	3,521	33,864	3.6
Ireland	5,126	48,767	3.6
Panama	2,152	19,702	3.5
Austria	5,340	47,744	3.5
Brazil	1,673	14,871	3.5
Italy	4,335	35,807	3.4
Luxembourg	12,083	95,176	3.3
Trinidad and Tobago	4,111	31,196	3.2
Norway	8,890	64,274	3.1
Argentina	2,890	20,222	3.1

Finland	5,961	40,401	3.0
Turkey	3,054	19,236	2.9
India	842	5,224	2.9
Netherlands	7,634	47,240	2.9
France	7,057	39,374	2.7
Morocco	1,312	7,163	2.7
Belgium	8,087	43,668	2.7
Peru	2,057	10,993	2.7
Ecuador	2,052	10,968	2.7
Israel	6,267	33,270	2.6
Iceland	8,354	42,876	2.6
Denmark	9,473	44,924	2.5
Philippines	1,424	6,659	2.4
Sweden	10,002	44,598	2.4
Costa Rica	3,223	14,186	2.3
United Kingdom	9,263	40,242	2.3
Switzerland	13,960	58,469	2.3
Colombia	3,179	12,599	2.2
Ethiopia	336	1,323	2.2
Mauritius	4,665	17,942	2.1
Canada	11,248	42,352	2.1
Sri Lanka	2,765	10,342	2.1
Bolivia (Plurinational State of)	1,661	6,013	2.0
Mexico	4,422	15,853	2.0
United States	14,655	52,292	2.0
Pakistan	1,333	4,646	2.0
Uruguay	6,259	20,396	1.9
Australia	13,310	43,071	1.9
Guatemala	2,374	6,851	1.7
New Zealand	12,402	34,735	1.6
Venezuela (Bolivarian Republic of)	5,862	14,134	1.4
South Africa	5,337	12,128	1.3
Uganda	854	1,839	1.2
Nigeria	2,623	5,501	1.2
Honduras	2,207	4,424	1.1
Kenya	1,590	2,769	0.9
Nicaragua	3,404	4,453	0.4

Democratic Republic of the Congo	1,839	1,217	-0.6

The fastest-growing countries in the group of 55 from 1950 to 2014 were Egypt and Japan, which grew at an annualized rate of 4.5 and 4.2 percent respectively. This rate is quite impressive given that Japan experienced fairly slow growth in the past two decades and is a testament to how fast Japan grew early in the sample period. Rapid growth in Japan led the per capita income to increase more than tenfold, from more than $2,600 in income per capita in 1950 to more than $35,000 in 2014 (2011 US$). Such a result is amazing in its own right, but it becomes even more impressive when contrasted with countries that were not nearly so fortunate. For example, in the Democratic Republic of the Congo, income per capita shrank during this period, with residents worse off by this measure in 2011 than their grandparents had been 60 years earlier. The statistics do not fully account for some technological advancements, so the number likely underestimates improvements in living standards. Nonetheless, the stakes involved surrounding issues of economic growth are clear.

Some countries in Table 2 grew faster in the latter half of the 20th century in part because they had to rebuild following World War II. Many European countries fall into this category. As will be seen in the review of the Solow model in chapter 2, a country can grow fast simply by destroying its capital stock. Such growth is not a good strategy for improving people's well-being, however, because it means initially lowering the *level* of income per capita. Both the level and the rate of economic growth are important for living standards.

Figure 1 shows the distribution of per capita growth rates across the countries listed in Table 2. For the years 1950–2014, most countries experienced 1.0–3.9 percent growth per year. The US grew at about 2 percent per annum during these years.

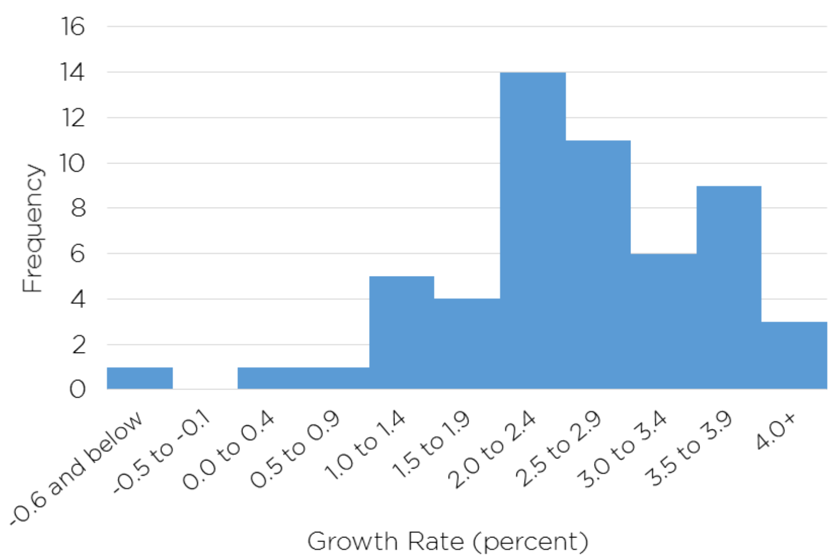


Figure 1: Annualized Growth Rates in Real GDP Per Capita for 55 Countries, 1950–2014
Source: Feenstra, Inklaar, and Timmer 2015.

A natural question to ask when looking at Figure 1 is the following: Given that economic growth in the range of 3–4 percent per year is clearly possible, what can policymakers do to help achieve and maintain such a rate? A central purpose of this essay is to shed light on this important question, with a particular emphasis on the role that regulatory policy can play as a promoter or an inhibitor of economic growth.

In the sections that follow, this essay will first explore one of the most widely-used growth models in economics: the Solow model. This model should look familiar to most students who have taken an undergraduate macroeconomics course; it is a workhorse of modern macroeconomics. The essay goes on to present a classification scheme to better understand the different outcomes that are possible following shocks to variables in the Solow model. With a classification scheme in place, newer growth models are reviewed, highlighting how regulation can affect the key variables in each model and, by extension, affect the growth path of the economy. Some of the remaining unsolved puzzles in growth theory, specifically those related to the roles of institutions and populations as contributors to growth, will also be discussed. Armed with this framework, regulators should be better able to achieve their goals while keeping this nation's economy on a healthy, sustainable growth path.

1.2.2 The Solow Model

To understand how individual regulations or groups of regulations affect economic growth, a model of economic growth is first needed. Models are necessary to make sense out of the complexity of the real world. Models simplify the world, thereby allowing better understanding of the forces that shape reality. The logical model to start with is the most famous of all growth models: the Solow model. The model was developed by American economist Robert Solow (1956) and Australian economist Trevor Swan (1956) in the mid-20th century. This essay uses a technology-augmented variant of the model, similar to versions in Charles I. Jones (2001), and David Romer (2011). Technology augmented means that technology is an input in the production process that works by increasing, or augmenting, the productivity of labor.

Throughout this essay, variations are used of the famous Cobb–Douglas production function. That production follows a Cobb–Douglas form is a common assumption in many economic growth models. The function, developed by mathematicians Charles Cobb and Paul Douglas (1928), takes a form such as $Y = K^{\alpha}L^{1-\alpha}$, where Y represents the total output of the production process, and K and L represent capital and labor, respectively, which are the main inputs into the production process. The parameter α is the output elasticity of capital. It explains how output changes as the amount of capital used in production changes. Under conditions of perfect competition, α also represents the fraction of total output that is paid to capital.

Cobb–Douglas production functions are widely used in part because they capture very important real-world phenomena. For example, when $\alpha < 1$, there are diminishing returns to capital and labor. This characteristic means that, as an economy adds more capital or labor to the production process, the additional output generated from each additional unit of input diminishes. This assumption is widely believed to be a realistic portrayal of actual production processes. For example, the first tractor put to work on a farm probably increases daily output by a substantial amount, but the fourth, fifth, or sixth tractor might not be of much use at all. This example highlights the phenomenon of diminishing marginal returns.

By allowing the exponents on capital and labor to sum to 1, the production function exhibits constant returns to all factors of production (even while there are diminishing returns to individual factors). In other words, if the level of capital alone is doubled, output less than doubles (i.e., diminishing returns to capital), but if all inputs together are doubled (in this case, both capital and labor), output exactly doubles. Another convenient property of Cobb–Douglas production functions is that the elasticity of substitution between capital and labor is exactly 1, which means that a rise in the relative price of capital or labor leads to an equivalent decline (in percentage terms) in the relative demand for the input.

All of these assumptions can be relaxed, of course, and changes in assumptions about the production process will have important implications for how government regulation changes output in any given model. To begin, however, things are kept simple, assuming that production is explained by the equation

$$\boldsymbol{Y = f(K, AL) = K_t{}^{\alpha}(A_t L_t)^{1-\alpha}, 0 < \alpha < 1,} \tag{1}$$

where Y is the total output, K is the amount of physical capital in the economy, A is an index of the state of labor-augmenting technology, and L is the number of people employed in the labor force. The parameter α is capital's share of output, given that this model assumes perfect competition. Because both α and $1 - \alpha$ are less than 1, but together sum to 1, this model exhibits diminishing returns to scale in the input factors K, A, and L and constant returns to scale in all factors of production.

The levels of A and L at any given time t are explained by the equations

$$A_t = A_0 e^{gt} \tag{2}$$

and

$$L_t = L_0 e^{nt}, \tag{3}$$

such that A and L grow at the constant rates g and n, respectively, and begin from the levels A_0 and L_0.

It turns out that economic growth can be defined in two ways: *intensive form* (i.e., changes in output per unit of some input, such as labor) and *extensive form* (changes in total output). Quarterly releases of GDP growth by the Bureau of Economic Analysis relate to extensive growth. These are the numbers that regularly appear in newspapers. Economists typically work with the Solow model in intensive form, however, meaning variables are evaluated per unit of production input. Typically, the production input is labor, so variables of interest are divided by the number of workers in the economy. For example,

$$k_t = \frac{K_t}{A_t L_t} \tag{4}$$

is the equation for capital in intensive form. Changes in per capita national income, like those presented in Table (2), are also a measure of intensive growth. Here, k is the level of capital per *effective worker*, meaning capital per unit of technology-augmented labor. A key reason we care about effective workers is that workers' pay is based on their total productivity. In the real world, it is difficult to separate the productivity of an individual worker from the productivity of the technology that makes the worker more effective. For example, a firm cannot tell how much a worker contributes to its profit margins versus the electricity the worker uses at his or her desk, the car that transports him or her to work, or the computer he or she uses to write reports. It is assumed that workers capture these benefits of technology in their wages. This assumption is reasonable because wages have tended to track productivity very closely over time.

A key component in the Solow model is the *capital accumulation equation*, which describes how the stock of capital per effective worker evolves over time:

$$\dot{k}_t = sy_t - (n + g + \delta)k_t. \tag{5}$$

Here, s is the fraction of national income that is saved, and δ represents the depreciation rate of capital. It is assumed that all savings in society are automatically invested in new capital. The variables n, s, g, and δ are *exogenous variables* in the model, which means these variables are determined outside the model itself and are simply given from the outset. Economists use models to predict changes in *endogenous variables*—that is, variables that are explained within the system. In the Solow model, the most important endogenous variable is probably output per worker.

The variable $\dot{k}$ is the derivative of capital per effective worker with respect to time. In other words, it explains how much the capital stock changes at each point in time t. Equation (5) indicates that the change in the stock of capital per effective worker at time t is equal to what is added to the capital stock from investment (i.e., the fraction of income that is saved) minus what is needed to maintain *break-even investment* (i.e., the investment required to maintain a constant level of capital per effective worker). To break even, investment at time t must add enough new k to offset labor force growth, technology growth, and capital depreciation.

An economy is at its *steady state* level of capital accumulation when $\dot{k} = 0$. At this point, capital per *effective* worker is constant, and it is fairly easy to show that output per *actual* worker solves to

$$\frac{Y_t}{L_t} = A_t \left(\frac{s}{n + g + \delta}\right)^{\frac{\alpha}{1-\alpha}} \tag{6}$$

in the steady state. Although capital per effective worker is constant in the steady state, capital per actual worker is not. It grows at rate g in the steady state, because all variables on the right-hand side are constant except for A, which is growing at rate g by assumption.

Equation (6) is also an equation for the *balanced growth path* of output per worker. An economy is operating along a balanced growth path when all variables in the model are growing at constant rates. A balanced growth path is achieved in the Solow model when the economy is at its steady state, with all per-*effective*-worker variables growing at the rate of 0; all per-*actual*-worker variables growing at rate g, as shown for Y/L in Equation (6); and the aggregates K and Y growing at rate $n + g$.

A central finding of the Solow model is as follows: *Growth in output per worker along a balanced growth path is determined by the growth rate of technology*. This finding is not to say that other variables in the model, such as the savings rate or the labor force growth rate, are not important. Rather, permanent changes in these other variables influence the *level* of output per worker along a balanced growth path. Remember that levels and growth rates are both important. The other variables also influence growth rates as part of *transition dynamics*—that is, times when an economy is not operating along a balanced growth path. It turns out that economies are usually in the transition phase, not operating along a balanced growth path but instead moving toward one.

During transition periods, output per worker can grow faster than the technological growth rate g because the capital stock is growing. The absolute value of the growth rate of capital per effective worker, defined as $\frac{\dot{k}}{k}$, is larger the further the economy is from the steady state. To calculate how fast the capital stock is growing, each side of capital accumulation Equation (5) is divided by k. This gives the instantaneous growth rate, $\frac{\dot{k}}{k}$, on the left-hand side of the equation. Figure 2 illustrates the growth rate dynamics of k and shows that, the further the economy is from the steady state value of capital per effective worker k^*, the larger will be the absolute value of the growth rate of k.

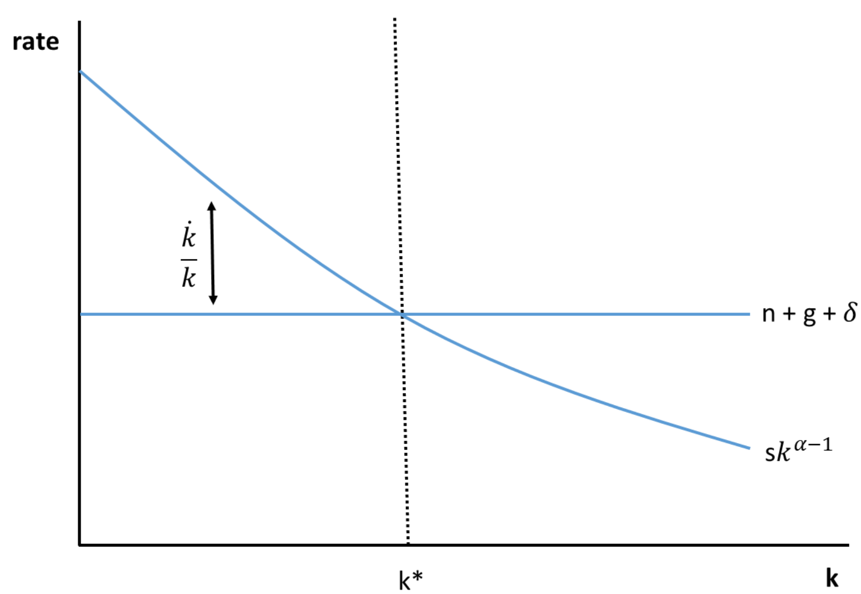


Figure 2: Acceleration and Deceleration of Growth in Capital per Effective Worker
Source: Author's illustration.

If all economies are assumed to have the same values for n, s, δ, and g, one implication of the Solow model is that economies that are further away from the steady state will grow faster than economies that are closer to the steady state. Eventually, however, all economies will converge to the same steady state. This convergence toward a common steady state is known as *absolute convergence*. If the values of n, s, δ, and g, differ, economies will experience *conditional convergence*, meaning they will converge with one another conditional on the fact that they have different underlying fundamentals and thus different steady states. When faster growth occurs in one country relative to another because the first country is further from its steady state, such growth is called *catch-up growth.*

The level of output per worker during the transition to a steady state can also be determined.[1] At all times, the level of output per worker is described by

$$\frac{Y_t}{L_t} = \left[\frac{s}{n+g+\delta}\left(1-e^{-\lambda t}\right) + k_0^{1-\alpha} e^{-\lambda t}\right]^{\frac{\alpha}{1-\alpha}} A_t \,, \tag{7}$$

where $\lambda = (1-\alpha)(n+g+\delta)$, which is the rate of convergence to the steady state.

Whereas Equation (6) describes output per worker along the balanced growth path, Equation (7) describes the *actual* path of output per worker at all points in time. Understanding where the economy is heading means understanding how the actual path

[1] See C. I. Jones (2000) and Romer (2011, 26) for specifics on how to obtain Equation (7).

of output per worker differs from where it would be the balanced growth path. To solve for the time it takes an economy to transition halfway to its new balanced growth path, the equation $e^{-\lambda t} = 0.5$ need only be solved by plugging in the appropriate values for λ and solving for t. Some empirical estimates put the value of λ at about 0.02,[2] which would imply it takes an economy about 35 years to converge halfway to the steady state. For obvious reasons, this value is known as an economy's *convergence half-life*.

1.3 Classification of Growth Effects

1.3.1 Shocks to Key Variables Induce Growth Effects

A central finding of the Solow model is that permanent changes to the growth rate of output per worker result only from permanent changes in the growth rate of technology, g. As will be shown here, this statement is not strictly true. Recurring shocks to other variables in the model can produce growth rate effects in the Solow model. For now, however, it is fair to say that the most straightforward manner by which the growth rate of output per worker can change permanently in the Solow model is through permanent shocks to the growth rate of the technology index, A (i.e., changes in g).Alternatively, one could say that permanent changes in the balanced growth path rate of growth of output per worker are only caused by changes in the growth rate of technology in the Solow model.

The variable g is exogenous in the Solow model, so the growth rate of the economy in this model is actually determined outside the model itself. Permanent shocks to the other exogenous variables in the Solow model—the savings rate, the labor force

[2] For example, see Barro and Sala-i-Martin (2004, 59).

growth rate, and the depreciation rate of capital—produce *level effects*. When thinking about the effects of regulations on economic growth, one should consider which of these key variables are affected (or shocked) by a particular policy and, by extension, what type of corresponding growth effect (level, growth rate, etc.) takes place with respect to output per worker. This distinction between the initial shock and the resulting effect is key to understanding growth changes.

This chapter reviews what is meant by level effects and growth rate effects, with one additional type of change added to the list—*transitory growth effects*. These effects are changes in output per worker that later reverse themselves. As will be shown, there are connections between these three types of changes that make them hard to distinguish in the real world.

First, consider an economy that is operating along a balanced growth path. Such a situation occurs when an economy has reached its steady state level of capital per effective worker in the Solow model. The balanced growth path will change when a key variable in the model is hit by a shock. A shock could be caused by a policy, such as a regulation or a tax, or by other forces, such as an invention, a war, or a natural disaster. Here, shocks are thought of primarily as regulations, but note that other kinds of shocks exist as well.

Shocks change the equilibrium-balanced growth path of an economy, setting output per worker on a new course. The economy will experience transition dynamics until the new balanced growth path is achieved, at which time output per worker will grow at a constant rate, determined by the growth rate of technology. Sometimes, the new balanced growth path will be the same as the old one before the shock hit (as is the case following transitory growth effects). Other times, the balanced growth path will be entirely different from the one the economy was on before the shock hit.

The *long run* in the model is defined as the time it takes for an economy to converge to its new balanced growth path after a shock occurs. The *short run* is the transition period after an economy leaves its initial balanced growth path but before it converges to the new balanced growth path. Technically, an economy *never* reaches its new balanced growth path after a shock to a variable in the model. The economy only converges toward its new balanced growth path asymptotically. This convergence should be obvious when one thinks about the half-life equation. Much as nuclear material never completely loses its entire radioactivity, an economy never fully converges to a new balanced growth path. Rather, it gets closer and closer to the balanced growth path without ever reaching it. In this sense, an economy is *always* in the short run. This situation does not mean the long run is not important. The long run describes the trajectory the economy is on, determining where the economy is heading. Furthermore, at some point, short-run transitional changes become so small that they can be safely disregarded as inconsequential.

1.3.2 Growth Rate Effect

Figure 3, panel a, illustrates an economy that begins along a balanced growth path. Initially, all per-effective-worker variables are growing at a rate of 0, and all per-actual-worker variables are growing at a constant rate g. At time t_0, this economy experiences a shock to the variable g in the model.[3] Here the y-axis measures output per worker on a log scale, with time plotted on the x-axis. Levels of output per worker are plotted in log form, so the growth rate of output per worker is simply the slope of the blue line.

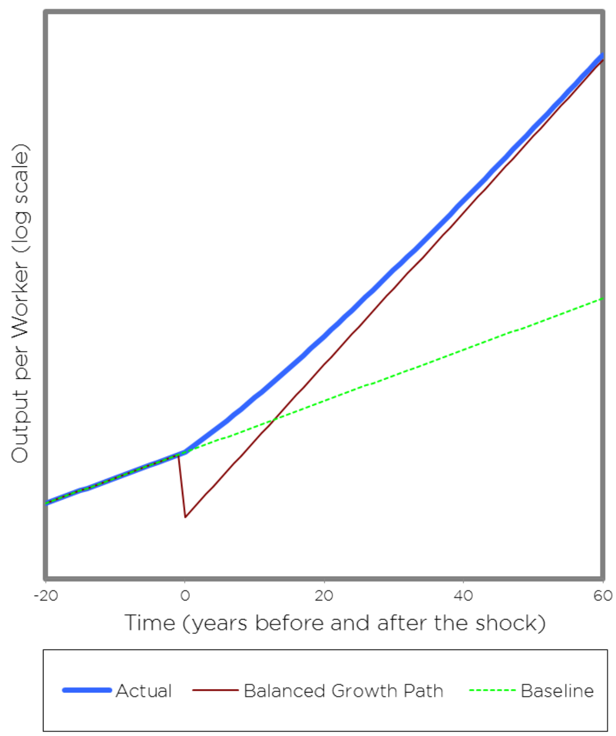
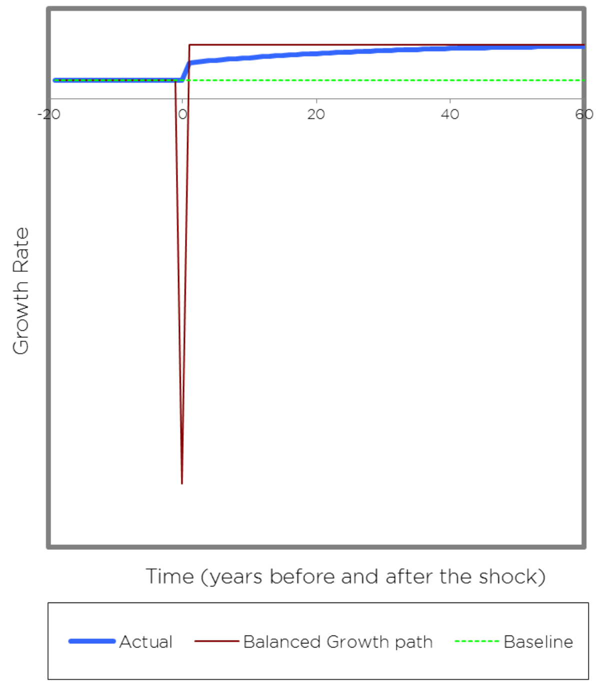


Figure 3: Growth Rate Effect
a. Levels

[3] Figure 3 as well as similar figures in this chapter were produced by modifying an impressive Excel version of the Solow model built by University of California at Berkeley economist Bradford DeLong (2006).



b: Growth Rates
Source: Author's illustration, created using DeLong (2006).

The blue line in Figure 3, panel a, represents the actual path of output per worker over time. In year 0, this economy experiences a shock to the variable g. The dashed green line shows how the level of output per worker would have continued absent the shock. This green line is the *baseline scenario*. Understanding how public policies affect economic growth means understanding how policies change output per worker relative to the baseline scenario. The red line shows the balanced growth path at all points in time. The difference between the blue and green lines is the change in output per worker resulting from the shock to g.

Again, it is important to distinguish between the shock itself and the resulting effect of the shock. The shock affects the exogenous variables in the model, whereas the resulting growth effect will be expressed in changes in output per worker. Sometimes shocks will be temporary, and other times they will be permanent. The same goes for changes in output per worker. Sometimes the effect will be temporary and other times permanent.

In the case of Figure 3, panel a, a permanent shock to the variable g in the Solow model permanently affects the path of output per worker. At the time of the shock, the balanced growth path and the actual path of output per worker diverge. To understand why this happens, logs can be taken of Equation (6) to obtain

$$ln\left(\frac{Y_t}{L_t}\right) = lnA_0 + gt + \frac{\alpha}{1-\alpha}ln(s) - \frac{\alpha}{1-\alpha}ln(n-g-\delta). \quad (8)$$

The variable g exerts influence on the log of the balanced growth path of output per worker in two ways. First, g changes the level of the technology index, A. Next, g changes the level of break-even investment of capital per effective worker. A higher g implies a higher level of break-even investment, meaning more capital is needed to offset new and better technology just to keep capital per effective worker constant.

The first effect of g raises output per worker, and the second effect depresses output per worker. In the very short run, this negative *level effect* of g actually outweighs the *growth rate effect* that g exerts by raising A. Figure 3, panel b, illustrates this effect more clearly by plotting the growth rates associated with the lines in Figure 3, panel a.

Although the initial effect of an increase in g is to put downward pressure on output per worker, the effect is quickly swamped by the upward pressure that g exerts on the level of A. The capital stock of an economy cannot immediately jump to its new balanced growth path because the capital stock is fixed in the short term. Because it takes time to adjust to the shock, once the capital stock starts to change, the upward growth rate effect of g outweighs downward-level pressures.

Permanent changes in the growth rate of technology have radical implications over time. Such changes result in revolutions—for better or worse—in living standards. When the growth rate of output per worker changes permanently, the laws of compounding take hold and the gap between the actual level of output per worker and the level of output per worker under the baseline scenario widens by a greater and greater amount over time.

What kinds of things might induce such effects? The growth rate of GDP per person has actually remained remarkably steady over time, at about 2 percent per year in the United States since the late 19th century (C. I. Jones 2015). Similar evidence can be presented for other advanced economies. This lack of any significant variation in the long-run growth rate is an important empirical finding in considering how public policies, such as regulation, affect growth rates. Growth appears to be fairly resistant to policy changes, at least in the very long run. Over shorter time horizons, growth rates vary widely, however. In the past 60 years, annual US growth rates in real GDP per capita have been as high as 11 percent and as low as −4 percent, as is demonstrated in Figure 4.

Business cycles are a main reason for this variation, but one has to wonder whether policy might have contributed to these wide swings as well.

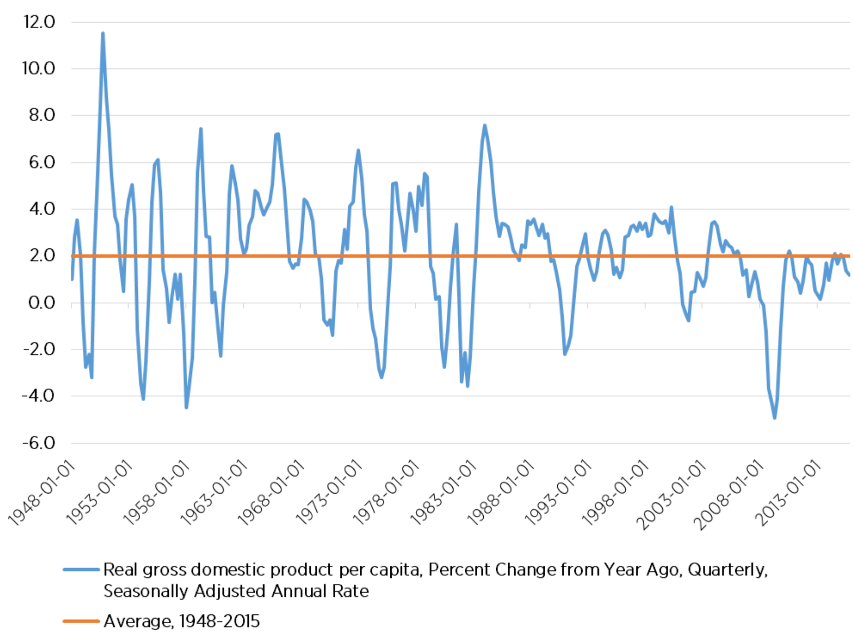


Figure 4: US Real GDP Per Capita Growth Rates, 1948–2015
Source: Author's illustration, based on data from the US Bureau of Economic Analysis, 2016.

For a policy to affect long-run growth rates, it must permanently change the productivity growth rate, i.e., the growth rate of the technology index, *g*, in the Solow model. Any single policy is unlikely to do this, although it is conceivable that the cumulative effect of many policies might impact productivity in this way. New discoveries or inventions might also permanently raise the productivity growth rate of workers. However, most inventions will increase only the *level* of the technology index, A, in the Solow model. Only technologies with sweeping, economy-wide effects could conceivably raise g permantly, and such technologies are likely to be incredibly rare, if they exist at all.

The only technologies that might come close to having such effects are general-purpose technologies (GPTs), which have been defined in many ways.[4] In this essay the definition in Lipsey, Bekar, and Carlaw (2005, 98) is used: "A GPT is a single generic technology, recognizable as such over its whole lifetime, that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects." Jovanovic and Rousseau (2005), who build on the work of Bresnahan and Trajtenberg (1995), suggest that GPTs have three main features: (a) *pervasiveness*—the GPT should spread to most sectors; (b) *improvement*—the GPT should get better over time and, hence, should keep lowering the costs of its users; and (c) *innovation spawning*—the GPT should make it easier to invent and produce new products or processes.

For a single technology to permanently raise the growth rate of all technologies, it must have endless uses. For a GPT to do so, it must be broad enough that it leads to further innovations in other areas. Perhaps the best example of a GPT is electricity. Society seems to never run out of new ways to use electricity, and without electricity there would not be other GPTs, such as computers or the Internet. Historical examples of GPTs are given in Table 3.

[4] See Lipsey, Bekar, and Carlaw (1998) for an in-depth discussion of the definition of GPTs.

Table 3: Historical Examples of General Purpose Technologies
Note: GPT = general purpose technology.
Source: Lipsey 2005.

No.	GPT	Date
1	Domestication of plants	9000–8000 BC
2	Domestication of animals	8500–7500 BC
3	Smelting ore	8000–7000 BC
4	Wheel	4000–3000 BC
5	Writing	3400–3200 BC
6	Bronze	2800 BC
7	Iron	1200 BC
8	Waterwheel	Early medieval period
9	Three-masted sailing ship	15th century
10	Printing	16th century
11	Steam engine	Late 18th to early 19th century
12	Factory system	Late 18th to early 19th century
13	Railway	Mid-19th century
14	Iron steamship	Mid-19th century
15	Internal combustion engine	Late 19th century
16	Electricity	Late 19th century
17	Motor vehicle	20th century
18	Airplane	20th century
19	Mass-production, continuous-process factory	20th century
20	Computer	20th century
21	Lean production	20th century
22	Internet	20th century
23	Biotechnology	20th century
24	Nanotechnology	21st century

It is hard to say whether any regulations have prevented the discovery, invention, or widespread adoption of a GPT. Even if this has occurred, it’s unclear whether this has reduced the growth rate significantly in developed economies. Nonetheless, just because growth has been fairly stable in the recent past does not mean it will be stable in the future. During most of the time that the human race has been on Earth, per capita income growth has been closer to 0 percent. Only since the industrial revolution have developed countries experienced per capita income growth on the order of 2 percent, suggesting that no one should assume that annual increases in living standards are automatic. This fact suggests that policymakers should be careful about blocking or delaying implementation of new technologies—especially if the technologies have the potential to be GPTs. Even short delays in the adoption rate of technologies that permanently raise productivity can have permanent effects. As an illustration, Figure 5 shows how a delayed growth rate effect compares against a world in which there is no delay.

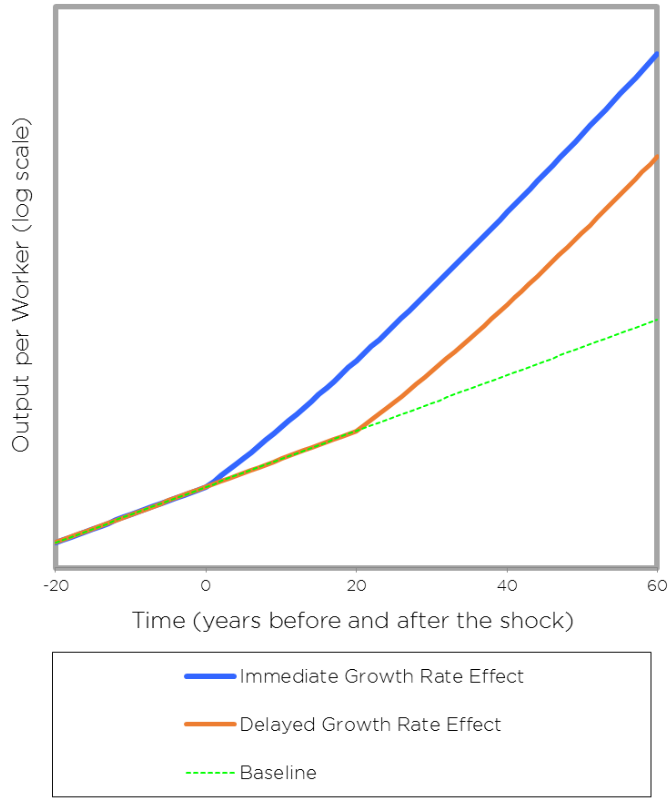


Figure 5: Immediate vs. Delayed Growth Rate Effect
Source: Author's illustration, created using DeLong (2006).

The blue line plots output per worker when a permanent increase in the productivity growth rate occurs in year 0. The orange line plots what happens if the change occurs 20 years later. As the graph makes clear, the effects of any delay are permanent. Every year, every day, even every minute that goes by without the technological breakthrough lowers the level of output per worker permanently.

There are also important redistributive consequences of growth rate effects. Technological innovations that are disruptive, like many GPTs, are likely to be heavily resisted when introduced, especially by interest groups that might be harmed or displaced by change. The Luddites, British textile workers who feared the new textile equipment that was developed during the Industrial Revolution, are a famous example of an interest group that was displaced by a beneficial new technology. In retrospect, their concerns seem almost comical. But in the emotion of the moment, and to those who are directly impacted, concern about disruptive technological change is legitimate.

Interest groups, as well as the public more broadly, often initially respond more powerfully to the negative aspects of new technologies than to the positive aspects. In part, this may be because of the psychological phenomenon of loss aversion (i.e., people tend to respond more strongly to losses than to equivalent gains), but resistance could also be a rational response to the incentives at hand. If new technologies are disruptive at first and the benefits come only later, the harms of new technologies fall on the present generation, and the greatest beneficiaries are future generations. No doubt, new technologies also benefit people in the present, but the compounding effects of productivity improvements will be most profound years in the future.

The kinds of sweeping, dramatic growth rate effects described here may be more the domain of theory than practice. It is unlikely that any single policy, unless it prevents or encourages a massive technological revolution, can influence economic growth rates in the manner described above. Even most GPTs probably do not raise economic growth rates permanently, although models exist of single-GPT driven growth.[5] The constancy of per capita growth rates over time is further evidence of this. Nonetheless, volatility in

[5] See Lipsey, Bekar, and Carlaw (2005) p. 379-384 for examples of such models.

short-run growth rates suggests that regulatory policy may still be important. The most likely way is by changing the level of output per worker, rather than by directly altering the economy's growth rate.

1.3.3 Level Effect

Standard Level Effects

This section discusses two forms of *level effects*, which occur when the level of output per worker is permanently shifted higher or lower. Unlike growth rate effects that compound over time, a level effect changes the level of output per worker by a uniform amount in every period along a balanced growth path. Such effects are caused by permanent shocks to the variables n, s, or δ in the Solow model. Here, changes produced by permanent shocks to these variables are referred to as *standard level effects* to distinguish them from temporary shocks to technology, which also induce level effects. and which will be discussed in the second half of this section.

Level effects do not compound over time like growth rate shocks. Rather, they produce fixed (positive or negative) changes in long-run output per worker that are felt indefinitely into the future. Figure 6, panel a, illustrates a positive level effect on output per worker.

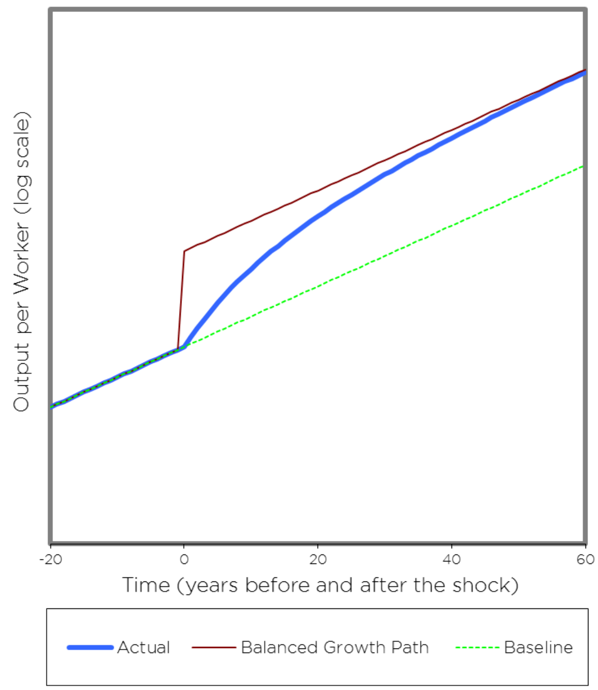
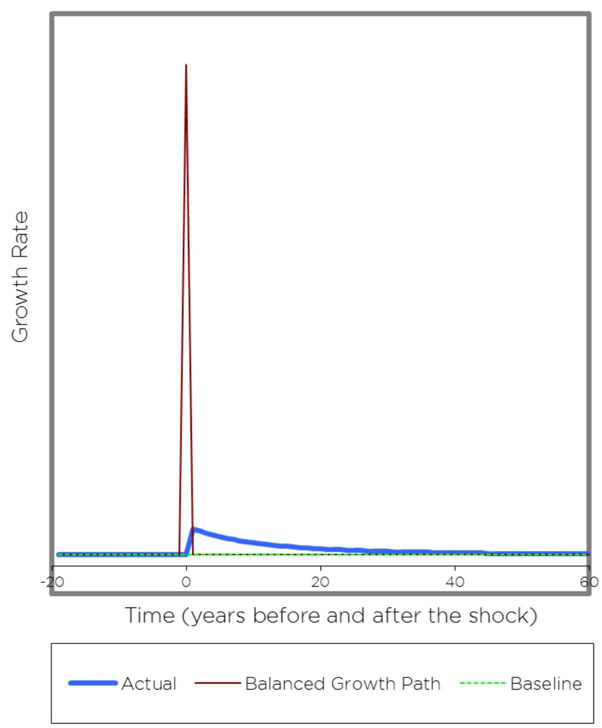


Figure 6: Standard Level Effect
a. Levels



b. Growth rates
Source: Author’s illustration, created using DeLong (2006).

Although level effects do not change an economy’s growth rate in the long run, they do change growth rates during transition periods before an economy reaches its new balanced growth path. Initially, an economy’s growth rate rises after a positive level effect; then growth slowly returns to its original level; and in the long run, the growth rate remains what it was before the shock occurred, as shown in Figure 6, panel b.

In the Solow model, several types of shocks can induce standard level effects: a permanent shock to the labor force growth rate *n*, a permanent shock to the savings rate *s*, or a permanent change in the depreciation rate of capital δ. Any regulation that influences these variables permanently will produce a level effect. From a policy standpoint, several factors that might influence the labor force growth rate, *n*. For instance, some regulations might indirectly affect the population’s fertility rate. Legalized abortion, for example, might have this effect. Public pension programs may discourage couples from having children (as individuals learn they can rely on other people’s children to support them in old age rather than their own). Alternatively, other policies, such as a child tax credit, might incentivize families to have more children.

Stricter immigration restrictions would raise output per worker in the Solow model if it slowed labor force growth. Yet, other policies, like unemployment or disability insurance, might encourage people to stay out of the labor force altogether. All else equal, if the fraction of people working in the population falls, this would raise output per *worker* but might also lower output *per capita*, since the total population hasn't changed, but everyone would now be supported by fewer workers.

These population-related outcomes from the model might be counterintuitive, but the reasoning is simple. The Solow model assumes diminishing returns to production factors, so anything that increases the number of workers along a balanced growth path will reduce output per worker. In this way, population growth is a fairly negative development in the Solow model. It might even be viewed as a Malthusian result of the model, after the classical economist Thomas Malthus, who also had a fairly pessimistic view of how the growth rate of a population impacts living standards. As will be shown in chapter 4, however, not all growth models share this pessimistic perspective about the relationship between the population and living standards.

Importantly, for a policy to result in level effects, the change must be strong enough to alter the growth rate, not just the level, of the labor force. (Policies that only change the level of the labor force are relevant to transitory growth effects, discussed later in this chapter.) Therefore, a significant policy, or set of policies, will be needed to permanently change the growth rate of the labor force. Such policies, if they exist, will relate primarily to immigration and fertility, topics largely outside the scope of this essay. Similarly, the depreciation rate of capital, δ, will be assumed to be determined by factors mostly unrelated to public policy, although the rate of innovation may have spillover effects that influence how quickly capital depreciates.

It is far more likely that regulations induce level effects by influencing s through changes in consumption and investment behavior. For example, policies that prompt individuals to contribute more to their 401(k) accounts might increase the national savings rate if the added savings are not offset by less investment elsewhere. However, restrictions on investments of various kinds might discourage people from saving and might reduce output per worker.

As with growth rate effects, important distributional factors must be considered with level effects. Because the gap between long-run output per worker and output per worker under the baseline scenario is constant over time, an immediate level effect that reduces output per worker by $1,000 today will also reduce output per worker by $1,000 next year, the year after, and so on. If income is rising over time—say due to economic growth—the change will feel far more significant today than it will in the future. One thousand dollars, even adjusted for inflation, may feel inconsequential as a fraction of income to an American 100 years from now. As a fraction of this year's median income, $1,000 is quite significant.

Because of this distributional effect, policies that produce positive level effects can be expected to have progressive redistributive consequences across time in the sense that the policies provide gains that are a larger fraction of income to the present poorer generation than to future richer generations. Policies that produce negative level effects have regressive redistributive consequences. In either case, level effects will seem to be of more consequence to people in the present than to people in the future, assuming that income levels are higher in the future This distributional consequence is an important contrast between growth rate effects and level effects. Growth rate affects produce consequences that will feel most pronounced in the future.

Given their attention to short-run factors, it may well be that policymakers and voters alike are more concerned with producing positive level effects and avoiding negative level effects than they are with producing a policy framework that produces growth rate effects.

Technology-Induced Level Effects

Temporary shocks to g in the Solow model also produce level effects. Such an outcome is almost identical to the level effects produced by permanent shocks to n, s, or δ, although the exact path that the level of output-per-worker follows is not identical in the two cases. Figure 7 illustrates such a technology-induced level effect. A technology shock arrives at time t_0, lasts for one year, and then abruptly reverses.

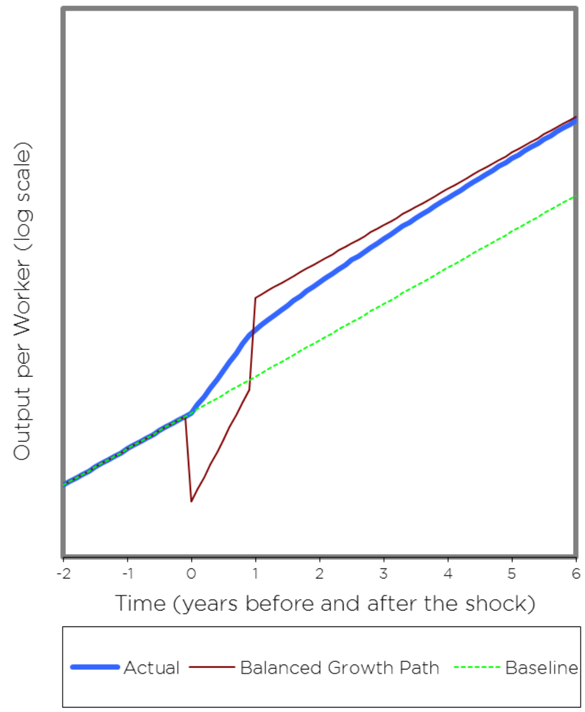
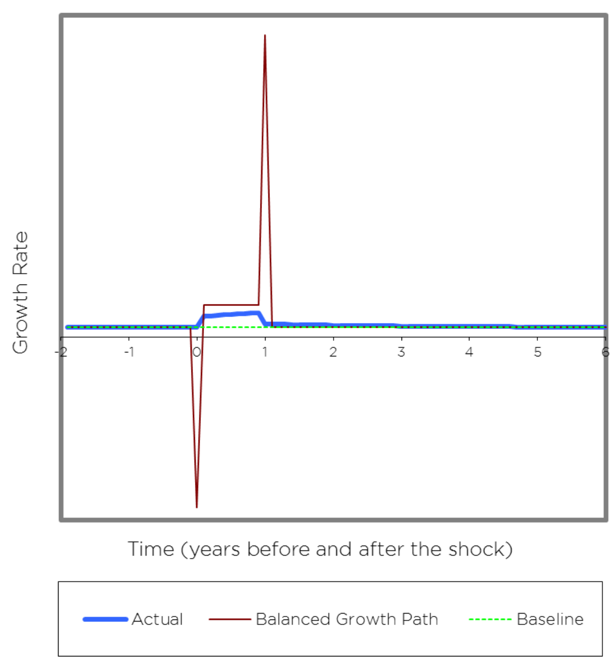


Figure 7: Technology-Induced Level Effect
a. Levels



b. Growth rates
Source: Author's illustration, created using DeLong (2006).

One can imagine temporary shocks to *g*, such as presented in Figure 7, occurring for many reasons. First, a powerful new technology might get started but then never fully take off. If some critical mass of use is necessary before people fully adopt a new technology, such a technology might begin to increase productivity, but then the change would reverse. Technologies often have a life cycle that eventually runs out as well. Consider a simple fax machine, which initially improved business productivity since it reduced the time required to send documents long distances. Over time, however, marginal uses of the facsimile (like spamming by marketers) likely did not add much value. Eventually, with the advent of email, fax machines became largely obsolete.

Most technologies—maybe even most GPTs—experience diminishing returns. With diminishing returns, new discoveries will not induce *permanent* technology shocks but instead *temporary* shocks, which result in level effects rather than growth rate effects. In most cases, it is reasonable to think that shocks to *g* are temporary, not permanent.

As discussed earlier, technological improvements from an increase in *g*, improve worker productivity. It is well-documented that regulations often affect the productivity of firms in a negative manner.[6] This result is fairly straight forward. Regulations create

[6] See, for example Conway et al. 2006; Égert 2016; Erlandsen and Lundsgaard 2007; Garicano, LeLarge, and Van Reenen 2013; Nicoletti and Scarpetta 2003.

added costs for firms. Those regulations may have benefits, but typically the intended outcomes of regulation are not to increase firm output. Rather, as managers devote time to understanding and complying with regulations, the cost per unit of output increases. Hence, by definition the productivity of the firm is lowered.

Some regulations induce laborers to work harder, thereby boosting productivity. For example, policies that create unemployment, such as minimum wage laws, can have the side effect of causing workers who retain their jobs to work harder. With their wages now above the market clearing wage, such policies can boost employee morale, prevent shirking, and lower turnover costs for firms (Yellen 1984). The existence of structural unemployment might also increase worker productivity out of fear of being cast into the ranks of the unemployed. Of course, it is unclear whether such policies will actually increase output per worker since fewer workers are employed. Such policies also create winners and losers, so it is ambiguous whether social welfare is improved in aggregate.

Another interesting implication of level effects is what happens when they are delayed. When a new productivity-boosting technology comes around, a delay in adopting it will make no difference to output per worker in the long run. This result is shown in **Error! Reference source not found.**, which illustrates a level effect that takes place at time t_0, compared against the same effect after a 10-year delay. With or without a delay, as long as the technology is eventually adopted and the technology experiences diminishing returns or some reversal of adoption, output per worker will look the same in the long run under both scenarios.

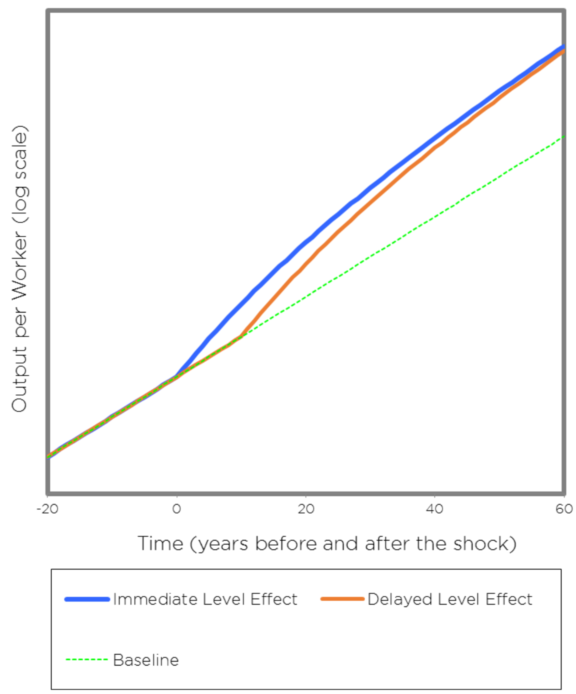


Figure 8: Immediate vs. Delayed Level Effect
Source: Author's illustration, created using DeLong (2006).

Recall from Figure 5 that delaying a beneficial growth rate change leads to permanent losses in income. If a policy were to cause this effect, it would have devastating consequences. As **Error! Reference source not found.** demonstrates, however, a counterproductive policy producing a negative level effect can be reversed with no long-run effect. This result suggests that level effects are often reversible, while growth rate effects are not. Of course, this is true for policies that produce both positive and negative effects.

These findings are relevant to the precautionary principle in public policy. That principle has been described as “the belief that new innovations should be curtailed or disallowed until their developers can prove that they will not cause any harms to individuals, groups, specific entities, cultural norms, or various existing laws, norms, or traditions” (Thierer 2016, 1).

Advocates of the precautionary principle argue for delaying new technologies until they are proven safe. This approach to policy affects economic growth very differently depending on the type of technology being affected. With sweeping revolutionary technologies, such as some GPTs, a huge tradeoff may have to be required if the precautionary principle is taken seriously. Even if a technology has the potential to result in catastrophic outcomes (e.g., nuclear power), delaying adoption of a new GPT could also have catastrophic outcomes to living standards for future generations. Such tradeoffs between risks and benefits are important to consider.

With smaller innovations or with technologies that are expected to run into diminishing returns eventually, the precautionary principle has lower opportunity costs. Delays in small innovations will make little difference to output in the longer run. But this result does not mean there are no consequences of delay. Rather, for less consequential technologies, the relative benefits and costs of each new technology must be carefully weighed when making judgment calls about how quickly to adopt. Upfront risks must be balanced against the benefits of a new technology, keeping in mind that it is the present generation of citizens who will realize the benefits most profoundly.

If the potential exists for large catastrophic consequences from a small or diminishing-returns new technology, delaying the technology may be a sensible idea until more is known. However, if the downside risks are small and the technology is likely to bring much utility to the current generation, delay does not make as much sense. Furthermore, a permanent delay, such as banning a technology, results in permanent losses regardless of whether the technology would produce level or growth rate effects.

Therefore, the precautionary principle should be taken seriously, but along a continuum. The shorter the implementation delay, the more limited the potential applications of a new technology (i.e., the less pervasive, the less room for improvement in the technology, and the less innovation spawning the technology is likely to be); the higher the downside risks, the more the precautionary principle might be reasonable. By contrast, the longer the delay is likely to be, the more wide-ranging and applicable the new technology is (i.e., the more it is like a GPT); the lower the downside risks, the more the precautionary principle is unreasonable.

1.3.4 Transitory Growth Effect

The final type of change discussed here is a *transitory growth effect*, which results from a temporary shock to n, s, or δ that eventually reverses itself. Transitory effects do not change output per worker in the long run; rather their effects eventually wither and slowly disappear.

If the labor force growth rate increases in a single year and then returns to its previous rate, this would produce the type of effect seen in Figure 9. A sudden burst in immigration might temporarily boost the growth rate of the country's labor force. If, after a year, the growth of the labor force returns to its previous rate,[7] then in the long run output per worker will return to the same balanced growth path as before the burst. Something similar would occur if the growth rate of the capital stock were to fall suddenly for only a single time. The transition dynamics are such that the growth rate initially turns negative before balancing off, speeding up, and ultimately returning to its initial level.

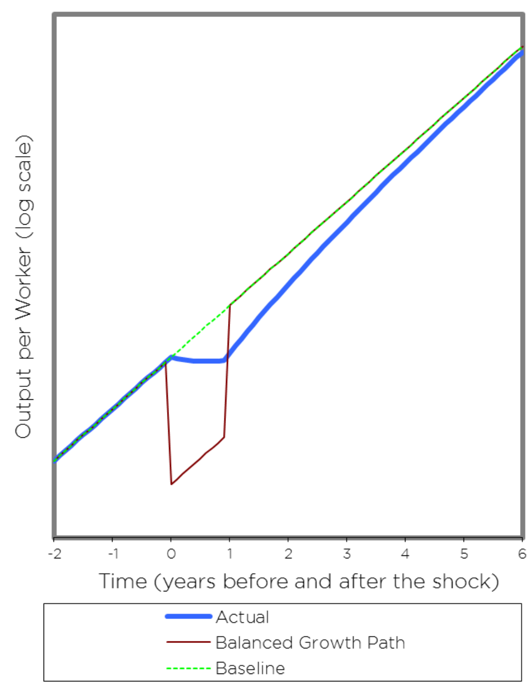
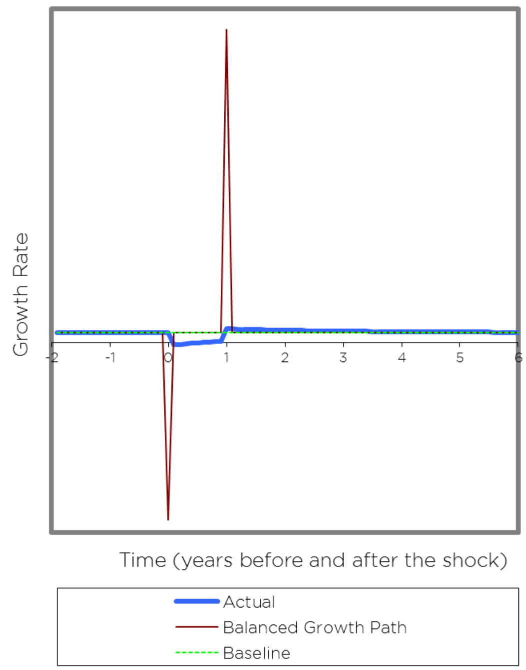


Figure 9: Transitory Growth Effect
a. Levels

[7] Note, however, that the level of the labor force has permanently increased.



b. Growth rate
Source: Author's illustration, created using DeLong (2006).

As with level effects, delaying temporary shocks will not matter in the long run. Output is eventually no different than if the shock never occurred. Nonetheless, Short-run effects such as these can be very large in magnitude, and their effects can last for years owing to the long convergence half-life found in most economies. Furthermore, delaying a transitory growth effect will have important distributional consequences. A shock that occurs today will affect a different group of people than a shock that occurs 10 years from now. Even if in the long run such effects do not matter much, they can matter a great deal to the people who are directly affected.

Because regulations govern the flows of both global capital and migration, regulations can indeed produce such effects. The types of regulations that induce transitory effects mirror many of those that produce level effects. The only difference is whether the shock caused by the regulation is temporary or permanent. For example, regulations that reduce net investment but that are later reversed will reduce output per worker in the short run. Temporary regulations that encourage investment will boost output in the same way. Similarly, eliminating some regulations already on the essays could produce transitory growth effects depending on whether a given reform has temporary or permanent effects. If rules that encourage firms to invest abroad are suddenly repealed, a flood of investment might rush back into the country in a very short period of time. There would be a temporary investment boom that would later reverse itself, eventually returning investment to its former growth path.

These findings also have relevance to the expenditures made by firms when engaging in compliance activities. Recurring compliance costs that displace investment by firms, such as the ongoing costs of maintaining compliance departments, will lead to downward level shifts in output per worker. Yet, one-off drops in investment, such as the one-time cost of a government information collection request, will produce transitory effects. Of course, astute observers will note that many compliance expenditures actually show up in GDP, since filling out forms and having lawyers to draw up documents are both market activities.

It is debatable whether compliance activities should be included in GDP. Remember, GDP measures the market value of final goods and services, and compliance expenditures look a lot like spending on intermediate goods where the final good consumed by the public is whatever outcome the compliance activity is intended to bring about. If a plant installs pollution control equipment, the final good consumed by the public is cleaner air, not the pollution control equipment. The final good, cleaner air, is like a good that is sold to the public at zero price. Recall that GDP is not a measure of overall human well-being; rather, it is a measure of income. Although clean air provides benefits to the public, it does not directly contribute to national income. Thus, a more accurate measure of national income might exclude compliance expenditures from GDP and treat them more as something along the lines of charity. Although this exclusion may be controversial, it is an example of how income and welfare do not always move in the same direction.

It is also important to distinguish between (a) output losses resulting from declines in investment and (b) any additional losses resulting from productivity declines as a firm's attention is diverted from production activities and toward compliance activities. Compliance activities can affect s and g simultaneously as businesses (a) reduce investment when they are forced to spend resources on compliance and (b) also suffer productivity losses as effort is diverted from production activity. Keeping these different shocks and outcomes distinct is critical. To assist in this endeavor, the next sections explore more formally how regulatory shocks can influence multiple variables in a growth model at the same time.

1.3.5 Interrelations of Growth Effects

We have seen how shocks to individual variables in the Solow model can produce very different types of changes, depending on whether a given shock is permanent or temporary. Similarly, the different kinds of growth effects can be viewed as permanent or temporary versions of other effects. For example, a level effect can be viewed as a series of permanently recurring transitory growth effects. This relationship is illustrated in Figure 10. Level effects occur when temporary shocks occur every period in perpetuity, each one producing a transitory growth effect. The colored lines in Figure 10 show the balanced growth paths associated with each new shock; the blue line represents the actual path of output per worker as the economy is subjected to a series of permanently recurring transitory effects.

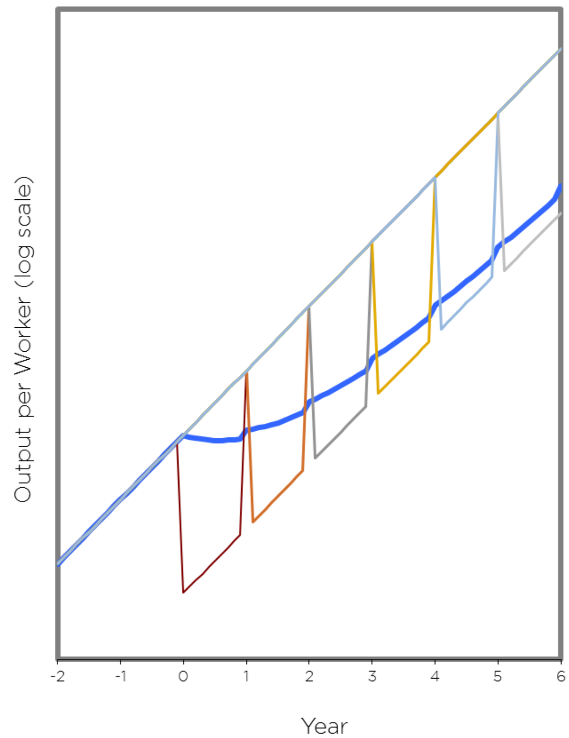
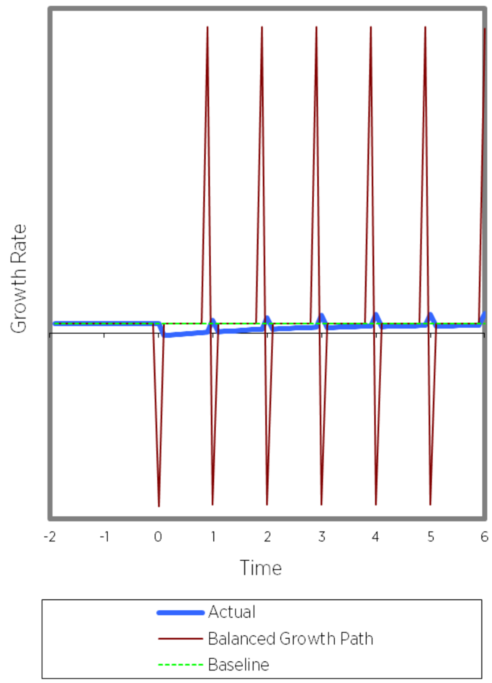


Figure 10: Recurring Transitory Effects
a. Levels



b. Growth rates
Source: Author's illustration, created using DeLong (2006).

Because of this relationship between transitory and level effects, it might be inferred that growth rate effects can be thought of as a series of permanently recurring level effects, which is indeed the case. Recurring, and permanent, shocks to the variables *n*, *s*, or *δ* in the Solow model produce growth rate effects when the shocks build on one another. Thus, the variables that are often claimed to "only" produce level effects in the Solow model—*n*, *s*, or *δ*—can actually produce growth rate effects as well. To do so, however, shocks to these variables must permanently prevent the economy from reaching its balanced growth path, as shown in Figure 11. A series of unrelated temporary shocks to productivity will also result in these kinds of growth rate changes.

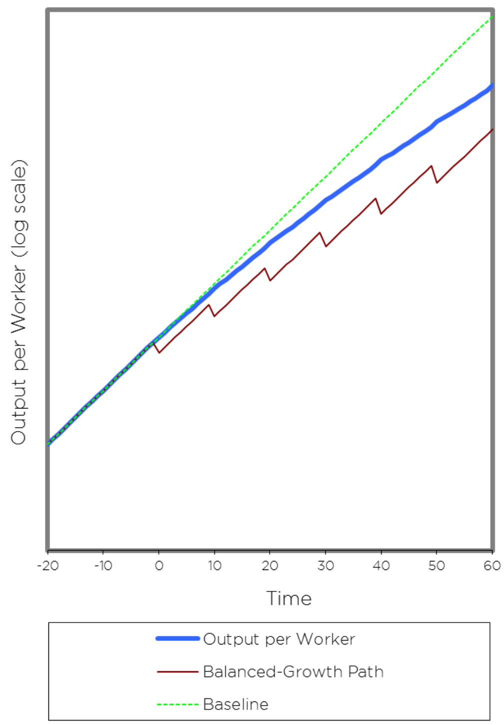
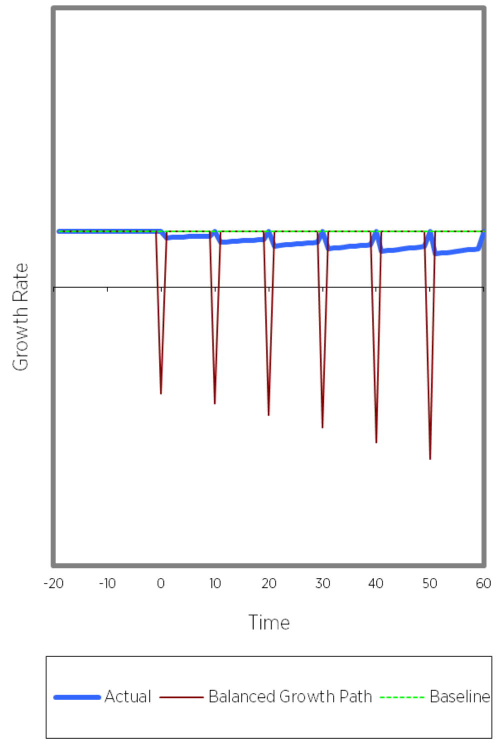


Figure 11: Recurring Level Effects
a. Levels



b. Growth Rates
Source: Author’s illustration, created using DeLong (2006).

Of course, permanently recurring rises or declines in population growth, the savings rate, or the depreciation rate are difficult to achieve on an ongoing basis. A society cannot save more than 100 percent or less than 0 percent of its income, so it is impossible to permanently increase the savings rate by 1 percent each year in perpetuity. People would have to stop eating at some point. That said, small changes in the three rates are realistic, and the effects could last for years or decades, such that each effect feels permanent even if at some point the economy eventually reaches its balanced growth path. The lesson here is that it is the cumulative effect of a series of public policies—many completely unrelated—that is most likely to induce growth rate effects. A single policy in isolation will rarely achieve such an outcome.

1.4 How Regulations Enter the Economic System

1.4.1 Regulatory Complexity and Interaction Effects

Another reason the cumulative effect of all regulations may have the most consequences for growth rates is that new regulations interact with existing ones, resulting in effects larger than the new regulations would create on their own. Consider the simple case where there are only two rules on the books—one old and one new. Both regulations could have an effect on output when acting in isolation, but there is also the potential for an interaction effect between the two regulations once both are in place at the same time.

Interaction effects among regulations have been compared to dropping pebbles in a stream (Mandel and Carew 2013). The first pebble may not slow the flow of water in a noticeable way, but the thousandth pebble might, and the millionth pebble might stop the flow altogether. This example is true despite the fact that the millionth pebble might be of little consequence if it were the first pebble dropped in the water. When hundreds of thousands of regulations are on the books, adding a single new one can produce much larger effects than one would expect from looking at the single new regulation in isolation.

Anticipating the possible interaction effects of hundreds of thousands of restrictions that are in the legal code is a daunting task. The task becomes ever more difficult when considering the how effects of regulations change over time. For example, a regulation issued in period t would have effects in periods $t+1$, $t+2$, and so forth.

Interaction effects of the regulation may differ in every period. Indeed, individual regulations may even interact with themselves across time, a phenomenon known as serial correlation.

One implication of interaction effects is that added complexity itself may induce changes in output per worker, although such effects are poorly understood. If one is inclined to think that greater complexity is more likely to lower rather than increase output per worker, a strong rationale is thus provided for capping the size of the regulatory code at some manageable level. One way to achieve this would be to remove an old regulation every time a new one is put in place to ensure the code does not grow indefinitely.

Microeconomic analysis will probably not be able to estimate the effects of regulatory complexity, but macroeconomists may have more success. Indeed, macroeconomists have already begun looking at the cumulative impact of all regulations on growth, and the results are profound. One study finds that the cumulative burden of regulations has slowed the growth rate of GDP in the United States by approximately 2 percent every year since 1949 (Dawson and Seater 2013). The same study finds that regulations also affect other key growth determinants, such as total factor productivity and capital and labor services. A clear lesson for regulators is that they should be more careful as regulations continue to be put in place, because there are likely to be additional unintended effects, both positive and negative, as the code grows larger.

1.4.2 The Innovation Spider Web

Regulations restrict behavior and limit the range of opportunities available to people. This constraining aspect of regulation is why some recent measures of regulation count restrictions—words such as *shall*, *prohibited* and *may not*—that appear in the *US Code of Federal Regulations*. As of this writing, there are more than 1 million restrictions in the *US Code of Federal Regulations* (Al-Ubaydli and McLaughlin 2015). Limiting choices can be beneficial if some choices would result in undesirable harms. At other times, restricting choice prevents improvements from transpiring that would otherwise occur.

Here, such improvements can be thought of as innovations that enhance worker productivity. In the Solow model, this would be anything that raises the level of the technology index, A. Future sections of this essay will show that productivity enhancements through innovation come in many forms: formal education and job training, informal learning through work experience and specialization, new products, quality improvements, and knowledge transfer and imitation.

Formal education and training includes, for example, completing a course in computer programming to learn new skills. By contrast, informal learning through experience or specialization takes place when, for example, a worker on an assembly line learns how to make finicky machinery he or she operates run smoothly. Both innovations might increase daily output at a factory, even with no new labor or capital added to the production process. New products increase both the number of goods and services that consumers may purchase and the number of production inputs available to produce more consumer goods. Quality improvements occur when an old product, such as rotary telephones, is replaced with a new and better version—smartphones. Finally, knowledge transfer and imitation occur through the sharing of information. For example, when a US company opens a factory in China, plant managers might teach the new employees methods of production that were developed in the United States.

Because innovations come in all shapes and sizes, it may seem odd to lump them together in one category. There is also likely to be some overlap in the categories just described. The one characteristic all these innovations have in common, however, is that they increase total factor productivity. Furthermore, many such enhancements relate to the discovery and use of new knowledge. Sometimes knowledge discovered was never known to another human being before. More often, however, knowledge exists in certain times and places and must be rediscovered or transferred to new individuals to be put to good use. The diffusion of knowledge is what enhances productivity, drives economic growth, and raises living standards over time.

Regulation can play an important role in both advancing and stifling knowledge diffusion. We can think of knowledge as existing in a kind of innovation spider web, whereby discoveries are mapped according to the pathways that allow individuals to uncover new productivity-enhancing knowledge. Figure 12 illustrates an example of an innovation spider web. The black lines represent the various paths by which discoveries might be made, and the blue circles are the innovations themselves. Restrictions limit the number of discovery pathways that are available to society. In extreme cases, these restrictions make it impossible for specific innovations to be uncovered via any pathway.

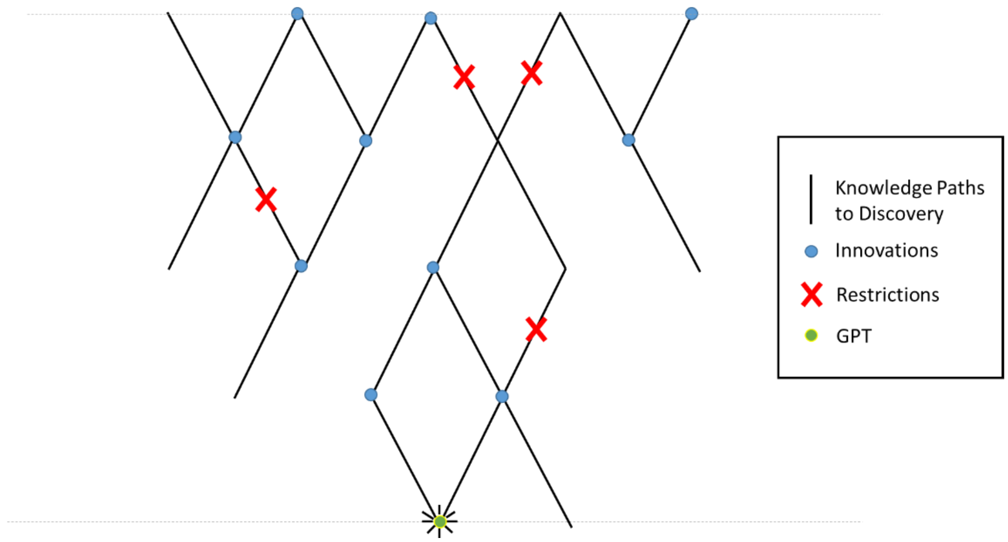


Figure 12: Innovation Spider Web
Source: Author's illustration.

Of course, it is not always bad if some innovations are sealed off and kept out of reach. Most people would not want to see options available that could lead to low-priced portable nuclear weapons. Such an invention might increase the productivity of terrorists; so not all productivity is good. Unfortunately, if society restricts discovery pathways, it can never be sure if access is being blocked to only harmful innovations or also beneficial discoveries as well. This predicament is part of the very nature of undiscovered knowledge.

Restrictions that block discovery paths need not be regulations, either. Culture and religion restrict behavior as well, and such restrictions can also block discovery pathways. The printing press, invented in the 1400s, met resistance from some leaders in the Roman Catholic Church in Western Europe as well as Islamic leaders in the Middle

East. Countries with greater openness to technological change have generally thrived economically compared with those that resist and reject technological change (Comin, Easterly, and Gong 2010).

Even market prices, which are often a useful guide for entrepreneurs, can be remarkably deceptive when it comes to innovations. Market prices will not reflect the opportunity cost of resources if some uses of those resources have not yet been discovered. For example, before the many uses for oil were discovered, it was likely viewed as useless sludge. Prices only reflect information that is known to people, but when people do not possess certain kinds of information, prices can be misleading.

Other forms of law aside from regulations, such as patent protections, also can restrict pathways to innovation. In some extreme cases, patent protections have led to what is known as a tragedy of the anticommons (Heller 1998). A tragedy of the anticommons occurs when multiple parties own overlapping property rights to the same resource, such as when multiple government authorities have taxing power to install tolls on a road. This overlapping authority can lead to underuse of the resource as authorities tax the resource beyond the socially optimal level.

In the case of patents, the tragedy of the anticommons plays out as a fragmentation of the market (Boldrin and Levine 2008). If different parties hold the rights to unique knowledge pathways along the spider web of innovations, the resulting scenario is that no one has the ability to reach certain innovations without first gaining the consent of many other parties. A coordination failure ensues when the transaction costs involved in gaining the consent of every unique patent holder exceed the expected returns to discoveries. The benefit-cost calculus to an individual may make innovation not worth the cost of obtaining permission, even though the benefits to society in general of innovation might vastly outweigh the costs.

In such cases, fewer innovations will be developed than are socially desirable. A similar scenario may even be occurring in the United States now. There have been large increases in the number of patents granted over time (Dourado and Tabarrok 2015), even though productivity growth remains relatively low compared with historical averages. It might be that the innovations these patents are protecting either (a) do not have much productivity-enhancing effect or (b) the patents themselves are stifling productivity growth by granting monopoly privileges to patent trolls rather than encouraging innovation.

1.4.3 Specialization and the Extent of the Market

Usually, when regulations block access to new innovations, the innovations in question probably have a narrow range of highly specific uses, as opposed to being general-purpose technologies, which are very rare, at best occurring a few times in a century. Smaller innovations will be more targeted. As a result, it might be surmised that preventing society from having access to smaller discoveries might have limited repercussions, but that is not necessarily the case. Even highly specific technologies with narrow uses can result in significant social losses when whole classes of products or production techniques are prevented from being developed.

One of the interesting aspects about the age we live in is that production processes tend to require more and more highly specific inputs over time. This is simply a quality of technological progress. As production becomes ever more complex, inputs in the production process become more highly specialized, which is true of inputs such as workers and also of equipment that may be perfectly tailored for the particular task at hand. In complex processes, a single missing or faulty element in the production chain can cause the entire production line to fail. This situation has come to be known as the O-ring theory of economic development (Kremer 1993a), so named because the 1986 explosion of the *Challenger* was due to failure of the space shuttle's O-rings.

Few would have guessed that one small component in a complex piece of machinery like a space shuttle could be so critical, but this phenomenon is true in all kinds of production processes. An automobile is useless without brakes. A personal computer is useless without a memory card. In the same way, preventing even small, highly specific innovations from being developed can have widespread ripple effects if new lines of production depend on that small innovation as a critical input.

A key insight of trade theory in recent decades has been that more open trade gives firms access to a greater variety of more highly specialized production inputs. Access to such resources allows firms greater opportunity to specialize and differentiate their products, leading to increasing returns to scale in some cases (Krugman 1980). Therefore, regulations in the form of trade restrictions can limit the ability of firms to specialize and to take advantage of increasing returns to scale where the phenomenon exists.

Thus far, we have mainly explored the role that regulation plays with respect to knowledge-generation. Knowledge is the key ingredient in enhancing productivity, and regulation places limits on what kind of knowledge can be uncovered. However, regulations also affect knowledge generation through changes in investment activity, which itself results in discoveries. The next sections looks more closely at how changes in saving and investment behavior contribute to economic growth. In the jargon of economists, key variables such as the savings rate and the level of technology will be endogenized to explain their evolution in the economic system. Previously, we treated these parameters as given. By doing this, deeper analysis is possible of the fundamental causes of economic growth than the Solow model has allowed.

1.5 Models of Economic Growth

1.5.1 The Ramsey–Cass–Koopmans Model

As insightful and influential as the Solow model has been over the past half-century, the model is too simplistic to explain some of the most important aspects of economic growth. The central force behind long-run growth in the Solow model—technology—is determined exogenously. This result is not satisfactory. Over the past few decades, economists have worked with the Solow model's core insights to build more sophisticated models that go further in explaining the most important drivers of growth.

Perhaps the second most famous growth model in economics is one that brought together the work of Ramsey (1928), Cass (1965), and Koopmans (1965) —aptly named the Ramsey–Cass–Koopmans growth model. Most of the key takeaways from the model mirror those from the Solow model. For example, along a balanced growth path, the economy grows at rate g, the exogenously determined rate of technological progress. The major difference between the Solow model and the Ramsey–Cass–Koopmans model is that the latter is built on microfoundations. Whereas the Solow model focuses only on economy-wide aggregates, such as the national savings rate, the savings rate in the Ramsey–Cass–Koopmans model is endogenously determined as a result of optimization behavior at the individual (or household) level. Specifically, individuals optimize utility, such that their consumption is described according to the function

$$\frac{\dot{c}_t}{c_t} = \frac{r_t - \rho}{\theta}, \tag{9}$$

which states that the consumption path of the representative agent grows at a rate that adjusts to account for the gap between the interest rate r at time t and the consumer's rate of time preference, ρ, taking into account the consumer's degree of relative risk aversion, θ. Consumption is a *control* variable in the model in that it is the variable that the optimizing agent controls to bring about equilibrium. Both ρ and θ are important new variables in the framework because they help determine the degree to which the representative agent is willing to save and invest.

As in the Solow model, permanent changes in the savings rate still affect the level of output per worker along a balanced growth path, but now there is a microeconomic explanation for what causes a change in the savings rate. A permanent rise in ρ means that the representative agent becomes more impatient. Compared to before the change, the agent now values present consumption relatively more than future consumption and will shift consumption forward in time accordingly. The national savings rate falls as a result, producing a downward shift in the level of output per worker. Therefore, shocks to the parameter ρ in the Ramsey model have effects just like shocks to s in the Solow model. Permanent shocks to ρ produce level effects, whereas temporary shocks produce transitory growth effects.

The same is true for the coefficient θ, which explains how much risk the agent is willing to bear and also the degree to which the agent's marginal utility declines as consumption rises. A rise in θ means the agent becomes more risk averse and is thus less willing to undergo swings in consumption to take advantage of the gap between the interest rate, r, and the agent's rate of time preference, ρ. A higher θ causes the agent to save less as the agent smooths consumption more across time. This pushes down the level of savings and investment and decreases the level of output per worker in society.

It is certainly plausible that some regulations might induce the public to be more shortsighted than otherwise. Policies that create principal agent or moral hazard problems might influence ρ and θ—for example. If managers at firms expect to be bailed out if they get into trouble, these managers might be willing to take on more risk than is optimal and may be more shortsighted. It is more likely that policies influence the rate of return in Equation (9), however. For example, government borrowing might drive up interest rates or taxes on investments might drive a wedge between the rate of return earned on investments and the rate of borrowing to pay for the investments. Regulations that change the rate of return on financial assets will influence the consumption behavior described in Equation (9) because optimizing individuals will adjust their saving and consumption as interest rates move closer or further away from the individual's rate of time preference.

1.5.2 Human Capital Models

In the Solow model, technology augments human labor so as to make it more productive, thus making it the primary determinant of rising wages. Many economists believe that human capital, which broadly refers to people's knowledge, education, and skills, can also augment labor so as to make it more productive. Human capital is the first formal form of knowledge that will enter the models reviewed here. One can think of it as having previously been included in the technology index, A, and now its effects will be isolated from other labor-augmenting influences. A strong correlation between the level of human capital and GDP per capita across countries provides empirical support that the contribution of human capital to growth is meaningful.

The two most famous attempts to incorporate human capital into an economic growth model are from Lucas (1988) and Mankiw, Romer, and Weil (1992). These two models take slightly different approaches toward endogenizing human capital into a growth model, but both begin from Cobb–Douglas and Solow origins.

Lucas's approach is to assume there is a tradeoff between using time to develop job skills and using time to produce output. Time is divided between these two activities, and a society can only gain more output at the expense of less education and training and vice versa. C. I. Jones (2001) presents the following simplified version of the Lucas approach, using the Cobb–Douglas production function:

$$Y = K^{\alpha}(AH)^{1-\alpha}, \tag{10}$$

where $H = e^{\psi u}L$. Here, u represents the fraction of time that laborers spend acquiring new skills, so $1 - u$ is the fraction of time spent working in production activities. The labor force, L, is defined as $(1-u)P$, where P is the total population, so this expression describes how the labor force shrinks as people take time off to obtain new skills.[8] The variable H is the level of human capital–adjusted labor, and the term ψ is the payoff for each additional unit increase in time spent obtaining skills. The level of technology is again represented by the index A; however, in this case, human capital augments labor, and technology augments human capital–adjusted labor.

The solution for the balanced growth path of output per worker in the Lucas model is

$$\frac{Y_t}{L_t} = \left(\frac{s}{n+g+\delta}\right)^{\frac{\alpha}{1-\alpha}} e^{\psi u} A_t. \tag{11}$$

The average level of human capital per worker, $e^{\psi u}$, is a constant. As a result, there is very little difference between this model and the traditional Solow model. A look back at Equation (6) demonstrates how closely Lucas's human capital model follows the Solow model. The new parameters, ψ or u, become new standard variables, in that any permanent shock to either ψ or u will produce standard level effects in the model, whereas temporary shocks will produce transitory growth effects. The rate of output per worker still grows at the rate of technological progress, g.

There are limits to how much of a shock to u is feasible given that laborers cannot spend more than 100 percent of their time developing skills. Society also gives up production with increases in the fraction u, so there are likely to be diminishing returns to

[8] This description is clearly an oversimplification, since some people who are not in the labor force will be doing things other than obtaining new skills.

developing skills. The 10th year of education may produce valuable training, but the 15th year probably less so, the 20th year even less, and so on.

If governments want to raise output per worker through increases in the amount of time spent obtaining job skills, they must consider what society gives up as more time and resources are devoted toward skills development. If time is better spent producing output than obtaining training, human capital returns might be so low as to make more investments in schooling counterproductive. In other words, investments in human capital should pass a benefit-cost test.

Policymakers may have more luck increasing ψ, the human capital payoff. Increasing ψ requires that knowledge improves over time such that the same amount of time devoted to training produces more human capital. More highly skilled teachers might accomplish this, as might more useful information in textbooks. The key is to improve the current state of knowledge or the mechanisms of transmitting knowledge to the young and uneducated.

Mankiw, Romer, and Weil (1992) take a slightly different approach in their model of human capital. Unlike the Lucas model, where human capital augments labor, these authors assume human capital is a separate input in the production function such that

$$Y = K^{\alpha} H^{\beta} (AL)^{1-\alpha-\beta}. \tag{12}$$

Note that the assumption of constant returns to scale is maintained, so α and β are between 0 and 1 and together sum to less than 1. Solving for the equation for output per worker along the balanced growth path yields

$$\frac{Y_t}{L_t} = \left[\left(\frac{s_K}{n + g + \delta} \right)^{\alpha} \left(\frac{s_H}{n + g + \delta} \right)^{\beta} \right]^{1-\alpha-\beta} A_t. \tag{13}$$

Here, s_K is the fraction of savings dedicated to physical capital accumulation, and s_H is the fraction of savings dedicated to human capital accumulation. Both forms of capital are assumed to depreciate at the same rate, δ. In this model, human capital is like physical capital in that it is generated by forgoing consumption—that is, saving more. Recall that in the Lucas model human capital results from forgoing production.

As in the Lucas model, permanent shocks to human capital produce level effects. Both models suggest that a more educated labor force will (with all else equal) be associated with a richer country, and this turns out to be the case empirically. Figure 13 plots the relationship between human capital levels in 2011, as measured by the Barro–Lee Educational Attainment Dataset and the level of output per worker across countries in the same year. There is a strong correlation between the two variables, with R^2 = .30, meaning human capital differences explain about one-third of the variation in output per worker across countries. There are some outliers in the model, such as Qatar, Brunei, and Luxembourg. Qatar's wealth and Brunei's wealth are both largely explained by natural resources such as oil, whereas Luxembourg is a very small country with an unusually large financial sector. Beyond such outliers, the relationship between human capital and the wealth of a nation is strong.

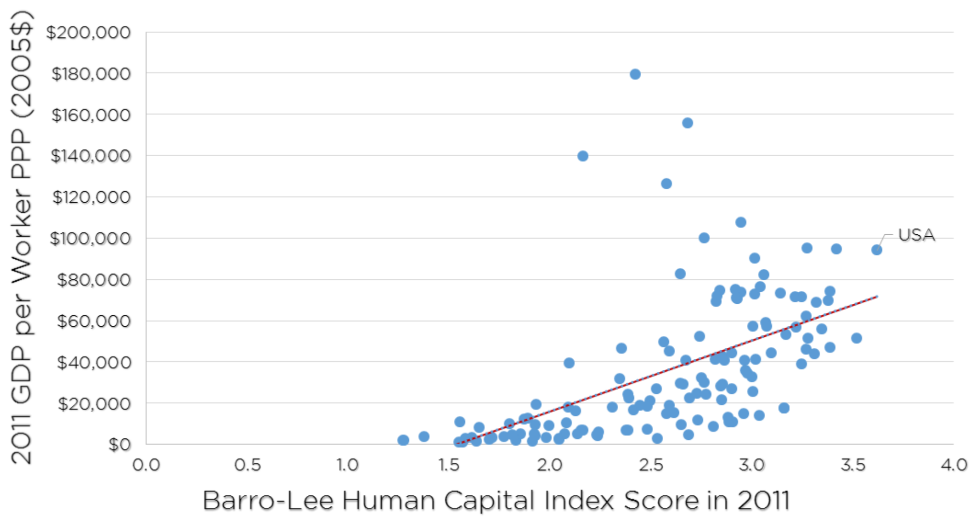


Figure 13: Output per Worker and Human Capital, 2011
Source: Feenstra, Inklaar, and Timmer 2015.

As with the Solow model, sustained, long-run growth in the Lucas and Mankiw–Romer–Weil models is determined exogenously by the rate of technological progress. However, these models can easily be modified so that positive spillover effects of human capital lead to endogenous growth, i.e., a growth rate that is determined within the model itself as opposed to by external factors. Lucas (1988) presents a version of his model that includes such spillover effects, and the results look very similar to the AK model of economic growth and the learning-by-doing model discussed next. Nonetheless, when maintaining the assumption of a constant returns to scale production function and no spillovers, permanent shocks to human capital or its determinants produce level effects rather than growth rate effects.

1.5.3 AK Models

Until now, the models examined here have assumed that long-run growth is determined exogenously by the growth rate of technology. Now we turn to models where growth rates are generated endogenously—that is, within the model itself. Models of this type are known as *endogenous growth models,* and the simplest way to endogenize growth is to eliminate the assumption of diminishing returns to input factors in the production function. The most basic model that does this is the AK model, which assumes constant returns to capital.

The most well-known example of an AK model comes from Barro (1990). In his model, production is ruled by the function

$$y = f(k) = Ak, \tag{14}$$

where A is a constant and y and k are the intensive forms of output and capital. Since A is a constant, its growth rate is 0, that is, $\frac{\dot{A}}{A} = g = 0$. As a result of this assumption, production still exhibits constant, and not increasing, returns to all inputs. In other words, if all inputs in production are doubled, aggregate production is still only doubled because the level of technology remains unchanged. It is easy to show that the marginal product of capital, $f'(k) = A$; so that along a balanced growth path output per worker grows at the rate of consumption growth:

$$\frac{\dot{c}}{c} = \frac{1}{\theta}(A - \rho). \tag{15}$$

Equation (15) is simply a restatement of Equation (9) from the Ramsey–Cass–Koopmans model. Only now the marginal product of capital is equal to A.

In the AK model, permanent increases in the savings rate have growth rate effects. This makes the AK model very similar to a precursor of the Solow model known as the Harrod–Domar growth model. Developed by Harrod (1939) and Domar (1946), this model explains growth as being largely driven by savings and capital formation. At a more granular level, the AK model implies that changes in the coefficient of relative risk aversion, θ, and in the rate of time preference, ρ, also have growth rate effects through their influence on savings behavior, as would a change in the interest rate, which in equilibrium is equal to A.

Limited empirical evidence supports the AK model because countries that save more do not always grow faster (Barro and Sala-i-Martin 2004, C. I. Jones 2001). Another problem is that there is no convergence in the AK model. The assumption of diminishing returns to capital is what causes convergence in the Solow model; so relaxing this assumption eliminates the tendency for countries to converge. This presents a problem for the AK model because in the real world there is significant evidence of convergence across countries, especially in the conditional sense.

Interestingly, there are also no transition dynamics in the AK model. Changes in the model take place instantly; so after experiencing a shock, the economy moves seamlessly to a new balanced growth path. In this sense, the world inside the AK model is always in the long run. The AK model becomes more plausible if capital is thought of broadly as encompassing human as well as physical capital (Barro and Sala-i-Martin 2004, Rebelo 1991). The model should probably be viewed even more broadly than this, however. AK models can be thought of as a general class of models that produce endogenous growth. As shown in the next section on learning-by-doing models, other models in the AK family also lead to endogenous growth without requiring this broad view of capital.

A central result of the AK model is that policies, including regulations, can produce growth rate effects. Any permanent change in savings behavior leads to changes in the growth rate of output per worker without any corresponding increase in technology. Encouraging prudent savings and investment or subsidizing job training and skills development produce hugely beneficial results in this model. The result is interesting because growth that is driven by human capital and investment may not be as disruptive as technology-driven growth. Mandel (2004, xix) refers to this kind of capital-driven growth as "cautious growth," because it is less upsetting and disorderly than "exuberant growth" that is based on disruptive technological innovation. Thus, public policies that encourage investment may generate growth while also maintaining more support from the public.

The AK model is also consistent with empirical evidence that levels of capital do explain a fair amount of growth differences across countries. *Growth accounting* refers to the practice of disentangling the different determinants of growth. Those who have tried to quantify the contributions of capital, labor, and technology in a growth accounting framework do find a relationship between capital accumulation and growth rates. For example, capital formation is thought to explain about 30 percent of differences in cross-country growth rates (Caselli 2005). Solow model proponents might claim that it is only short-run growth and that in the long run only technology will still determine growth rates. Nonetheless, the long run takes a long time to arrive, and given the consistent relationship found between savings, capital formation, and growth, it seems that the basic insight of the AK and Harrod–Domar models—that capital formation matters for growth rates—is correct.

1.5.4 Learning-by-Doing Models

Once economists drop the assumption of diminishing returns to capital, the door is opened to constant returns at the factor level and increasing returns to scale at the aggregate production level. Dropping the assumption of diminishing returns also allows inclusion of other interesting elements into a growth model, such as externalities. In economics, an externality refers to an attribute of a product that is unpriced in the market. This attribute affects third parties who are not participants in an economic exchange. The standard example is pollution, whereby buyers and sellers of a good (e.g., electricity) do not take into account the effect of their actions on others (e.g., breathers of polluted air). Therefore, the cost of the externality is not accounted for in the market price of the good (in this case, electricity).

In growth models, externalities are included by allowing changes in one variable to affect other variables. These can also be thought of as multiplier or spillover effects. One of the first models to include such spillover effects was the *learning-by-doing* model, developed by Frankel (1962) and Arrow (1962). In the learning-by-doing model, the production process, rather than formal education, engenders learning, which leads to increases in productivity. This kind of informal learning process is a second form of knowledge-generating innovation that will be explored here. In the 1930s, engineers noticed that the labor hours required to produce a single airplane fell as the number of airplanes built increased. In other words, as aggregate output grew larger, there was a corresponding increase in productivity that could not be accounted for by the standard inputs of labor and capital. The learning-by-doing model was an attempt to explain this phenomenon by showing how worker productivity increases as a result of experience.

Learning by doing is similar to a phenomenon noticed by the 18th-century economist and political philosopher Adam Smith—that specialization tends to increase productivity. When workers divide production into different tasks and everyone specializes in a particular task, workers become more productive. This was the insight behind Henry Ford's famous moving assembly line for the production of his Model T cars.

In the learning-by-doing model, knowledge generation is a positive externality resulting from capital formation. Each firm faces the production function like the following:

$$Y_i = aI{K_i}^{\alpha}{L_i}^{1-\alpha}, \tag{16}$$

where Y_i, K_i, and L_i represent the firm-specific output, capital, and labor for each firm, i, and a is a constant level of technology that is distinct from the knowledge generated in the capital accumulation process. Each of the n firms in the economy uses an identical fraction, $\frac{K}{n}$, of capital and $\frac{L}{n}$ of labor, and each firm produces $\frac{Y}{n}$ of the economy's output. Additionally, every firm takes the average level of knowledge per worker, I, as given in its production function: $I = \left(\frac{K}{L}\right)^{\gamma}$, meaning knowledge per worker is a function of the level of capital.

There is no subscript with the I term because I is a social variable that is given to everyone, as opposed to a variable that varies across firms. Each firm's investment in capital makes a small contribution to I, but no firm takes its individual contribution into account when deciding how much output to produce. In other words, I is a public good. It represents the stock of nonexcludable and nonrival public knowledge, which is an accidental byproduct of the production process. Once produced, knowledge is immediately and freely available to everyone.

Astute observers will notice that in the special case where $\gamma = 1 - \alpha$, the aggregate production function simplifies to

$$Y = aK, \tag{17}$$

which is the AK model again. The learning-by-doing model is therefore a special case of the AK model; as with the AK model, a permanent change in the savings rate produces growth rate effects. With respect to regulation, this implies that rules that reduce saving and investment lower growth rates.

The learning-by-doing model might also be thought of as embodying a version of the 18th-century economist Adam Smith's famous invisible hand theorem. Smith noticed how individuals acting in their own self-interest can unintentionally bring about results that advance the public interest. Each firm in the learning-by-doing model, acting to maximize its own profits, accidentally contributes to the public good through its contributions to the stock of public knowledge. This accidental byproduct of production increases average knowledge per worker in the economy, thereby increasing output per worker (and hence wages) unintentionally.

There is something very appealing about modeling economic growth as an accidental byproduct of human exchange. As described in the section on remaining puzzles in growth theory, there is still a great deal about growth that economists cannot explain. If growth is truly an unintended consequence of human interaction, this might be why, historically, identifying the causes of economic growth has been so difficult.

The learning-by-doing model also corresponds nicely with many insights from the Austrian School of Economics. Nobel laureate Friedrich A. Hayek (1984 [1968]) describes competition in the marketplace as a “discovery procedure,” whereby firms discover new knowledge as they take part in the competitive market process. Knowledge, once uncovered, spreads throughout the economy by imitation and learning.

Knowledge is also very difficult to measure. If tacit knowledge of the sort developed through learning and experience is a core determinant of growth, it is understandable that economists do not find answers waiting for them in the aggregate statistics.

If growth really is largely an accidental byproduct of the production process, it poses problems for policymakers. Perhaps they might be able to stimulate the capital formation process through tax incentives or subsidies, but it is unlikely that they will be able to replicate the process by which new knowledge is uncovered. That takes competition, experimentation, and trial and error, which together suggest a role for policy in fostering a competitive market. One way to do this would be to remove barriers to competition, such as breaking up monopolies. Removing regulations that impose costs on new entrants in an industry, such as occupational licensing restrictions, is another way to enhance competition.

There may be another role for policy as well. An interesting implication of the learning-by-doing model is that a decentralized free market economy does not produce a Pareto efficient outcome. Pareto efficiency refers to a situation whereby no one can be made better-off without making another person worse off. The model deviates from Pareto efficiency because the private marginal product of capital for each firm and the social marginal product of capital diverge. Thus, each firm underinvests in capital, and I remains below its socially optimal level. Even if firms could come together to agree to each invest in the optimal amount of capital each period, there is a strong incentive for firms to shirk from the agreement because there is an incentive to free ride off the investment efforts of others. The growth rate of output per worker will be below its optimal level for these reasons.

Under laissez-faire, there is too little investment in the learning-by-doing economy relative to an ideal state, so investment could either be subsidized (directly or through tax credits) or policymakers could impose a tax of some kind (preferably of the lump sum form to avoid distortions) on consumption. The key question will be at what level to impose the tax or subsidy. This information may be unknowable. Furthermore, government already taxes and subsidizes countless forms of investment to varying degrees, so it is difficult to know whether there is too much or too little investment at any given time. Additionally, most taxes are not implemented in a lump sum. Thus, taxation efforts to bring social and marginal costs into alignment will produce economic distortions in their own right that must be weighed against any social benefits that result from improving market efficiency.

Although not discussed in detail here, the constant returns to scale models examined thus far assume perfect competition. This assumption means that prices of outputs equal the marginal costs of those outputs, and that factor inputs are paid their marginal products. Because the externality in the learning-by-doing model is completely nonexcludable, that is, firms are unable to exclusively use the knowledge they generate, perfect competition can still be assumed in this model. As shown in the next section, however, when firms are able to internalize some fraction of the knowledge they produce (i.e., they are able to find ways to exclude other firms from using the knowledge), the assumption of perfect competition is no longer tenable.

1.5.5 Models That Endogenize Technological Change

The first generation of endogenous growth models used the capital formation process to explain growth in the economy. Examples include the Harrod–Domar model, the learning-by-doing model, and the AK model. Not surprisingly, some scholars also sought ways to endogenize the mysterious technological change variable in growth models. They sought to model technological advances in society, including the process of generating new knowledge. Unlike the learning-by-doing model, where knowledge creation is an accidental byproduct of production, in this new generation of growth models, scholars explain knowledge creation as a purposeful activity on the part of firms. The most famous model to do this is the model of P. M. Romer (1990), whose work led to a revival of growth theory that came to be known as new growth theory.

Romer's growth model contains two sectors, an approach that can be traced to Uzawa (1964). One sector of the economy produces final goods intended for consumers, whereas a second sector—the research and development (R&D) sector—invents new durable capital goods that are used as inputs in the sector producing final goods. These durable capital goods might be thought of as new ideas, new designs, or new templates that expand society's ability to produce final goods for consumers. In other words, innovation in Romer's model shows up as a wider variety of goods in the marketplace.

In the model, some fraction a_L of the labor force is employed in the R&D sector—these people might be thought of as researchers—while the fraction $1 - a_L$ of the labor force is employed in the production of final goods for consumers. The technology index A represents the number of ideas, templates, or designs produced by the R&D sector, which has the production function

$$\dot{A}_t = \tau[a_L L_t]^{\lambda} {A_t}^{\emptyset}, \tag{18}$$

where τ is a measure of the productivity of researchers, and the parameter λ explains how adding new researchers affects the rate of change in new idea creation. For example, if $\lambda > 1$, there are increasing returns to adding new researchers, so each new researcher makes all existing researchers more productive.

The number of researchers has increased considerably in recent decades, as has the amount spent on R&D, so empirical evidence suggests there are not increasing returns to adding researchers (C. I. Jones 1995). Otherwise, increases would be seen in growth rates of developed countries. The parameter λ is more likely to lie below 1 for this reason and may even turn negative in some cases where the marginal researcher actually undermines the pursuit of knowledge rather than advances it.

Increasing the fraction of the labor force engaged in R&D has a two-pronged effect in the Romer model. First, output immediately falls as workers shift from producing final goods to conducting research. This first effect means there is an immediate drop in output per worker. Next, the growth rate of technology immediately rises as more research is conducted. This second effect produces a positive growth rate effect.[9]

The parameter $\emptyset$ can be positive or negative. It represents how the existing stock of ideas affects the difficulty of discovering new ones. If researchers are "standing on the shoulders of giants," to borrow a phrase from Isaac Newton, then previous discoveries make future discoveries easier and $\emptyset$ is greater than 0. For example, the discovery of electricity certainly facilitated development of the lightbulb and the personal computer. Conversely, if past discoveries make future discoveries harder, $\emptyset$ is negative. This might be the case if all the technological low-hanging fruit has been picked and further innovations require greater and greater investments (Cowen 2011).

As in the Solow model, output per worker grows at the same rate as A. However, in the Romer model the growth rate of technology is not always a constant along a

[9] For a description and an illustration of this point, see Weil (2013, 234).

balanced growth path. Specifically, the growth rate of the growth rate of technology is described by the function

$$\frac{\dot{g_{A_t}}}{g_{A_t}} = \lambda n \; + \; (\emptyset - \; 1) g_{A_t}, \tag{19}$$

where the growth rate of g_A at time t depends on two factors: (a) the labor force (weighted on the basis of the returns to adding new researchers) and (b) the growth rate of technology at time t (weighted by whether having a higher level of technology makes new ideas easier or more difficult to discover). These attributes of the model are a major departure from models discussed heretofore in this essay.

A core reason for this departure is because for any value of $\emptyset$ greater than 1, the growth rate of the economy will be increasing over time. Romer developed his model in part because he thought growth rates were increasing over time and he was seeking a way to explain this phenomenon (P. M. Romer, 1986). Looking back through history—centuries as opposed to years or decades—there is evidence that the growth rate of the developed world may be gradually increasing. This largely follows from the fact that growth was stagnant throughout most of human history.

However, C. I. Jones (1999) points out that any value of $\emptyset$ equal to or greater than 1 produces counterintuitive results with respect to population growth because just increasing the *level* of the labor force results in a growth rate effect. Changes in the growth rate of the labor force result in exponential increases in growth rates. Such a finding is sometimes referred to as a *scale effect* in the literature. A scale effect occurs when there are increasing returns to scale in certain variables in a growth model. Recall that the AK model reviewed earlier assumed constant returns to capital and not increasing returns, so there was no scale effect.

Population scale effects are unlikely to hold in the real world because these kinds of returns are just not seen in the empirical data. For this reason, $\emptyset$ probably lies below 1, which means that the growth rate of output per worker is determined largely by the growth rate (and not the level) of the labor force, *n*.

In the more realistic case where $\emptyset < 1$, changes in the growth *rate* of the labor force produce growth rate effects, and changes in the *level* of the labor force result in level effects. And because growth is primarily determined by the labor force growth rate in the model, and this variable is itself an exogenous variable, models such as the Romer model have come to be known as *semiendogenous* growth models. Technological change has been endogenized in the model, but the growth rate along a balanced growth path is still determined by forces outside the model. Creative destruction and quality ladder models, which are discussed next, are also classes of semiendogenous growth models.

There is another interesting property of the Romer model. When $\emptyset < 1$, growth in per capita income is a stationary, mean-reverting process. This just means that growth rates tend to be fairly constant over time, which is consistent with the empirical data (at least over the past century or two). But if $\emptyset$ lies below, but very close, to 1, the economy will behave *almost as if* $\emptyset = 1$ for long stretches of time. That is, the closer $\emptyset$ is to 1, the longer will be the transition to a new balanced growth path. As a result, short-run changes in growth rates as part of the transition dynamics from a level effect could last for very extended periods of time, perhaps even decades. Thus, the distinction between level effects and growth rate effects may be hard to decipher in the real world (Cochrane 2015). This could explain why factors like saving and capital accumulation appear to influence growth rates in growth accounting exercises, even though models like the Solow model suggest there should be no long-run effect.

The Romer model is unique from the previously reviewed models in another important way. The nature of knowledge in the Romer model is very different from the pure public-good form of knowledge seen in the learning-by-doing model. This is perhaps the most important contribution of P. M. Romer (1990). In his model, as in the real world, firms deliberately invest in new technologies, so there must be some financial incentive for them to do so. Recall that in the learning-by-doing model, new knowledge is a public good that is instantly available to all other firms. It had to be accidental in the model to explain why firms would create new knowledge at all.

The fact that firms do invest in R&D in the real world suggests at least some fraction of new knowledge is not a pure public good. Otherwise, every other firm would get a free ride off the knowledge-creation efforts of others, and there would be no incentive to invest in R&D. Firms must be able to keep some new discoveries to themselves—at least for a period of time—and this provides sufficient incentive to partially overcome the problem of free riders. The excludable component of knowledge might be the result of secrecy, or it could follow from deliberate policy interventions, such as patent protections.

That some knowledge is excludable undermines a fundamental assumption of the models examined thus far—the assumption of perfect competition. If firms are engaging in large up-front R&D expenditures to generate discoveries, pricing cannot possibly equal marginal cost. The first unit of production will be very expensive when large R&D investments are required to produce it, but costs will fall dramatically with each additional unit produced. Consider a pharmaceutical where the first pill costs a billion dollars to produce but the second pill costs just a penny. If all firms set prices equal to marginal cost, any firm that engages in R&D will quickly go out of business in this kind of market.

Many growth theorists have thus switched to models of monopolistic competition of the sort developed by Chamberlin (1933) and Dixit and Stiglitz (1977), and this switch in turn implies that there are two kinds of distortions in the economy, that is, deviations from Pareto efficiency. First, if firms set prices above marginal cost (which must be true to explain how firms that engage in R&D stay in business) then these firms must have some monopoly power. When a firm has monopoly power this means it will restrict output to maximize profits, and aggregate output lies below the socially optimal level. Second, too little output also implies too little demand for inputs, such as R&D, and because R&D drives growth in the Romer model, growth rates will be below the socially optimal level.

These findings suggest several possible roles for government. First, there may be a role in designing intellectual property protections. Without adequate protections, firms may lack the incentive to invest enough in new technologies because they cannot internalize the benefits of these new technologies. The more nonexcludable an invention is, the more likely there is a role for such protections to play. But there is a tradeoff to consider between incentivizing investment and the losses to society from monopoly restrictions on output. Furthermore, recall that patent protections can also result in a tragedy of the anticommons if patents are over-issued.

There may also be a role for government in subsidizing R&D. If firms are underinvesting in R&D, the government could encourage it through direct subsidies or through tax credits. Subsidies to final-goods producers would accomplish the same end by increasing demand for R&D inputs. However, before the government rushes in and begins subsidizing R&D, there are several factors to consider. When subsidies are financed by any means other than lump sum taxes, the taxes will create distortions that must be weighed against the benefits of the subsidies. Second, investment in R&D must actually be productive. Historically, governments have not had a better track record than the private sector at picking investment projects (OECD 2003). This suggests that R&D tax credits that give private firms control over the selection of projects may be more effective than having government invest directly in new research.

Aside from the scale effects that can arise in the Romer model, the model also has some other unrealistic features. Countries that remain mired in poverty often have high population growth rates, whereas many rich countries have low or stagnant population growth rates. Population growth does not appear to be a sufficient condition for economic growth. At the global level, population growth and economic growth move more closely together (Barro and Sala-i-Martin 2004; Kremer 1993b), but this may be because higher incomes allow more people to be sustained on Earth, rather than the other way around (that is, a bigger population causing higher incomes). Or, it may be that human capital–adjusted population growth is what really matters.

A final lesson from the Romer model relates to free trade. There are clearly benefits from engaging with greater numbers of people. Expanding the network of people that firms interact with means expanding the network of ideas. A larger market also implies greater demand for new ideas, which incentivizes idea creation. Adam Smith suggested in his book *The Wealth of Nations* that incomes in countries are dependent on the size of the market. A larger market allows for more specialization not just in physical production but in idea production as well.

1.5.6 Creative Destruction and Quality Ladder Models

In the Romer model, innovation shows up as changes in the number of products available. Economists have developed other classes of models to account for innovation in the form of quality improvements that occur over time. The most important contributions in this literature are from Grossman and Helpman (1991), who developed a theory of “quality ladders” in economic growth, and Aghion and Howitt (1992), whose “creative destruction” model of growth explains obsolescence (i.e., the process of new products replacing old ones over time).

Quality ladder models treat products as if they are on a race up a ladder. Each time an entrepreneur develops an improvement, the product moves up one rung on the ladder. Creative destruction models, named after the term coined by Austrian-born economist Joseph Schumpeter (1942), incorporate how old products become obsolete and disappear from the market over time as new and better products are developed.

As in the Romer model, firms have some monopoly power in both models. Several externalities are also present. First, when an innovation occurs, consumers pay the same price for a better product. The result is a spillover benefit to consumers as products move up each rung of the quality ladder. Second, producer profits decline for rivals when a firm innovates and takes the business of its competitors. This externality is known as *business stealing*, and it creates a misalignment of incentives because the benefits of innovation are permanent for consumers but only temporary for producers.

At first glance, business stealing looks like only a pecuniary externality—that is, an externality resulting from a price change that is a pure transfer from one party to another—but in fact other spillover effects arise. If businesses are not fully compensated when other innovations build on the quality improvements they developed, firms will be discouraged from investing in an optimal level of R&D.

Here is how this can happen. Consider Isaac Newton and Gottfried Wilhelm Leibniz, who are both credited with having developed calculus. Neither of these individuals was compensated during their lifetimes for the millions of ways in which calculus is put to use today. In an ideal world, these individuals would have been compensated, because so many aspects of modern life would not be possible without these innovations from the past. Without a compensation scheme for past inventors, expect there to be too little incentive to innovate. Even worse, competition will reduce the expected duration of monopoly rents accrued from innovation—so the more firms that are competing, the more inventors will be discouraged from inventing and the more firms will be discouraged from spending on R&D.

Interestingly, there can also be too much R&D in creative destruction and quality ladder models. Much like there is social waste when firms compete for transfers from the government (an activity known as *rent-seeking*), there can also be social waste if competition drives firms to overinvest in R&D as they seek to capture the monopoly profits of their rivals.

Creative destruction and quality ladder models again demonstrate the importance of finding the right balance between intellectual property rights protections and monopoly power. Too little intellectual property protection could mean that firms will not invest in R&D enough, whereas too much protection could encourage wasteful competition for transfers. One solution that has been proposed is to force innovators to compensate their immediate predecessors (Barro and Sala-i-Martin 2004). However, quality improvements are notoriously difficult to measure, and even if they could be measured perfectly, it is hard to know which ideas formed the basis for succeeding innovations. A predecessor payment scheme is likely to prove impossible to implement. Unfortunately, these problems have no simple solutions.

1.5.7 Technology Transfer

In the economic growth models explored here thus far, innovation is driven by the creation of *new* knowledge or by the creation of new products or improvements in the quality of existing products. Innovation that is based on the creation of new knowledge comes from formal education and training, on-the-job experience, and R&D. However, firms and individuals also have the ability to imitate innovations created elsewhere. When businesses are not operating along the *technological frontier* (i.e., using the latest and best technology), they have the option of either creating innovations themselves or imitating the innovative practices of others. The process of transferring technological knowledge through imitation is known as *technology transfer*.

Usually, technology transfer models are applied to countries. For example, middle and lower-middle-income countries such as China and India may be able to grow quickly by simply adopting the practices and technologies generated elsewhere, such as in Western Europe and the United States. The same phenomenon applies to individuals and firms, however. Firms can be divided into *leading firms* that operate along the technological frontier, with *follower firms* that lag. Firms operating at the technological frontier have no choice but to innovate by creating new knowledge if they want to grow. But follower firms have the option of imitating the technologies developed in leading firms if they do not want to create their own innovations. And just as there are costs associated with invention, there are costs associated with imitation, although in general imitation will prove easier than innovation. The costs of imitation include the time and effort it takes to copy a product design, to adjust a product design to fit the preferences of different consumers, and to adopt the modes of production of one industry or region to new ones.

There are also likely to be diminishing returns to imitation. Some innovations are very easy to copy. On one hand, it might be easy to imitate a dating website and create a similar website targeted to a new demographic. On the other hand, supply chain management techniques in factories might be much harder to copy and may not have the same payoffs if workers respond differently in different industries. In other words, some technologies have limited applicability outside a single narrow use, or they may simply be too costly to copy because of their complexity.

Just as diminishing returns to capital create convergence among economies in the Solow model, diminishing returns to innovation create convergence as well. The further a country or firm is from the technological frontier, the faster that country or firm will grow. This type of growth might be deemed *technological catch-up growth*, to be distinguished from the traditional *capital-based catch-up growth* found in the Solow model. In essence, imitation is another way to increase g in the Solow model. The rate of technological progress will be fast in firms or countries that begin from a low level of technology.

Human capital also plays a role in technology transfer. Some technologies are relatively easy to learn, such as operating a soft-serve ice cream machine. Other technologies take years of schooling to master, such as computer programming or statistical analysis. So, again, human capital has a tendency to augment labor in ways that facilitate growth.

Because knowledge has attributes of a pure public good, technologies invented in one industry can have spillover effects in other industries, as in the learning-by-doing model. At the country level, technological advances in developed countries act like a form of foreign aid to follower countries. Cell phones are an obvious example of a technology that has raised living standards for some of the poorest individuals in the world. Such a result was probably not the intention of those who created cell phones, but it is nonetheless a social benefit that should be recognized.

If policymakers in rich countries think they have a duty to assist individuals in poor countries—and many people would agree they do—one of the best things they can do is to spur innovation at home. For regulators, this means that the costs of blocking new innovations, and the benefits of nurturing new innovations, extend beyond a country's borders. Considering that it is poor people in the developing world who stand to gain the most from technological advances in the present, this benefit provides a strong reason for allowing socially beneficial innovations to arrive as soon as possible.

Policymakers should also seek to extend intellectual property rights abroad through treaties or to find ways to encourage more foreign direct investment at home and abroad. When property rights are protected, firms in leading countries may find it easier to protect their innovations abroad. Or firms might protect their investments by purchasing the foreign firms that use their inventions. Not only does this help secure property rights, it speeds the process of technology diffusion. Workers in follower countries learn by following the practices of leading countries by working directly for them.

Regulatory complexity also can discourage technology transfer. If a country's regulatory code is too complex, it creates a hurdle to investing. Investors are already more likely to invest in their own countries because of a home bias effect. If they do not understand a foreign country's legal code or think it will be arbitrarily enforced, investment will be discouraged.

Some follower countries or firms may see a short-term benefit from using the inventions of others without compensating them. In the long run, however, this practice is likely to discourage foreign direct investment, slow technology transfer, and lead to reciprocal stealing when followers eventually become leaders themselves.

1.5.8 Elasticity of Substitution

The production function that forms the foundation for the economic growth models discussed thus far is the famous Cobb–Douglas production function. One of the useful features of this function, indeed one of the main reasons it was first conceived of by Cobb and Douglas (1928), is its assumption of constant unit elasticity of substitution between capital and labor.

The elasticity of substitution describes the change in relative demand for capital and labor when there is a change in the relative cost of these inputs. Unit elasticity means that for every 1 percent rise in the ratio of prices between labor and capital, $\frac{w}{r}$, where w represents the wage rate for labor and r is the rental rate on capital, there is a corresponding 1 percent rise in the ratio of aggregate capital to labor demanded in the economy, $\frac{K}{L}$. Unit elasticity of substitution between capital and labor is a convenient assumption because it simplifies the math in the model, but it is not likely to be true most of the time in the real world.

In the early 1960s, a new class of production functions was developed to relax the unit elasticity assumption (Arrow et al. 1961). Production functions of this class are known as constant elasticity of substitution production functions, and an example of such a function is

$$Y = A[\alpha K^{\vartheta} + (1 - \alpha)L^{\vartheta}]^{\frac{1}{\vartheta}}, \tag{20}$$

where $\vartheta < 1$ and the elasticity of substitution between capital and labor is defined as $\sigma = 1/(1-\vartheta)$. The parameter α is a share parameter between 0 and 1. In the special case where $\vartheta = 0$, there is the Cobb–Douglas production function, where $\sigma = 1$. This result can be shown by taking the limit of Equation (20) as $\vartheta \to 0$. Similarly, as $\vartheta \to -\infty$, the production function approaches the fixed proportions production function made famous by Leontief (1941).

An interesting result to emerge from constant elasticity of substitution production functions is that growth models based on these functions produce endogenous growth when ϑ lies between 0 and 1. In such cases, there is high substitutability between capital and labor (i.e., σ is greater than 1). When this happens, the property of diminishing returns to capital per worker gradually vanishes as capital per worker asymptotically approaches infinity (Barro and Sala-i-Martin 2004). With a high enough savings rate, changes in the savings rate produce growth rate effects rather than level effects. But this property of the model also violates fundamental assumptions of most growth models, known as the Inada conditions (Inada 1963). These conditions state that as $k \to 0$, $f'(k) \to \infty$, and as $k \to \infty$, $f'(k) \to 0$. In other words, this assumption states that the marginal product of capital per worker diminishes as capital per worker grows, and the marginal product of capital per worker grows as capital per worker shrinks.

When ϑ lies between 0 and 1, meaning there is a high elasticity of substitution, $f'(k)$ approaches a positive constant as $k \to \infty$. Recent empirical estimates suggest σ is likely to be less than 1 (Chirinko 2008), so the Inada conditions likely hold in the real world, which is also consistent with conventional wisdom. But there are reasons to believe these conditions may be fragile. For example, the theory of wealth inequality proposed by Piketty (2014) depends on an elasticity of substitution that is greater than 1 (Rognlie 2015). Piketty argues that wealth inequality increases in an economy where $r > g$, a condition known as the transversality condition. That $r > g$ is a standard assumption in growth models and is believed to be true in the real world, at least in healthy economies (Abel et al. 1989). Piketty seems to believe that wealth inequality is a natural outgrowth of a capitalist economy, but another possibility is that wealth inequality is an outgrowth of an economy with high substitutability between capital and labor.

Many forces might bring about this situation. Technology can facilitate substitution between capital and labor. For example, supermarket cashiers can be replaced by self-service cash registers and tollbooth operators can be replaced by E-ZPass technology. This creates an interesting bridge between the *wealth* inequality theory of Piketty and the *income* inequality theory of Cowen (2013). Cowen's story of inequality is based largely on the idea that individuals with job skills that are complementary to new technologies are likely to earn high incomes in the future, whereas individuals whose skills are substitutes for new technologies will earn lower incomes. Thus, rather than inequality being a natural outcome of a capitalist economy, it might be a natural outgrowth of improvements in technology in the 21st century.

Regulators may not have much control over the long-run progress of technology, but they can certainly influence the ability of firms to substitute capital and labor. For example, they could lower the elasticity of substitution by making it more difficult to fire workers. Empowering labor unions might have this result. However, such protections could backfire if firms are discouraged from hiring workers in the first place due to high labor costs. Firms might respond to labor protections by simply going straight to replacing workers with machines.

Regulators clearly influence the relative prices that affect aggregate demand for capital and labor as well. Everything from workplace safety regulations to rules mandating that employers provide such benefits as maternity leave or health insurance will influence the price of labor and encourage capital substitution. Raising the minimum hourly wage is also likely to speed up the process of automatizing human labor. Of course, capital is taxed and regulated to varying degrees as well, which encourages substitution towards labor. Which production input is given preferable treatment in the aggregate is unclear, although in recent decades the share of national income going to labor has declined and the share going to capital has risen. Whether this is a direct result of policy is unclear.

A key question is whether the elasticity of substitution is indeed rising. Rognlie (2015) makes the important point that it is *net* elasticity of substitution—that is, the elasticity of substitution *after depreciation* is taken into account—that matters for inequality purposes. Unfortunately, most estimates in the literature are estimates of gross elasticity. Rognlie assumes that net elasticity must be lower than gross elasticity, but his argument hinges on the assumption that capital depreciates whereas labor does not.

If we assume that labor's share of income is augmented by human capital, as seems likely, it is not clear whether the assumption that labor does not depreciate is realistic. For example, unemployment can cause the erosion of worker skills over time.

Technology can also erode worker skills. For example, those who know how to repair typewriters will probably have trouble finding a job that requires this skill today.

Some human capital models of growth even include a depreciation factor. The model of Mankiw, Romer, and Weil (1992) assumes that human capital and physical capital depreciate at the same rate. Therefore, a shock could influence growth rates through effects on the depreciation factor. For example if technology is causing labor to depreciate faster—that is, job skills are eroding more quickly over time because of technology—the net elasticity of substitution could actually be *above* the empirical estimates of gross elasticity found in the literature. If the net elasticity of substitution is high enough, the Inada conditions could be violated, and inequality could rise.

Whether this situation has been true in the past, is true now, or will be true in the future is unclear. However, it does suggest that inequality may be another factor influenced by public policy. If technology is raising the elasticity of substitution and regulation has a tendency to favor capital over labor, regulations could be contributing to wealth inequality through the mechanisms described here.

1.6 Remaining Growth Puzzles: The Roles of Institutions and Population

1.6.1 The Role of Institutions

Despite significant advancements in the theory and empirics of economic growth, many mysteries remain. Macroeconomics is a notoriously difficult discipline because it seeks to explain so much complexity with so few variables and relatively few data. The process of uncovering the sources of growth has largely been about experimenting with different variables that for theoretical reasons seem important. Over time economists have been able to weed out the variables that appear to be less important and identify those with more explanatory power.

To see how far economic growth theory has come, consider that more than 50 years ago the economist Nicholas Kaldor (1961, 178) highlighted six “stylized facts” about economic growth. Stylized facts are accepted empirical observations that researchers seek to explain. Kaldor’s six facts centered on the contribution of capital accumulation to economic growth. When Kaldor wrote his paper, savings and capital accumulation were thought to be the most important contributors to growth. Economists have since learned that these factors can explain only a fraction of the growth differentials observed across countries.

In 2010, economists C.I. Jones and P. M. Romer (2010) updated Kaldor’s list, highlighting the new stylized facts that require explaining by the next generation of growth economists. Their list includes ideas, institutions, population, and human capital. Economists are now in general agreement that these four factors matter for economic growth, but the micro-level mechanisms by which these factors influence macro-level growth remain poorly understood. Furthermore, it is not clear if these inputs themselves are what fundamentally drives growth or whether these variables are correlated with or caused by something more fundamental.

In recent decades, much attention in the economic growth and development literature has focused on the role of institutions. When economists talk about institutions, they are referring to the rules that constrain human economic and social behavior. The late Nobel laureate Douglass North (1991, 98), whose major contribution was to make institutions more central to economic theory, defined institutions as “rules of the game.” Rules bind human behavior and shape the incentives people face in their economic lives. Some rules are formal, such as laws written by legislatures or regulations written by administrative agencies. Other rules are informal, such as social and cultural norms that pressure us to be kind to our neighbors or to tell the truth. Institutions are extremely important for economic development, so important that some prominent economists call them a “fundamental cause of long-run growth” (Acemoglu, Johnson, and Robinson 2005).

Because institutions are so important, economists sometimes add an index of social infrastructure to the production functions in their models to estimate the contributions of institutions to economic growth. Such indices attempt to measure things such as the strength of property rights in a country, the rule of law, credible contract arrangements, the level of corruption, social levels of trust, and the degree of rent-seeking in society.[10] It turns out social levels of trust can explain some differences in growth rates (Zak and Knack 2001). Trust is also negatively correlated with regulation (Aghion et al. 2010). Of course, all such indicators are very difficult to measure, so economists must be creative when they collect data on social infrastructure measures.

A difficulty that arises with using indices of social infrastructure is that these measures tend to be correlated with other factors, such as culture or religion. Max Weber (1930 [1904]), the German sociologist, thought the wealth of nations was determined in large part by Protestant values. An economist regressing output per worker on an index of

---

[10] For examples of this approach, see Hall and Jones (1999) and Knack and Keefer (1995).

social infrastructure in Switzerland might find a strong correlation between institutions and growth. But the fundamental cause of economic growth might simply be the Swiss culture that produces both strong institutions and steady growth. Most likely, there is a feedback loop between culture and institutions whereby culture shapes institutions and institutions shape culture (Alesina and Giuliano 2015).

There is a similar debate about the role that geography plays in economic growth. It turns out that latitude is highly correlated with GDP per capita (Bloom and Sachs 1998). For hundreds of years, observers have noticed that countries near the equator tend to be less developed than countries farther from the equator.[11] Jared Diamond (1997) is one of the best-known scholars to argue for the importance of geography in economic development. He contends that geographical happenstance determined mightily which groups were able to adopt certain technologies, develop agriculture, and generate immunities from diseases. See Figure 14.

[11] See, for example, Montesquieu ([1748] 1989), who was an early observer of this fact.

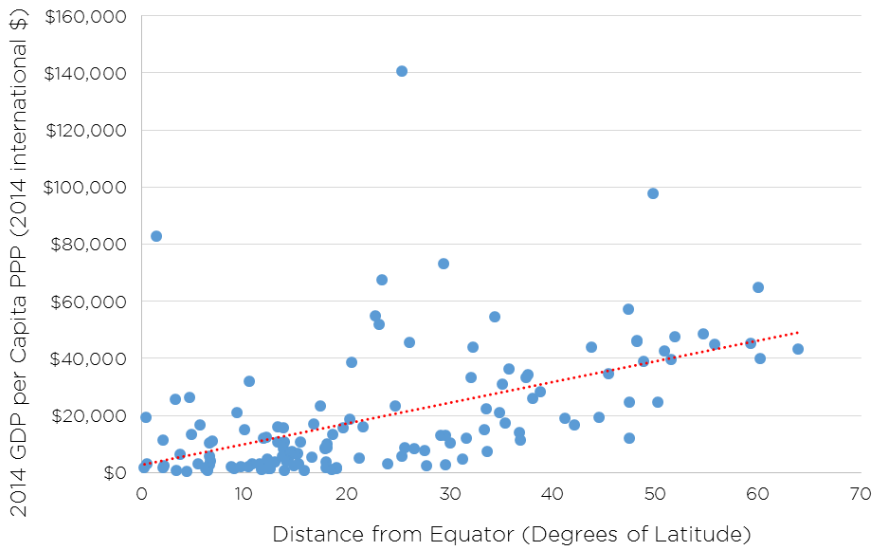


Figure 14: GDP Per Capita and Latitude, 2014
Sources: World Bank Development Indicators and Hall and Jones (1999).

Acemoglu, Johnson, and Robinson (2001) think that colonialism explains the link between institutions and geography. They argue that Europeans during the colonial period set up more *inclusive* institutions in places that had lower disease rates and a more hospitable environment for settlement. Alternatively, in those places that were less hospitable for economic development, such as South America, Europeans created *extractive* institutions to expropriate wealth from those regions—regions where they never planned to settle permanently. These inclusive and extractive institutions persisted long after colonists had left, either facilitating long-run growth or entrenching a culture of rent-seeking and corruption.

The debate about institutions and geography is far from settled. Some economists, such as Rodrik, Subramanian, and Trebbi (2004), argue that the effects of geography are weak, and they still operate primarily through institutional quality. Other economists, such as Sachs (2003), point to examples where geography has had direct effects on income per capita without any link to institutions. For example, many debilitating diseases, such as malaria, are far more common in areas near the equator. Distance from a coast also matters. Indeed, Adam Smith noted that cities with access to water tended to have higher living standards, which he attributed to access to global markets.

In all likelihood, geography does have direct effects on growth while also contributing to growth through institutions. But even more fundamental forces could be driving institutions. Comin, Easterly, and Gong (2010) point out that technology adoption practices from as far back as 1000 BC are strongly correlated with income per capita and technology adoption practices today. Using migration data to control for the historical places of origin of modern populations, these authors find that certain peoples, for cultural, historical, or perhaps even genetic reasons, have been more open to adopting new technologies. For whatever reason, these tendencies seem to have persisted for hundreds, even thousands, of years.

This observation suggests that something much deeper may be going on than just cultural forces. Spolaore and Wacziarg (2013) point to some of the transmission mechanisms by which our ancestors might have passed on traits that support economic development, some of which are biological. G. Jones (2012) shows how cognitive skill is associated with technology diffusion, which comports with the idea that human capital and technology transfer are closely related. Perhaps intelligence even influenced human migration patterns thousands of years ago, leading to a connection between intelligence and geography. In fact, time preference and geography appear to be linked (Galor and Özak 2014). Furthermore, patience contributes to savings and capital formation, and high-intelligence people tend to be more patient (G. Jones 2015). At some point, economists may need to endogenize the parameters ρ and θ for underlying genetic characteristics.

Another possibility is that better nutrition improves cognitive skills, which facilitates human capital development and, by extension, spurs growth. Better health is strongly associated with higher income (Smith 1999), and negative health outcomes, such as child mortality, tend to fall with higher income. This is demonstrated in Figure 15. Perhaps one reason humans lived in grinding poverty for so many thousands of years was simply because they were not healthy enough to build strong institutions or to devote time to inventing. With adequate nourishment came opportunities to invent, to develop skills, and to build social infrastructure. This complex history suggests that health, geography, culture, cognitive skills, institutions, patience, ideas, and growth are all linked.

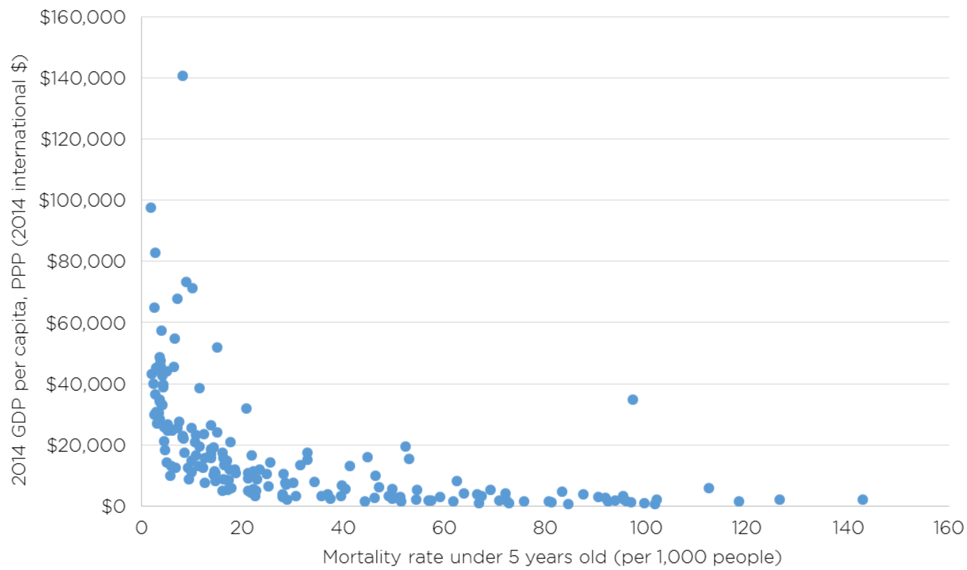


Figure 15: GDP Per Capita and Child Mortality, 2014
Source: World Bank Development Indicators

Unfortunately, the lessons for regulators here are far from clear. Respecting private property rights, enforcing contracts, and resisting the temptation to expropriate wealth all facilitate good institutions and improve growth. It certainly cannot hurt for policymakers to aspire to these goals. But it is also not clear that in places that lack good institutions the solution is simply to plant new institutions in place of old ones. Underlying forces, like the roots of a weed, may prevent healthy institutions from arising in the first place. Simply pulling the weed from the dirt will not change the underlying fundamentals that caused the weed to grow in the first place.

A better option is for regulators to embrace a culture that respects new technologies. Whatever the underlying causes, cultures that are more open to new technologies tend to thrive. Regulators should resist the demands of interest groups that are displaced by new technologies and should work to explain to the public the benefits of new technologies, even when the benefits carry certain risks. Regulators who encourage safe experimentation with new technologies will promote growth more than those who act as gatekeepers to technological change.

1.6.2 The Role of Population

As far back as the late 18th century (Malthus 1798), debate has raged among economists about whether a rising population raises living standards or promotes poverty. The Solow model takes the extreme position that faster population growth lowers the level of output per person, whereas other models, like the Romer model, go to the opposite extreme.

The truth probably lies between these two positions, but where along the continuum the world lies is unclear. Empirically, the relationship between population and growth is vague. Many countries with large populations have historically grown slowly, whereas economies with slow population growth often grow quickly. Only at the global level is the relationship between population and output per worker fairly stable, and even this relationship might be misleading if is the correlation exists because it is rising income that allows more people to inhabit the earth.

An additional bias permeates the literature on the economics of ideas: only good ideas result from having more people. But of course, people come up with terrible ideas as well. Sometimes bad ideas can take civilization down wrong turns, thereby leading to incredible destruction and misery. Communism is a particularly salient example of a bad idea that has captured the imaginations of the people and destroyed millions of lives.

Sometimes the problem is not whether there are too many or too few people but instead whether a fixed number of people are allocated optimally across professions. Many people are not able to be as productive as they could be because they cannot, for one reason or another, enter the profession where they would be most productive. One reason might be discrimination (Hsieh et al. 2013). Policies that limit the free movement of people can also lead to a suboptimal allocation of people in the labor force. In cases where freedom of movement is necessarily limited, such as internationally, public policies that promote trade might improve the allocation of resources without requiring people to move.

Microlevel distortions of these kinds can actually lower total factor productivity at the macrolevel, thereby lowering growth rates (C. I. Jones 2013). The channels by which micromisallocations lead to macroeconomic inefficiencies remain poorly understood, but some kind of spillover effect is an obvious possible explanation. The best option for policymakers is to, wherever possible, avoid creating misallocations, discourage discriminatory practices, and encourage trade across regions where movement is necessarily limited.

1.7 Conclusion

Many economists would love to claim that all a society needs to spur faster economic growth is more investment in R&D and more immigration. Indeed, some economists do make such proclamations. But nothing is so simple when it comes to economic growth. This essay has classified the different kinds of growth effects to present a framework for understanding the growth implications of public policy. But even the distinction between growth rate effects and level effects is not so clear in the real world. Most likely there are diminishing returns to many inputs in production, including labor, R&D, capital, and human capital. As a result, it is very hard to increase growth rates sustainably over the long term. This may be a core reason why very long run growth rates have held constant in higher income countries over the past century and a half.

Nonetheless, there are some takeaways from the growth models surveyed in this essay. The first lesson is twofold: (a) innovation matters and (b) a culture that embraces innovation should be promoted to a great extent. This essay has identified a number of sources of innovation, including the creation of new products, formal education and job training, informal tacit learning through experience and specialization, quality improvements, and knowledge transfer through imitation. Regulators should seek to nurture and promote these sources of innovation, avoiding encouraging fear of new technology, so as to support a culture of progress through technological change.

Another lesson of this essay is that the cumulative effect of all policies is likely to matter most for economic growth. A single policy by itself probably will not have growth rate effects unless it encourages or discourages the adoption of a general purpose technology (GPT). Together, however, all policies working in concert can interact in ways that influence growth. This fact is particularly important because the regulatory code in the United States has consistently grown over time. As Figure 16 demonstrates, regulation in this country has been growing, both in the number of pages in the *Code of Federal Regulations* and the number of regulatory restrictions contained in the code. The regulatory system has become more complex over time, which means there are also likely to be more significant unintended consequences of policy. Going forward, policymakers need to address an important problem: how to control the growth of regulation.

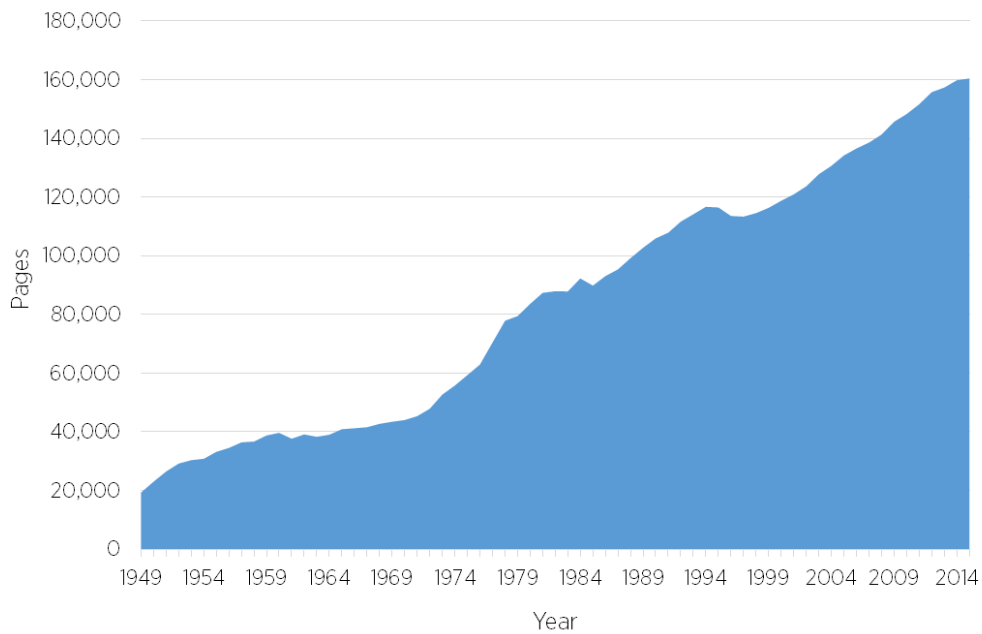
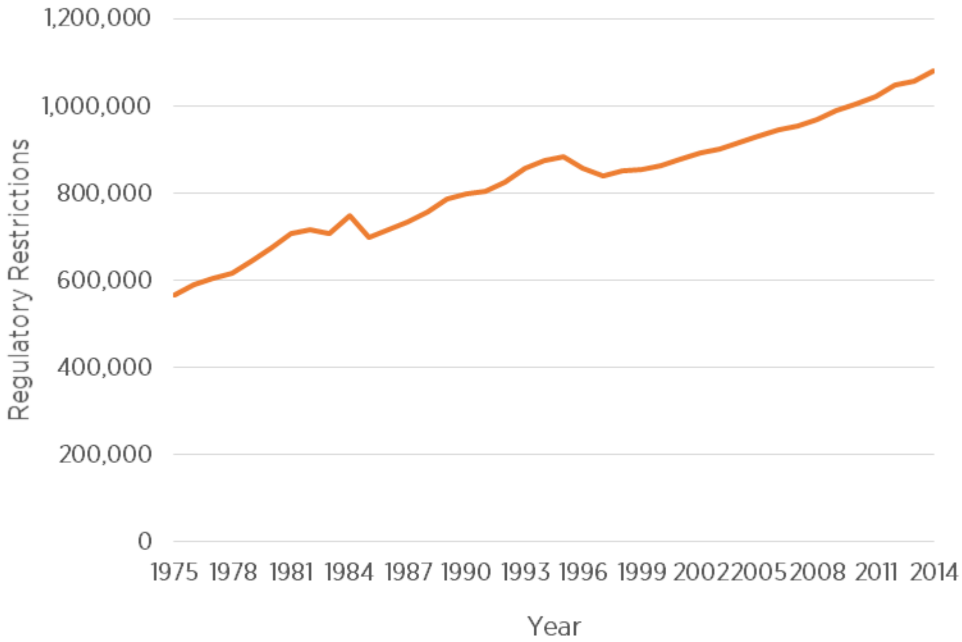


Figure 16: Growth of US Regulation Over Time
a. *US Code of Federal Regulations*, adjusted page count, 1949–2015
Note: Page count adjusted to exclude materials unrelated to regulation and to account for stylistic changes in the code over time.
Source: Dawson and Seater 2013.



b. Regulatory restrictions in the *US Code of Federal Regulations*, 1975–2014
Source: Al-Ubaydli and McLaughlin 2015.

Regulators must also approach potential GPTs with great care. They should seek to create a climate whereby discovery of new GPTs is more likely to occur and where the development and diffusion of potential GPTs is not stifled. Nanotechnology and biotechnology are two possibilities for what the next GPTs might be. Although GPTs can also create disruptions to particular subsets of the population, the long-run benefits generally vastly outweigh the costs, and the benefits should be explained to the public whenever possible.

Capital accumulation is another important contributor to economic growth. Thus, there is wisdom to the idea that a penny saved is a penny earned. However, there is such a thing as too much of a good thing. If the optimal level of capital is exceeded, such as when the elasticity of substitution between capital and labor rises high enough, there may be reasons to limit capital accumulation. Regulators should take great care to avoid unfairly favoring investment relative to consumption. More investment will generally lead to the kind of cautious growth that is more palatable to the public than disruptive technological innovations. However, ultimately the levels of consumption determine standards of living, so encouraging consumption can also make sense.

Human capital accumulation clearly matters, but job skills can come in many forms—from formal training to on-the-job experience. Obtaining both forms of skills requires tradeoffs. Competition policy is also important. Regulations that stifle competition by preventing new firms from entering an industry, setting maximum or minimum pricing, restricting quantities, or granting arbitrary monopoly privileges to firms or industries stifle the learning-by-doing process, which is important to economic progress. Trade also encourages growth by expanding the size of markets, encouraging specialization, and transferring knowledge from one part of the globe to another. Intellectual property protections are also desirable within reason.

New technologies and global trade will not benefit everyone equally, however. There are important distributional consequences to consider. Over time, it may be becoming easier to substitute people with machines, and businesses will be attracted to those areas where labor is cheapest. At the very least, regulators should seek not to exacerbate income inequality that arises from these forces. They should not give an unfair advantage to capital over labor and they should not create incentives whereby the best way to get ahead is through political connections and rent-seeking as opposed to serving customers.

Even if any single regulation is unlikely to produce effects that show up in the GDP data, regulations should nurture and not stifle the factors known to be important for growth. These factors include productivity, investment, competition, human capital, and innovation.

After 60 years of modern economic growth theory, our ability to predict the growth implications of public policies may be better than many people realize. Yet given the vast uncertainty confronting regulators, a strong sense of humility is required. Critical puzzles in growth theory remain to be fully explained, and it is probably easier for regulators to slow economic growth than it is for them to accelerate it. Perhaps the framework presented here will help bridge theory and practice by providing a theoretical foundation for regulators. Such a foundation could move theory closer to solving the remaining puzzles while improving the design and implementation of the regulations that in practice govern our lives.

References

Abel, Andrew B., N. Gregory Mankiw, Lawrence H. Summers, and Richard J. Zeckhauser. 1989. “Assessing Dynamic Efficiency: Theory and Evidence.” *Review of Economic Studies* 56 (1): 1–20.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. “The Colonial Origins of Comparative Development: An Empirical Investigation.” *American Economic Review* 91 (5): 1369–1401.

———. 2005. “Institutions as a Fundamental Cause of Long-Run Growth.” In *Handessay of Economic Growth*, vol. 1A, edited by Philippe Aghion and Steven N. Durlauf, 386–472. Amsterdam: Elsevier.

Aghion, Philippe, Yann Algan, Pierre Cahuc, and Andrei Shleifer. 2010. “Regulation and Distrust.” *Quarterly Journal of Economics* 125 (3): 1015–49.

Aghion, Philippe, and Peter Howitt.1992. “A Model of Growth through Creative Destruction.” *Econometrica* 60 (2): 323–51.

Alesina, Alberto, and Paola Giuliano. 2015. “Culture and Institutions.” *Journal of Economic Literature* 53 (4): 898–944.

Al-Ubaydli, Omar, and Patrick A. McLaughlin. 2015. “RegData: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations, 1997–2012.” *Regulation & Governance*.

Arrow, Kenneth J. 1962. “The Economic Implications of Learning by Doing.” *Review of Economic Studies* 29 (3): 155–73.

Arrow, Kenneth J., Hollis B. Chenery, Bagicha S. Minhas, and Robert M. Solow. 1961. “Capital-Labor Substitution and Economic Efficiency.” *Review of Economics and Statistics* 43 (3): 225–50.

Barro, Robert J. 1990. “Government Spending in a Simple Model of Endogenous Growth.” *Journal of Political Economy* 98 (5): S103–26.

Barro, Robert J., and Xavier I. Sala-i-Martin. 2004. *Economic Growth*, 2nd ed. Cambridge, MA: MIT Press.

Bloom, David E., and Jeffrey D. Sachs. 1998. "Geography, Demography, and Economic Growth in Africa." *Brookings Papers on Economic Activity* 29 (2): 207–95.

Boldrin, Michele, and David K. Levine. 2008. *Against Intellectual Monopoly*. New York: Cambridge University Press.

Bresnahan, Timothy F., and Manuel Trajtenberg. 1995. "General Purpose Technologies 'Engines of Growth'?" *Journal of Econometrics* 65 (1): 83–108.

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." In *Handessay of Economic Growth*, vol. 1A, edited by Philippe Aghion and Steven N. Durlauf, 679–741. Amsterdam: North-Holland.

Cass, David. 1965. "Optimum Growth in an Aggregative Model of Capital Accumulation." *Review of Economic Studies* 32 (3): 233–40.

Chamberlin, Edward Hastings. 1933. *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Chirinko, Robert S. 2008. "σ: The Long and Short of It." *Journal of Macroeconomics* 30 (2): 671–86.

Cobb, Charles W., and Paul H. Douglas. 1928. "A Theory of Production."*American Economic Review* 18 (1): 139–65.

Cochrane, John. 2015. "Smith Meet Jones." *The Grumpy Economist*. December 3. http://johnhcochrane.blogspot.ch/2015/12/smith-meet-jones.html.

Comin, Diego, William Easterly, and Erick Gong. 2010. "Was the Wealth of Nations Determined in 1000 BC?" *American Economic Journal: Macroeconomics* 2 (3): 65–97.

Conway, Paul, Donato de Rosa, Giuseppe Nicoletti, and Faye Steiner. 2006. "Regulation, Competition and Productivity Convergence" (OECD Economics Department Working Papers. Organisation for Economic Co-operation and Development, Paris).

Cowen, Tyler. 2011. *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better*. New York: Dutton.

———. 2013. *Average Is Over: Powering America Beyond the Age of the Great Stagnation*. New York: Dutton.

Dawson, John W., and John J. Seater. 2013. "Federal Regulation and Aggregate Economic Growth." *Journal of Economic Growth* 18 (2): 137–77.

DeLong, Brad. 2006. "Brad DeLong's Solow Growth Model Scenario Generator Spreadsheet." http://delong.typepad.com/print/20060829_Solow_growth.xls.

Diamond, Jared. 1997. *Guns, Germs, and Steel: The Fates of Human Societies*. New York: Norton.

Dixit, Avinash K., and Joseph Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67 (3): 297–308.

Domar, Evsey D. 1946. "Capital Expansion, Rate of Growth, and Employment." *Econometrica* 14 (2): 137–47.

Dourado, Eli, and Alex Tabarrok. 2015. "Public Choice Perspectives on Intellectual Property." *Public Choice* 163 (1): 129–51.

Égert, Balázs. (2016). "Regulation, Institutions, and Productivity: New Macroeconomic Evidence from OECD Countries." *American Economic Review* 106 (5): 109–13.

Erlandsen, Espen, and Jens Lundsgaard. 2007. "How Regulatory Reforms in Sweden Have Boosted Productivity" (OECD Economics Department Working Paper, Organisation for Economic Co-operation and Development, Paris).

Feenstra, Robert C., Robert Inklaar, and Marcel Timmer. (2013). *The Next Generation of the Penn World Table* (Working Paper No. 19255). National Bureau of Economic Research.

———. 2015. "The Next Generation of the Penn World Table." *American Economic Review* 105 (10): 3150–182.

Frankel, Marvin. 1962. "The Production Function in Allocation and Growth: A Synthesis." *American Economic Review* 52 (5): 996–1022.

Galor, Oded, and Ömer Özak. 2014. "The Agricultural Origins of Time Preference" (Working Paper No. 20438, National Bureau of Economic Research, Cambridge, MA).

Garicano, Luis, Claire LeLarge, and John Van Reenen. 2013. “Firm Size Distortions and the Productivity Distribution: Evidence from France” (Working Paper No. 18841, National Bureau of Economic Research, Cambridge, MA).

Grossman, Gene M., and Elhanan Helpman. 1991. “Quality Ladders in the Theory of Growth.” *Review of Economic Studies* 58 (1): 43–61.

Hall, Robert E., and Charles I. Jones. 1999. “Why Do Some Countries Produce So Much More Output per Worker than Others?” *Quarterly Journal of Economics* 114 (1): 83–116.

Harrod, Roy F. 1939. “An Essay in Dynamic Theory.” *Economic Journal* 49 (193): 14–33.

Hayek, Friedrick A. 1984 [1968]. “Competition as a Discovery Procedure.” In *The Essence of Hayek*, edited by Chiaki Nishiyama and Kurt R. Leube, 254–65. Stanford, CA: Hoover Institution Press.

Heller, Michael A. 1998. “The Tragedy of the Anticommons: Property in the Transition from Marx to Markets.” *Harvard Law Review* 111 (3): 621–88.

Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow. 2013. “The Allocation of Talent and U.S. Economic Growth” (Working Paper No. 18693, National Bureau of Economic Research, Cambridge, MA.)

Inada, Ken-ichi. 1963. “On a Two-Sector Model of Economic Growth: Comments and a Generalization.” *Review of Economic Studies* 30 (2): 119–27.

Jones, Charles I. 1995. “Time Series Tests of Endogenous Growth Models.” *Quarterly Journal of Economics* 110 (2): 495–525.

———. 1999. “Growth: With or Without Scale Effects?” *American Economic Review* 89 (2): 139–44.

———. 2000. “A Note on the Closed-Form Solution of the Solow Model.” http://web.stanford.edu/~chadj/closedform.pdf.

———. 2001. *Introduction to Economic Growth*, 2nd ed. New York: Norton.

———. 2013. “Misallocation, Economic Growth, and Input-Output Economics.” In *Advances in Economics and Econometrics*, Tenth World Congress, vol. II, Applied Economics, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel. Cambridge, UK: Cambridge University Press.

———. 2015. “The Facts of Economic Growth” (Working Paper No. 21142, National Bureau of Economic Research, Cambridge, MA).

Jones, Charles I. and Paul M. Romer. 2010. “The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital.” *American Economic Journal: Macroeconomics* 2 (1): 224–245.

Jones, Garett. 2012. “Cognitive Skill and Technology Diffusion: An Empirical Test.” *Economic Systems* 36 (3): 444–60.

———. 2015. *Hive Mind: How Your Nation’s IQ Matters So Much More than Your Own*. Stanford, CA: Stanford University Press.

Jovanovic, Boyan, and Peter L. Rousseau. 2005. “General Purpose Technologies.” In *Handessay of Economic Growth*, vol. 1B, edited by Philippe Aghion and Steven N. Durlauf, 1181–1224. New York: Elsevier.

Kaldor, Nicholas. 1961. “Capital Accumulation and Economic Growth.” In *The Theory of Capital*, 177–222. New York: MacMillan.

Knack, Stephen, and Philip Keefer. 1995. “Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures.” *Economics and Politics* 7 (3): 207–27.

Koopmans, Tjalling C. 1965. “On the Concept of Optimal Economic Growth” (Discussion Paper No. 163, Cowels Foundation for Research in Economics at Yale University, New Haven, CT).

Kremer, Michael. 1993a. “The O-Ring Theory of Economic Development.” *Quarterly Journal of Economics* 108 (3): 551–75.

———. 1993b. “Population Growth and Technological Change: One Million B.C. to 1990.” *Quarterly Journal of Economics* 108 (3): 681–716.

Krugman, Paul. 1980. “Scale Economies, Product Differentiation, and the Pattern of Trade.” *American Economic Review* 70 (5): 950–59.

Leontief, Wassily W. 1941. *The Structure of American Economy, 1919–1929*. Cambridge, MA: Harvard University Press.

Lipsey, Richard G. Cliff Bekar, and Kenneth Carlaw 2005. *Economic Transformations: General Purpose Technologies and Long Term Economic Growth*. New York: Oxford University Press.

Lipsey, Richard G., Cliff Bekar, and Kenneth Carlaw. 1998. "What Requires Explanation?" In *General Purpose Technologies and Economic Growth*, edited by Elhanan Helpman, 15–54. Cambridge, MA: MIT Press.

Lucas, Robert E., Jr. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22 (1): 3–42.

Malthus, Thomas Robert. 1798. *An Essay on the Principle of Population*. London: J. Johnson, in St. Paul's Church-yard.

Mandel, Michael J. 2004. *Rational Exuberance: Silencing the Enemies of Growth and Why the Future Is Better than You Think*. New York: HarperCollins.

Mandel, Michael, and Diana G. Carew. 2013. "Regulatory Improvement Commission: A Politically-Viable Approach to U.S. Regulatory Reform" (Policy Memo, Progressive Policy Institute, Washington, DC).

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107 (2): 407–37.

Montesquieu, Charles de Secondat. 1989 [1748]. Montesquieu: *The Spirit of the Laws*, translated and edited by Anne M. Cohler, Basia Carolyn Miller, and Harold Samuel Stone. New York: Cambridge University Press.

Nicoletti, Giuseppe, and Stefano Scarpetta. 2003. "Regulation, Productivity, and Growth: OECD Evidence" (Policy Research Working Paper WPS2944. World Bank, Washington, DC).

North, Douglass C. 1991. "Institutions." *Journal of Economic Perspectives* 5 (1): 97–112.

OECD (Organisation for Economic Co-operation and Development). 2003. *The Sources of Economic Growth in OECD Countries*. Paris: OECD.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.

Ramsey, F. P. 1928. "A Mathematical Theory of Saving." *Economic Journal* 38 (152): 543–59.

Rebelo, Sergio. 1991. "Long-Run Policy Analysis and Long-Run Growth." *Journal of Political Economy* 99 (3): 500–21.

Rodrik, Dani, Arvind Subramanian, and Francesco Trebbi. 2004. "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development." *Journal of Economic Growth* 9 (2): 131–65.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise in the Net Capital Share" (Brookings Paper on Economic Activity. Brookings Institution, Washington, DC).

Romer, David. 2011. *Advanced Macroeconomics*, 4th ed. New York: McGraw-Hill.

Romer, Paul M. 1986. "Increasing Returns and Long-Run Growth." *Journal of Political Economy* 94 (5): 1002–37.

———. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71–102.

Sachs, Jeffrey D. 2003. "Institutions Don't Rule: Direct Effects of Geography on Per Capita Income" (NBER Working Paper No. 9490, National Bureau of Economic Research, Cambridge, MA).

Schumpeter, Joseph. 1942. *Capitalism, Socialism and Democracy*. New York: Harper & Brothers.

Smith, James P. 1999. "Healthy Bodies and Thick Wallets: The Dual Relation between Health and Economic Status." *Journal of Economic Perspectives* 13 (2): 145–66.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70 (1): 65–94.

Spolaore, Enrico, and Romain Wacziarg. 2013. "How Deep Are the Roots of Economic Development?" *Journal of Economic Literature* 51 (2): 325–69.

Swan, T. W. 1956. "Economic Growth and Capital Accumulation." *Economic Record* 32 (2): 334–61.

Thierer, Adam. 2016. *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom* (revised and expanded edition). Arlington, VA: Mercatus Center at George Mason University.

Uzawa, Hirofumi. 1964. "Optimal Growth in a Two-Sector Model of Capital Accumulation." *Review of Economic Studies* 31 (1): 1–24.

Weber, Max. (1930 [1904]). *The Protestant Ethic and the Spirit of Capitalism*. London and New York: Routledge.

Weil, David N. 2013. *Economic Growth*, 3rd ed. London: Routledge.

Yellen, Janet. 1984. "Efficiency Wage Models of Unemployment." *American Economic Review* 74 (2): 200-205.

Zak, Paul J., and Stephen Knack. 2001. "Trust and Growth." *Economic Journal* 111 (470): 295–321.

2 The Social Discount Rate: A Baseline Approach

2.1 Introduction

One of the most important decisions in public policy analysis is selection of the social discount rate (SDR). Economists apply an SDR in benefit-cost analysis (BCA) when the benefits and costs of social projects accrue across different time spans. The SDR is the interest rate used to calculate the present value of intertemporally distributed benefit and cost flows, so that these flows can be compared to one another as apples to apples.

A higher SDR will mean a lower present value of future benefit and cost flows. This effect is most profound for benefits and costs that occur in the distant future. Even minor adjustments to the SDR can have huge effects on present-value calculations because of the power of compounding. The more benefits and costs are separated by time, the more sensitive the sign of the net benefits calculation will be to selection of the SDR. Thus, the SDR will matter most in cases where (1) large upfront costs produce flows of benefits in the distant future or (2) large upfront benefits produce flows of costs in the distant future. Some of the most pressing issues of our time, including mitigation of global climate change and the growing national debt, have long-run intergenerational consequences. As a result, the SDR is closely tied to questions of ethics, such as how much society should care—and be willing to spend—for the welfare of future generations.

To understand the practical relevance of the selection of the SDR, Table 4 presents estimates of the social cost of carbon (SCC) for the year 2020 using three SDRs. The SCC is an estimate of the monetized damages associated with the emission of an additional ton of carbon dioxide into the atmosphere. The numbers were calculated by averaging the output of three models, known as integrated assessment models, which estimate the effects of climate change. These estimates are inputs into the calculation of benefits of carbon dioxide reductions in government BCAs, where benefits are the avoided damages resulting from carbon dioxide emissions.

Table 4: Estimated Average Social Cost of Carbon Dioxide at Various Social Discount Rates
Source: Interagency Working Group on Social Cost of Carbon (2013)

Year	**2007$ per metric ton of carbon dioxide**		
	5.0%	**3.0%**	**2.5%**
2020	$12	$43	$65

As should be clear from the numbers presented in Table 4, selection of the SDR is critical. At a discount rate of 2.5 percent, the SCC is more than five times higher than it is at a 5 percent discount rate. An SCC that is five times higher means it is efficient to spend five times more on climate change mitigation strategies. One reason for this large difference is that the integrated assessment models used to calculate the SCC run simulations out as far as the year 2300, and damages in the year 2300 will be extremely sensitive to selection of the SDR. In fact, the SCC is so sensitive to the SDR that raising the SDR by a mere 0.5 percentage points, from 2.5 to 3.0 percent, causes the SCC estimate to fall by one-third.

There is no consensus among economists as to what the appropriate SDR should be. Economist Martin Weitzman (2001) surveyed more than 2,100 economists and found a wide variation of opinion about the discount rate to be used for projects designed to mitigate global climate change. Responses to his survey ranged from negative discount rates to discount rates over 20 percent, with a mean response of about 4 percent. Even among only the most renowned scholars in the profession, Weitzman found disagreement, although less disagreement than across the larger sample of the profession. More recently, Drupp et al. (2015) surveyed economists about their preferred long-term discount rate and found considerable variation in opinion.

Most economists agree that the SDR should be above zero, although exceptions exist. The surveys of Drupp et al. (2015) and Weitzman (2001) reveal that a minority of economists favors a zero SDR (and sometimes even a negative SDR) for long-term social projects. Cowen (2007), for example, argues for a zero SDR as a means to protect the welfare of future generations. According to Cowen (2004), zero discounting is not necessarily a recipe for government intervention in the economy, but instead it would necessitate policies that foster a rapid rate of economic growth. In earlier work, Cowen and Parfit (1992) also show that discounting is not necessary to address many of the concerns cited by economists who say discounting is required.[12]

More commonly, economists argue for a positive SDR when analyzing potential social projects. Although the connection is not generally made explicit, many rationales for discounting in social project analysis appear to have been carried over from neoclassical models of optimal consumption behavior and economic growth. For example, the rationales for discounting provided by the federal government can be explained with a simple formula known as the Ramsey equation, which states that individual agents that optimize utility discount future consumption flows as a result of impatience and because consumption generates diminishing utility as consumption rises. These rationales form a large part of the theoretical case for discounting in social project analysis.

However, Ramsey discounting presents a problem for use in BCA for several reasons. As this paper will show, the logic underlying the equation contradicts basic tenets of BCA, such as the assumption that one additional dollar is valued equally by everyone and that a benefit's value should be determined by its recipients. Furthermore, ethics-based value judgments are often necessary to calibrate the parameters in the Ramsey equation, making disagreement among economists virtually inevitable.

[12] Viscusi (2007) argues that certain anomalies can arise in economic analyses that do not use discounting. For example, total benefits and costs could approach infinite values as they extend into the infinite future, because costs could recur year after year in perpetuity or because willingness to pay tends to rise with income. However, limiting the time horizon is a much easier and more practical way to solve this problem. In fact, with an infinite time horizon, even using a positive discount rate produces an infinite flow of much smaller costs and benefits in net-present-value terms. Thus, a seeming anomaly in analysis can be resolved without discounting.

In the sections that follow, this paper describes some of the problems associated with Ramsey discounting and why its use in BCA is troublesome. In place of Ramsey discounting, an alternative approach to discounting based on the time value of money is recommended. Here, discounting is used as a way to compare alternative projects to a baseline, a method which avoids many of the ethics controversies that can arise with Ramsey discounting. The discounting rule that comes closest to this time-value-of-money approach is known as the *weighted average approach*. This paper recommends a modified version of the weighted average rule with a pure rate of time preference set to zero, and concludes with suggestions for how policymakers might update government guidelines on regulatory analysis in light of this new information.

2.2 Ramsey Rationales for Discounting

The current federal guidelines for regulatory analysis, presented in *Circular A-4* (OMB 2003), offer three rationales for discounting: (1) positive time preference, (2) diminishing marginal utility of consumption, and (3) opportunity cost of capital. Perhaps the easiest way to explain these rationales is with the Ramsey formula associated with the Ramsey–Cass–Koopmans economic growth model. The Office of Management and Budget (OMB) does not explicitly mention the formula in its guidance; however, the formula is a convenient vehicle for explaining the agency's reasoning, and many economists believe the Ramsey formula provides a useful framework for discounting (Arrow et al. 2012). The formula is written as

$$r(t) = \rho + \theta \frac{\dot{c}(t)}{c(t)}, \tag{21}$$

where the interest rate r at time t is equal to the representative agent's pure rate of time preference ρ plus the product of the consumption elasticity of marginal utility θ and the instantaneous growth rate of consumption $(\dot{c}(t))/(c(t))$ at time t. This equation is an equilibrium condition, where the marginal cost of using capital—the real interest rate r—equals the marginal physical product of capital, $f'(k)$, that is, the opportunity cost of capital, in equilibrium. The Ramsey rule recommends setting the SDR equal to r in this equation.

Let us consider these parameters one at a time as a basis for discounting. An SDR that is based on a positive pure rate of time preference parameter, ρ, assumes that utility of future citizens is worth less than utility of present citizens because society is impatient. People would rather consume today than in the future, and therefore they lose utility by waiting for future benefits to arrive. Since everyone in society is impatient, any aggregation method will necessarily generate a positive time-preference measure for society as a whole.

This reasoning requires that benefits and costs be valued from the perspective of those alive at the time a policy is implemented. True, individuals in the present are likely to be impatient when it comes to waiting for future benefits to arrive. However, individuals in the future will have no such impatience because the current generation's future will be their present. This problem is relevant across not only generations but also shorter time spans because people are continually being born and dying, and our future selves might be willing to pay different values for benefits than would our present selves.

One of the greatest achievements of BCA is its ability to value benefits and costs in terms of what the individuals who receive those benefits and costs would be willing to pay for them (or pay to avoid them). Discounting on the basis of time preference violates this premise because it forces analysts to value benefits and costs in terms of how present members of society value them, rather than how individuals in the future—those who will receive the benefits or entail the costs—value them.

Aside from this problem, serious questions of ethics arise from discounting when it is based on time preference. With the time preference rationale for discounting, the utility of people in the future is valued less than the utility of people today simply because of the passage of time. The British economist Frank P. Ramsey (1928, 543) himself suggested this is "ethically indefensible." Although individuals are impatient—something not disputed here—society as a whole does not share this impatience. Assuming the *social* discount rate behaves like an *individual's* discount rate is a fallacy of composition.

The parameter θ in the Ramsey equation describes how steeply marginal utility declines as consumption increases and it is also a measure of the representative agent's degree of risk aversion. Thus, it is sometimes referred to as the *coefficient of relative risk aversion*. Because future generations will presumably be richer than people alive today as a result of economic growth, the marginal utility generated by an additional unit of consumption is expected to be lower for people in the future than for those living now. The idea here is that analysts discount dollars rather than utility because dollars generate lower utility for richer citizens. Many environmentalists who have ethics concerns about discounting based on time preference still find this premise for discounting acceptable.[13]

Even with general agreement that there is diminishing marginal utility *within* individuals (ruling out the possibility of utility monsters with *increasing* marginal utility of consumption), neoclassical economists generally contend that interpersonal comparisons of utility *between* individuals are meaningless because utility is an ordinal concept. Economists can say that the second apple matters less to John than the first, but economists cannot say that the second apple to Susan matters less than the first apple to John. Utilities are rankings, and comparing one person's set of rankings to another person's provides little useful information. This concept has been acknowledged in neoclassical microeconomic theory going back at least to the British economist Lionel Robbins (1938).

Applying any value to the wealth effect term $\theta(\dot{c}(t))/(c(t))$ in the Ramsey equation is akin to applying distributional weights in cross-sectional BCA. Distributional weights are values assigned to benefits and costs accruing to certain subgroups in the population (e.g., the poor) to raise or lower the value of those benefits and costs in an analysis. Distributional weighting is a controversial practice for several reasons. It allows for potential Pareto improvements simply through the redistribution of wealth via pure transfers. This result leads to the conclusion that efficiency improvements are possible simply by equalizing the level of wealth across the population. Distributional weights are

[13] See Gollier (2013) for examples of how risk aversion and consumption smoothing across generations form a basis for discounting in policies that address global climate change.

also hard to defend because they are inherently arbitrary and require singling out certain groups for special treatment.

In general, neoclassical economists do not weight benefits and costs that occur at the same time according to income. They should not do so across time, either, unless they are willing to make interpersonal comparisons of utility and single out groups for special treatment.[14] If important intertemporal distributional issues are raised in an analysis, a more transparent way to present this information is to present undiscounted flows of benefits and costs in a separate distributional analysis that highlights impacts on subpopulations of interest, such as the present and future generations. Then decision makers who are more accountable to the public can decide what is a fair and equitable intertemporal distribution of wealth.

Aggregation problems are also associated with discounting dollars based on θ. Recall that in the Ramsey–Cass–Koopmans growth model, the Ramsey equation describes the behavior of a single representative agent. To calculate how the marginal utility of society changes in response to wealth increases, every individual's preferences must be aggregated to form a set of social preferences. This calculation is easy enough in a growth model where everyone has identical preferences. In the real world, however, preferences are far from identical. Nobel laureate Kenneth Arrow (1950) proved that it is impossible to convert ranked individual preferences into a function describing communitywide aggregate social preferences without the possibility of certain paradoxes

[14] Note, however, that it is *real* dollars that concerns us here, so making adjustments to dollars based on changes in the price level across time is reasonable.

arising. There is no guarantee that, in the aggregate, people's preferences will be as well-behaved as preferences are at the individual level.

Economist Stephen Marglin (1963, 109) describes the problem in more detail:

> There remains the problem of aggregating the time-preference maps of individuals for collective decisions into a single social time-preference map. This problem is a special case of the general problem of aggregating individual utility functions into a social welfare function. The more general problem has been investigated by Kenneth Arrow and others, and Arrow's negative conclusion that "democratic" aggregation is impossible unless we restrict the allowable class of individual preference functions or abandon one or more intuitively appealing axioms about preferences is too familiar to require elaboration.

What is the time-preference map Marglin speaks of? Elsewhere in the article he states that "the term 'time-preference map' should be carefully distinguished from the loosely used expression 'time preference.' The time-preference map refers to the entire functional relationship between, on the one hand, individual marginal rates of substitution of consumption at one time for consumption at another and, on the other hand, the levels of consumption at all different times" (Marglin 1963, 95-96).

Those familiar with the Ramsey–Cass–Koopmans growth model will know that $1/\theta$ is equal to the elasticity of substitution between consumption at any two points in time. So the time-preference map described by Marglin represents the marginal rates of substitution at all possible levels of consumption across time. By contrast, what Marglin calls the "loosely used expression 'time preference'" refers either to a particular marginal rate of substitution at a specific level of consumption or to the parameter ρ in the Ramsey equation.

The Arrow-Pratt coefficient of relative risk aversion,

$$\theta = \frac{-cu''(c)}{u'(c)}, \tag{22}$$

makes explicit that the coefficient θ is a function of utility (Simon and Blume 1994, 363). Because utility is a component of individual discount rates, the aggregation of individual discount rates to form an SDR can lead to the paradoxes identified by Arrow (1950). This result need not always occur, but it is a problem that cannot be ruled out. Indeed, one of the reasons for BCA's popularity is certainly that it replaced the need for aggregated social welfare functions that were unable to resolve the problems identified by Arrow.

These findings suggest that the appropriate value for θ might be zero in BCA. A value of zero for θ also implies that one additional dollar of consumption generates a constant and equal level of utility for all individuals. This result is intuitively pleasing for BCA analysts because this assumption—that the marginal dollar is equally valuable to everyone—is also a foundational assumption of BCA. Indeed, any value of θ other than zero creates inconsistency in BCA unless economists start applying distributional weights to benefit and cost flows.

Some economists sweep away these aggregation problems by assuming a modified "social" version of the Ramsey Rule, such as

$$r = \delta + \eta g, \tag{23}$$

where δ is society's rate of time preference, η is a measure of society's inequality and risk aversion, and g represents the growth rate of the economy. Gollier (2013) provides an excellent survey of these approaches.

Such an approach is sensible if we allow ourselves to think that society has a discount rate function and engages in optimizing risk and consumption-smoothing behavior just as individuals do. But if analysts are willing to go down this road, they might also be forced to consider whether BCA itself should be discarded and replaced with a social welfare function that policymakers seek to maximize. BCA has developed over time in part to avoid the subjective value judgments that analysts necessarily embed in the selection of a social welfare function.

Other economists recommend selecting an SDR on the basis of observable market interest rates.[15] These observable market interest rates could correspond to the left-hand side of the Ramsey equation (the opportunity cost of capital), allowing economists to forgo any consideration of the right-hand-side values. Acknowledging that capital has an opportunity cost means considering that it can have alternative uses, but this is also precisely what analysts seek to determine when they conduct BCA. They consider multiple alternative uses of public resources to identify where resources produce the highest social returns (i.e., the most efficient use of resources). As will be shown in the sections that follow, the cost of capital turns out to be more of a *method of obtaining* an SDR, but than it is a *rationale* for discounting. The SDR, when applied in this paper's recommended manner, focuses on only one alternative use of resources, rather than all alternative uses. This single use may or may not represent the opportunity cost of resources.

---

[15] In the past, this approach was sometimes referred to as the *descriptive* approach to discounting. See Arrow et al. (1996).

2.3 Gamma Discounting

Before discussing the recommended discounting approach of this paper, it is worth mentioning a slightly different approach called *gamma discounting*. Gamma discounting was developed by Martin Weitzman (1998, 2001), who noticed that the distribution of economists' beliefs about the proper SDR (as measured by surveys) resembled a gamma probability distribution. He points out that the discount factor e^{-rt} can be viewed as a special case in a gamma distribution of the form $g(r) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$. The first terms of the expression on the right-hand side of the equation represent a probability weight to be applied to a set of uncertain discount factors. Because economists cannot agree on the appropriate discount rate to use in project analysis, each discount factor has a certain probability of being the correct discount factor. Critically, Weitzman (1998, 2001) recommends taking the expected value of the probability-weighted discount *factors* (as opposed to discount *rates*). Doing so generates a certainty-equivalent average discount rate that declines over the term structure toward the lowest possible discount rate. This result flows from Jensen's inequality, which applies to concave expected net present value functions.

Gollier (2004) responds to Weitzman by pointing out that if one instead takes the expected net *future* value of social projects and evaluates payoffs in terms of net future value rather than net present value, one finds that uncertainty causes the discount rate to *rise* to its *highest* possible value. This puzzle, known as the Weitzman–Gollier puzzle, persisted until Gollier and Weitzman (2010) together showed that, after adjustments are made for the risk-aversion and consumption-optimization tendencies of individual agents, the discount rate declines over the time horizon, seeming to solve the contradiction.

However, if gamma discounting is to be applied to BCA, and if it is to avoid the Weitzman–Gollier puzzle, economists must again make certain assumptions about the risk-aversion and consumption-optimization tendencies of society. Whereas it is reasonable to make these assumptions about individual agents, especially in an optimal growth model, it is much less reasonable to assume the same about society as a whole. Again, the line between individual and societal preferences is blurred.

Even if economists are inclined to believe that society indeed has such tendencies as risk aversion, inequality aversion, and the like, they will likely never form a consensus as to how to calibrate these parameters in their models. Such calibrations depend on ethical questions more than empirical ones, and given what is at stake, such as how much to invest in mitigating global climate change, there will likely never be consensus on the matter. Furthermore, gamma discounting tends to require survey results to calibrate the gamma distribution function, and survey results can be unreliable. First, who should be surveyed? As seen earlier, the distribution of SDRs preferred by a select group of economists in Weitzman's 2001 paper differed significantly from the broader profession's views. Further, given the ethical nature of the question, perhaps a broader range of professions beyond just economists should be included in any survey. There is also the problem of time inconsistency. If this year's survey suggests that 3 percent is the appropriate SDR but next year's respondents decide 7 percent is the appropriate rate, should next year's policymakers terminate all projects that started this year but fail a benefit-cost test next year?

In the following section, an alternative discounting approach that bases the SDR on the logic underlying time value of money is presented. This approach produces intuitively pleasing results in that it is consistent for use in BCA and does not run into the same thorny aggregation and ethical controversies so common in Ramsey and Gamma discounting approaches.

2.4 An Alternative Approach Based on The Time Value of Money

2.4.1 The Time Value of Money

The core reason for discounting future cash flows in finance is the *time value of money* (TVM). TVM means that income earned sooner is preferable to income earned later. A leading money and banking textbook puts it this way: "If you are promised $1 of cash flow, for certain, ten years from now, this dollar would not be as valuable to you as $1 is today because, if you had the $1 today, you could invest it and end up with more than $1 in ten years" (Mishkin 2016, 112). In other words, people prefer to receive money earlier rather than later because money can be used to generate even more money over time.

Note that TVM does not say that money today provides more utility than money in the future. Nor does it say that people are impatient, so utility matters less in the future. TVM says only that more money is preferred to less and that getting money sooner rather than later is preferable because it results in more money in total (and presumably by extension more utility) at the end of the period.

Analysts discount future cash flows according to TVM because there is an implicit alternative asset or account in which money can be invested. Often this investment is thought to be a risk-free asset,[16] although it could be any alternative investment instrument. The role of the risk-free asset here is critical. First, it is an *implicit alternative investment*. Although analysts generally do not explicitly consider the cash flows from this alternative investment (they only discount the cash flows of the investment under consideration), it is possible to do so. Thus, discounting cash flows is about comparing the cash flows from one investment to the cash flows generated by an implicit alternative investment. Discounting is a rule of thumb that makes comparison easier. If the net present value of an investment is negative that means the implicit alternative asset (e.g., a risk-free asset) generates larger returns than the investment being evaluated. A negative net present value only means the return is negative relative to the implicit alternative investment. It does not necessarily mean returns are negative relative to no investment.

Next, the risk-free asset is a *displaced investment* whenever a decision is made to embark on a new investment. Thus, the discount rate under TVM accounts for how resources would have been used if an investment had not been made. Putting money in the implicit alternative investment (e.g., a risk-free asset) is the *baseline* investment scenario. The baseline scenario is not zero investment because it would be foolish to give

[16] The rate of return on a risk-free asset is not the only discount rate used in financial analysis, of course. Financial analysts may use hurdle rates, the weighted average cost of capital, or the capital asset pricing model to identify a proper discount rate. Often the discount rate is raised above the rate of return on the risk-free asset to account for the riskiness of an investment (as is done in the capital asset pricing model). This kind of risk adjustment is also possible for an SDR. Alternatively, adjustments for risk could be made in an uncertainty analysis of benefits and costs in BCA.

up free interest at no risk. Thus, discounting is a rule for comparing cash flows from different investment opportunities to a *baseline alternative investment scenario*. In this way, the practice of discounting can be thought of as an acknowledgment that the world being evaluated is not static. Failing to discount is to assume no investment in the absence of the investment under consideration.

Contrary to what is sometimes claimed, the opportunity cost of putting money in a financial investment is not necessarily investing in an implicit alternative asset like a risk-free asset. The definition of opportunity cost is the value of the *next-best* alternative forgone when undertaking an activity. The cash flows from a risk-free asset are only one alternative that should be considered when investing, and this alternative may or may not be the next-best alternative to the social project being considered. Thus, discounting under TVM only compares investments to one alternative—the most likely alternative. To determine the opportunity cost of an investment, a wide variety of alternatives must be considered.

2.4.2 Discounting as a Form of Baseline Analysis

One of the main differences between financial analysis of cash flows and BCA of social projects is that the two techniques compare investments against different baselines. Identifying the implicit alternative investment is relatively easy for financial analysts because market rates of return are usually available for risk-free assets or for other assets with comparable risk to the investment under consideration (i.e., whatever other asset is the *most likely* alternative investment). By contrast, the baseline scenario in BCA is the state of the world as it would have evolved in the absence of a social project. To know what this state of affairs looks like, one needs to know how resources would have been consumed and invested—and what the social returns on those uses would have been—in the absence of a policy. This is much harder to estimate than the baseline in financial analysis.

The discounting rule that perhaps comes closest to identifying the relationship between the baseline scenario and the SDR is known as the *weighted average* approach to discounting. Economists who endorse this approach divide resources displaced by social projects into two categories: resources that would have been consumed and resources that would have been invested. These economists use a consumption rate of discount based on the pure rate of time preference, and they use a higher investment rate of discount based on the cost of capital. The SDR is a weighted average of those two rates, weighted based on the mix of resources (consumption and investment) displaced by social projects. In other words, the SDR is weighted based on the sources of funding for social projects.

Harberger and Jenkins (2015) advocate this kind of approach. They note that when considering the weights to use in an SDR, “The profile of net benefits and costs that we analyze is really the difference between two moving pictures—one showing how the economy would evolve ‘with’ our project or program, and the other tracing a similar evolution ‘without’ it” (Harberger and Jenkins 2015, 8). Note the similarity of language between Harberger and Jenkins’s description of an SDR under the weighted average approach and OMB’s description of the baseline as “the best assessment of the way the world would look absent the proposed action” (OMB 2003, 15). The purpose of the SDR under the weighted average approach is to compare a world without a proposed regulation (i.e., the baseline) with a world in which the regulation is enacted.

Harberger and Jenkins (2015, 7) also note that their approach to discounting requires a "reinterpretation of the concept of opportunity cost." In their assessment, "rather than thinking of the opportunity cost of public funds as their 'best' alternative yield, this reinterpretation looked upon it as their 'likely' alternative yield" (Harberger and Jenkins 2015, **7**). This is why the rationale for the use of the opportunity cost of capital as a basis for discounting is misleading. The SDR does not compare an investment to its next-best alternative. Rather, it compares an investment to only one specific alternative, the most likely alternative, which is the baseline alternative.

The weighted average approach uses weights that approximate only how social projects displace consumption and investment flows *in general*. Therefore, an obvious problem with this approach is that resources will be used differently in different contexts. The returns on investments lost as a result of regulation will not be the same for every social project. However, it might be reasonable to think such returns are close to the marginal return on average. For example, if the expected return on an investment is significantly above market rates of return, firms can still borrow from capital markets to finance their investments, even after complying with a regulation.

It is therefore reasonable to assume that investments displaced by social projects are the ones on the margin, meaning they earn rates of return that are approximately equal to the cost of capital. No doubt there will be cases where government actions displace investments with higher or lower returns than the marginal rates—for example, when firms or individuals face credit constraints and cannot borrow capital or when businesses would have invested in projects that fail. Nonetheless, identifying a unique SDR for every social project—at least with present knowledge—will likely prove to be too difficult. The most practical way forward is to calculate an average rate of return on lost investments using market interest rates that reflect the cost of capital to firms.

Under the standard form of the weighted average approach, both consumption and investment are preferred sooner rather than later. On the one hand, producer surplus flows (i.e., profits) can be converted into even greater amounts of both producer and consumer surplus flows in the future, so these flows can have a compounding effect over time. Consumer surplus, on the other hand, is discounted because consumers are impatient, just as with the Ramsey rule.

Failure to discount producer surplus flows is a problem because it does not acknowledge the effect that compounding of lost investments has on economic growth. However, because consumption cannot be reinvested, there is no compounding effect over time. For reasons discussed earlier, it is reasonable to think that forgone consumer surplus flows should receive no special treatment on the basis of when the flows arrive. This would suggest setting the time preference parameter under a weighted average rule equal to zero.

This modified version of the weighted average approach produces some intuitively pleasing results. First, no special treatment is given to consumption on the basis of timing, so analysts are not suggesting anyone should die of cancer today just because Cleopatra enjoyed a second helping of dessert—to use an example from Cowen and Parfit (1992). On the other hand, the importance of economic growth is emphasized with the preference that producer profits arrive sooner rather than later. These intuitive results yield an approach that appears to address many of the concerns raised by Cowen and Parfit (1992), and later by Cowen (2007), related to economic growth and equity across generations. Thus, discounting to net-present-value, rather than being taken literally (such as implying that lives saved in the future are less important than the same benefits today), should instead be viewed as a rule for comparing returns on social projects to returns under a baseline scenario.

2.5 Implications for Regulatory Policy

The SDRs currently used by regulatory agencies in the United States are 7 percent and 3 percent. According to the Office of Management and Budget, which sets guidelines for regulatory analysis,

> The 7 percent rate is an estimate of the average before-tax rate of return to private capital in the U.S. economy. It is a broad measure that reflects the returns to real estate and small business capital as well as corporate capital. It approximates the opportunity cost of capital, and it is the appropriate discount rate whenever the main effect of a regulation is to displace or alter the use of capital in the private sector. (OMB 2003, 33)

Meanwhile, the lower 3 percent rate exists because

> The effects of regulation do not always fall exclusively or primarily on the allocation of capital. When regulation primarily and directly affects private consumption (e.g., through higher consumer prices for goods and services), a lower discount rate is appropriate. The alternative most often used is sometimes called the "social rate of time preference." (OMB 2003, 33)

What OMB describes in its guidance is something like the weighted average approach to discounting combined with the Ramsey rule (recall OMB lists diminishing marginal utility of consumption as one of its three reasons for discounting). OMB is concerned with resources that are displaced by regulation, as is the weighted average approach. Displaced consumption is discounted at 3 percent, the pure rate of time preference, whereas displaced investments are discounted at 7 percent, the opportunity cost of capital. Any weighted average of OMB's recommended rates would have to be bounded by 3 percent and 7 percent, so any social project that realistically passes a benefit-cost test at the 7 percent rate should improve economic efficiency. Perhaps OMB thinks something like 4 percent is the rate at which society should discount based on diminishing marginal utility (if following the Ramsey rule).

This paper recommends a slightly modified version of OMB's current approach. First, OMB appears to at least partially embrace the Ramsey rule, given its embrace of diminishing marginal utility as a rationale for discounting. OMB should reject this rationale for discounting on the basis that it is not consistent with standard BCA assumptions. By extension, OMB should reject Ramsey approaches to discounting and be explicit that the weighted average approach is the agency's recommended approach.

OMB may also want to consider whether an update to the 7 percent base-case discount rate is appropriate. Harberger and Jenkins (2015) recommend a discount rate of about 8 percent for advanced countries. They arrive at that number through a weighted average formula with an additional component for the marginal cost of foreign funds. For example,

$$SDR = f_1 r + f_2 \rho + f_3 MCFF, \tag{24}$$

where f_1, f_2, and f_3 are weights to be applied to the opportunity cost of capital r, the pure rate of time preference ρ, and the marginal cost of foreign funds (which presumably differs from the marginal cost of domestic funds). The weights represent the sources of funding to pay for social projects. Using the Harberger and Jenkins estimates of 8.6–10.8 percent for r (which are notably higher than the 7 percent rate OMB uses) and 6 percent for the $MCFF$, and weights of 0.5–0.6 for f_1 and 0.3 for f_3, equation (24) is recalibrated by setting the pure rate of time preference parameter ρ equal to zero.[17] The result is an SDR in the range of 6.1–8.3 percent, or 7 percent, which is approximately the midpoint of this range. Thus, the 7 percent base-case estimate of the SDR recommended by OMB is in line with the approach outlined in this paper.

If OMB wants to continue to present a range of discount rates, a reasonable lower bound on the SDR would be zero, reflecting the extreme case where only consumption is displaced by social projects. A reasonable upper bound of 10 or 11 percent would reflect the opposite extreme, where only domestic investment is displaced by social projects. Those who want assurance that interventions are improving social welfare can be fairly confident that a proposed project that realistically passes a benefit-cost test at the higher end of discount rates will improve economic efficiency, although there might be reasons for using still higher SDRs, such as due to the irreversibility of many social projects.

2.6 Conclusion

Economists both within and outside the federal government seem uncertain as to why they discount. At the very least, economists discount for a variety of reasons. The scope of issues that the SDR is used to resolve is astounding, considering it is just a single

[17] Harberger and Jenkins (2015) use time preference values in the range of 6.0–8.1 percent for advanced countries.

number. The SDR is simultaneously used to account for displaced consumption and investments that pay for social projects, the rate of time preference of society, the proper distribution of wealth across generations, society's degree of risk and inequality aversion, and more. Is it any wonder that economists cannot agree whether the appropriate number to account for all these things is 3 percent, 4 percent, or some other number?

This paper presents a case for abandoning discounting approaches based on the Ramsey rule, as well as those based on gamma discounting, for social project analysis. This conclusion is consistent with the results of a recent survey that found that "the prominence of the simple Ramsey Rule needs to be revisited" (Drupp et al. 2015, 4). Still, these approaches are useful in other contexts, such as explaining individual behavior or as assumptions about the behavior of representative agents in optimal growth models. This paper claims only that the Ramsey and Gamma discounting approaches pose problems for use in benefit-cost analyses of social projects.

In place of these approaches, this paper recommends using an SDR that is based on the TVM. With this approach, the SDR serves one function only: to compare returns on social projects to returns under a baseline scenario. In this manner, the no action alternative need not be explicitly considered in an analysis since the alternative of leaving resources in private hands is already implicitly considered through the practice of discounting.

The rule that comes closest to this approach is the weighted average rule. A modified weighted average rule that sets the pure rate of time preference at zero is the recommended approach of this paper. Following this approach with recent data yields a best estimate of the SDR of around 7 percent.

If and when OMB decides to update its guidelines on regulatory analysis, the agency should provide clarity on the reasons it believes discounting is necessary in regulatory analysis. This paper argues the agency should reject Ramsey approaches and endorse the weighted average approach to discounting. OMB's current recommended base case estimate of 7 percent is still reasonable, but the agency might want to consider recommending 0 percent as a lower bound on the SDR and 10 or 11 percent or higher as an upper bound. Such a range would address the inherent uncertainty surrounding the sources of funding for social projects.

References

Arrow, Kenneth J. 1950. "A Difficulty in the Concept of Social Welfare." *Journal of Political Economy* 58 (4): 328–46.

Arrow, Kenneth J., Maureen L. Cropper, Christian Gollier, Ben Groom, Geoffrey M. Heal, Richard G. Newell, William D. Nordhaus,Robert S. Pindyck, William A. Prizer, Paul R. Portney, Thomas Sterner, Richard S. J. Tol, and Martin L. Weitzman. 2012. "How Should Benefits and Costs Be Discounted in an Intergenerational Context? The Views of an Expert Panel." Discussion Paper, Resources for the Future, Washington, DC.

Arrow, Kenneth J., William R. Cline, Karl-Göran Mäler, Mohan Munasinghe, Raymond L. Squitieri, and Joseph E. Stiglitz. "Intertemporal Equity, Discounting, and Economic Efficiency." In *Climate Change 1995: Economic and Social Dimensions of Climate Change,* edited by James P. Bruce, Hoesung Lee, and Erik F. Haites, 125-44. USA: Cambridge University Press.

Cowen, Tyler. 2004. "Policy Implications of Zero Discounting: An Exploration in Politics and Morality." *Social Philosophy and Policy* 21 (1): 121–40.

———. 2007. "Caring about the Distant Future: Why It Matters and What It Means." *University of Chicago Law Review* 74 (1): 5–40.

Cowen, Tyler, and Derek Parfit. 1992. "Against the Social Discount Rate." In *Justice between Age Groups and Generations*, edited by James S. Fishkin and Peter Laslett, 144–61. New Haven, CT: Yale University Press.

Drupp, Moritz, Mark Freeman, Ben Groom, and Frikk Nesje. 2015. "Discounting Disentangled." Working Paper No. 172, Grantham Research Institute on Climate Change and the Environment, London.

Gollier, Christian. 2004. "Maximizing the Expected Net Future Value as an Alternative Strategy to Gamma Discounting." *Finance Research Letters* 1 (2): 85–89.

———. 2013. *Pricing the Planet's Future: The Economics of Discounting in an Uncertain World*. Princeton, NJ: Princeton University Press.

Gollier, Christian, and Martin Weitzman. 2010. “How Should the Distant Future Be Discounted When Discount Rates Are Uncertain? *Economics Letters* 107 (3): 350–53.

Harberger, Arnold C., and Glenn P. Jenkins. 2015. “Musings on the Social Discount Rate.” *Journal of Benefit-Cost Analysis* 6 (1): 6–32.

Interagency Working Group on Social Cost of Carbon. 2013. *Technical Support Document:—Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis—Under Executive Order 12866*.

Marglin, Stephen A. 1963. “The Social Rate of Discount and the Optimal Rate of Investment.” *Quarterly Journal of Economics* 77 (1): 95–111.

Mishkin, Frederic S. 2016. *The Economics of Money, Banking, and Financial Markets*, 11th ed. Essex, UK: Pearson Education.

OMB (Office of Management and Budget). 2003. *Circular A-4: Regulatory Analysis*. September 17.

Ramsey, F. P. 1928. “A Mathematical Theory of Saving.” *Economic Journal* 38 (152): 543–59.

Robbins, Lionel. 1938. “Interpersonal Comparisons of Utility: A Comment.” *Economic Journal* 48 (192): 635–41.

Simon, Carl P., and Lawrence Blume. 1994. *Mathematics for Economists*. New York: Norton.

Viscusi, W. Kip. 2007. “Rational Discounting for Regulatory Analysis.” *University of Chicago Law Review* 74 (1): 209–46.

Weitzman, Martin L. 1998. “Why the Far-Distant Future Should Be Discounted at Its Lowest Possible Rate.” *Journal of Environmental Economics and Management* 36 (3): 201–8.

———. 2001. “Gamma Discounting.” *American Economic Review* 91 (1): 260–71.

3 The Stock and Flow of US Federal Regulation: Evidence of Technology Shocks?

3.1 Introduction

A key concept in economic theory is the distinction between stock variables and flow variables. Economists are used to distinguishing between financial stocks and flows, such as wealth and income. Less often do they use these concepts to describe government policies. This paper contributes to the analysis of public policies by highlighting the disparate effects of the stock and flow of US federal regulation, focusing on how regulation predicts changes in the aggregate price level in the US.

Conventional wisdom suggests regulation should have no impact on the price level, at least in the long run. The quantity theory of money, for example, argues that the price level is determined by the total quantity of money in the economy, not by real factors such as the amount of oil extracted from the ground or the number of government regulations. According to this theory, if the money supply is doubled, the price level will double as well. Any increase in regulations might raise the prices of a few individual goods or services, but these price increases should be offset by declines in the wages of workers, the real incomes of consumers, and the profits of the producers of goods and services.

Real life is likely messier than the quantity theory would suggest, however.

Economists do not have perfect measures of “the” price level. Instead they have rough approximations of the price level that come from tracking movements in prices of products found in a basket of goods and services. Even if the “true” price level is determined by nominal, i.e. monetary forces, supply-side factors routinely influence prices of goods and services in basket of goods that make up price indices, as do demand-side factors unrelated to any changes in the money supply.

This is why economists distinguish between headline inflation, total changes in a price index such as the consumer price index, from core inflation, which removes volatile food and energy prices from the price index. Food and energy prices are generally determined by supply side, i.e. real, factors, and relative, rather than aggregate, demand. The total change in headline inflation then will always be determined by some mix of real and nominal forces. To identify just the nominal component of a price index change, one must remove price changes caused by other factors.

From a theoretical standpoint, then, real factors can and do raise headline inflation, at least in the short run. Thus, it is conceivable that regulation could impact the aggregate price level, as measured by price indices. It is ambiguous, however, whether the stock or the flow of regulation will have more significance. Legally, no requirement on the books is any more or less important than any other requirement; they all carry the force of law. Hence, one might surmise that the cumulative stock of all government regulations is the more significant factor. On the other hand, old regulations may be forgotten about, may be poorly enforced, could become obsolete over time, or might simply be low priorities for government officials who are focused on writing new rules, cementing legacies and tackling the most fashionable issues of the day. These are good reasons to believe new regulations may matter more.

New regulations are also likely to generate more attention from incumbent firms as they seek to avoid the costly compliance costs associated with new rules, while old, already-complied-with, rules may represent sunk costs. Furthermore, it’s conceivable that short run shocks to the flow of regulation impact prices of products but not relatively stickier wages, while in the long-run, wages adjust making regulation a relatively neutral force.

This analysis finds that both short-run regulatory shocks and the long-run stock of federal regulation significantly predict changes in the aggregate price level. There is a significant cointegrating relation between the long-run level of regulation and prices. Furthermore, the cointegrating relation appears to begin around 1960, when the level of federal regulation began to increase significantly. Similarly, output from structural vector autoregressions demonstrates significant short-run co-movements between regulation and prices. Regulation explains roughly 12 to 17 percent of the forecast error variance in the price level in the years following a regulatory shock. Findings are robust to a number of alternative model specifications, though significance levels vary across price indices as well as different model specifications.

These findings have important implications for the macroeconomic analysis of regulations. Regulation may be a form of technology shock, much like weather, oil prices, or disruptive technological innovations. However, regulation also appears to have effects like a demand shock as well. Micro-level evidence from spending on durable and nondurable goods provides clues as to the mechanisms by which regulation induces shocks, such as intertemporal substitution on the part of consumers and higher spending on relatively price inelastic goods when prices increase. Regulation may also be an overlooked factor whose influence should be removed from indices of core inflation that seek to eliminate price changes due to real factors.

3.2 Regulatory Macroeconomics

In recent years, a small but growing literature has emerged that attempts to quantity federal regulation for the purposes of macroeconomic analysis, though few studies explicitly distinguish between the effects of the stock of regulation and the effects of the flow of regulation. The US Code of Federal Regulations (CFR) is perhaps the most common source used to explain how the total stock of federal regulation evolves over time. The CFR is published on an annual basis and is where regulations are eventually captured once finalized. Some studies have used page-count data for the CFR as a proxy for the total stock of regulation, a technique that is continued in this paper. For example, Dawson and Seater (2013) use adjusted page-count data to find that federal regulation in the United States has slowed economic growth by an average of two percentage points per year since 1949.

Other studies have used computer programs to count the number of restrictions in the CFR (Al-Ubaydli & McLaughlin, 2015). Restrictions are words such as “shall”, “must”, and “prohibited,” that signify binding legal constraints on the public’s allowable activities. Still others have used staffing levels at federal agencies as a proxy for the level of regulation (e.g. Dudley and Warren (2016)). Each of these measures—page-counts, restriction counts, and staffing levels—represent measures of the stock of regulation at different points in time.

There have been similar attempts to evaluate regulation as a flow variable. For example, Crews (2016) is part of a series of annual reports that counts the number of new rules, executive orders, and other notices issued by the federal government each year. Much of this information is captured in the *Federal Register* (FR), a daily publication of federal bureaucratic activity. De Rugy and Davies (2009) uses page-count data in the FR as a measure of the flow of federal regulation, finding evidence of spikes in regulatory activity during the period between presidential elections and inauguration day (dubbed the "midnight" period). Budget expenditures of federal regulatory agencies have also been used as a proxy for regulatory flows (Dudley and Warren, 2016).

Benefit-cost analysis (BCA) is another a tool used to evaluate regulatory flows.[18] For example, each year the Office of Management and Budget (OMB) issues an annual report that aggregates regulatory agency estimates of benefit and cost flows for a decade's worth of major regulations (e.g. Office of Management and Budget (2016)).

There is a key empirical challenge to contend with when distinguishing between the different impacts of the stock and flow of policies. This challenge arises due to the time trend component of time series data. Regressing levels of two time series on one another can lead to spurious correlations. The usual way to address these problems is to remove the time trend component of a time series, i.e., convert the series from a

[18] Formally, BCA is a comparative statics exercise analyzing differences between the stock of social wealth at two discrete moments in time. For the purposes here, however, it can be thought of as describing flows of benefits and costs that accrue between equilibrium moments.

nonstationary to a stationary process. This can be done using a time series filter or alternatively one can take first differences of the data. These approaches help overcome *statistical* problems related to spurious regression, but *conceptually* the approaches are problematic if one is interested in how *the level* of a time series variable, in this case regulation, influences *the level* of another time series variable, such as a price index. One is throwing the baby out with the bathwater when one converts from levels to changes in that the magnitudes of these variables are lost, and magnitudes are what are needed when analyzing stock variables.

This paper overcome these problems by using cointegration and vector autoregression procedures to distinguish between the dual effects of the stock and the flow of time series data. These techniques have primarily been applied in areas outside of regulation, most notably in the field of monetary economics,[19] as well as in the literature on oil shocks.[20] This research extends these techniques to the analysis of regulation.

This analysis is also unique in that it focuses on how regulation impacts the price level. Most other empirical studies of the macroeconomic consequences of regulation have focused on economic growth rates or productivity. For example, Djankov et al. (2006) finds that improving the climate of business regulations can add as much as 2.3 percentage points to annual growth rates. Dawson and Seater (2013) finds that federal regulation has slowed economic growth by roughly two percentage points per year since 1949. Coffey et al. (2016) finds that US private GDP would have been $4 trillion larger in

---

[19] See Christiano at al. (1999) for an excellent summary of findings from the monetary VAR literature.
[20] See Hamilton (1983) and (2003), and Bernanke et al. (1997) for examples of studies using VAR techniques in the context of oil shocks.

2012 had federal regulation had been held at 1980 levels. Nicoletti et al. (2003) finds that liberalizing product market regulation improves productivity performance. Dawson (2006) and Crafts (2006) both find evidence that regulation lowers total factor productivity, and Égert (2016) finds that product and labor market regulations have reduced multifactor productivity in OECD countries.

By focusing on the price level, rather than on growth or the factors that contribute to growth, this study is harkening back to the period when supply-side factors were thought to be important contributors to inflation. Belief that the private sector could create inflation was so strong in the 1970s, in fact, that it led to the creation of a new federal agency, the Council on Wage and Price Stability, which was tasked in part with studying the inflationary impact of regulations (Hopkins & Stanley, 2015).

The logic underlying supply-side sources of inflation was something like the following.[21] The national income version of the equation of exchange is stated as,

$$MV = Py. \tag{25}$$

Here the left hand side of Equation (25) is simply total expenditures on final goods and services, while the right-hand side represents total income from final goods and services. It follows then that, *ceteris paribus*, anything that raises real income, y, will reduce the price level (since there will be the same number of dollars chasing more goods and services), and inversely anything that lowers real income will raise the price level, assuming nominal income is held constant.

In more recent years, these kinds of supply-side changes in the price level have

---

[21] This argument is based in part on Miller (2011).

come to be known as technology shocks. A common view is that monetary policy should not seek to offset changes in the price level due to technology shocks—only those related to shocks to velocity.[22] This is a central reason why economists, and in particular central bankers, are concerned with core, rather than headline, inflation. If regulation is a source of technology shock, it could be raising headline inflation in a manner similar to the prices of food and energy. Hence, central bankers would not want to offset these changes with monetary policy.

3.3 Methods

3.3.1 Cointegration

This study uses cointegration time series techniques to estimate whether a long-run steady-state relationship exists between the level of regulation and three measures of the price level. Early work by Granger (1981) and Engle and Granger (1987) developed the theory of cointegration, which defines two time series as cointegrated if each process is integrated of order one, or I(1), and the two series react to one another via a stationary I(0) process that links the two series. More formally, we say that two time series x and y are cointegrated if the equation

$$z_t = y_t - \theta x_t + \varepsilon_t \tag{26}$$

is stationary, i.e., ε is mean reverting, and where θ is a cointegrating coefficient.

Johansen (1988, 1991, 1995) designed tests for determining the number of cointegrating equations that exist between two time series. This first step involves

[22] See Selgin (1997) for a detailed description of this argument. Bernanke et al., (1997) makes a similar case that the Federal Reserve should not respond to certain technology shocks, such as oil shocks.

determining whether the two series have unit roots, which can be done using the augmented Dickey-Fuller unit root test. Assuming both equations are unit root processes (and are thus I(1) processes), one can then test for the "rank" of these equations using maximum likelihood tests to determine the number of cointegrating equations linking the two series.

A finding of zero cointegrating equations would mean the two series are not cointegrated, while two cointegrating equations implies that both series are I(0), or stationary. Alternatively, if there is one cointegrating equation then the coefficient of the cointegrating equation can be estimated to determine its magnitude and significance level. The cointegrating equation itself must be stationary or else the two time series will drift apart over time.

Establishing cointegration is useful for several reasons. First, many time series variables have unit roots or are otherwise nonstationary. Such data is inappropriate for use in standard OLS regression analysis since the time trend component of two nonstationary series can lead to spurious correlations. A typical response to this problem is to remove the time trend by taking first differences or by identifying deviations from a trend component of the series using time series filters. However, if one wants to understand the effects of the *stock* of regulation, i.e. the *level*, and not just the short-run effects of regulatory *flows*, then time series filters and first differences are conceptually inappropriate for the task.

Analysis of cointegration overcomes these problems because it uses levels of time series variables rather than changes. Interestingly, very little work has been done to study the effects of regulation in the context of cointegration. Dawson (2007) is one notable exception, which uses an older version of the dataset used in this study. That study found a cointegrating relation between intensive forms of regulation and total output. This analysis builds on that work by exploring whether regulation is cointegrated with aggregate measures of the price level.

3.3.2 Structural Vector Autoregression

To analyze short-run regulatory shocks, vector autoregression (VAR) techniques, as developed by Sims (1980) and refined in subsequent papers, are used. A simple VAR is of the form:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \beta_{p+1} C + u_t, \tag{27}$$

where y is a dependent variable regressed on p lags of itself, and the matrix C contains control variables as well as p lags of each control. Such a regression is referred to as a VAR(p) process, with dimension, K, where K is the number of variables in the regression. A VAR includes a series of OLS regressions of the form presented in Equation (27), where each control variable is also a dependent variable in a regression where the other $K - 1$ variables are used as controls.

The simplest form of a VAR is a reduced-form VAR, which estimates simple correlations between variables. This study uses the structural variation of a VAR—or SVAR—where causation is established through assumptions about the contemporaneous cause and effect relations among variables. In other words, identification is established by assuming some variables cause changes in other variables contemporaneously, i.e., within the same period in the analysis, while additional variables are assumed to cause changes to other variables only in subsequent periods. Such assumptions are usually based on argumentation and logical reasoning, as opposed to being based on empirical observation.

To date, very few studies have used VAR or SVAR techniques in the context of regulation. Again, Dawson (2007) uses these techniques and finds significant negative short-run responses in output in response to regulatory shocks. Michaels and DeVany (1995) uses VAR and cointegration to examine how regulation influences the extent of integration of natural gas markets. Several unpublished studies also use VAR to study regulations. Beard et al. (2011) looks at impacts on private GDP and employment using budget and staffing levels at regulatory agencies as proxies for regulation. Unlike the approach in this paper, which distinguishes between stock and flows, the Beard et al. study combines a stock variable (staffing levels) with a flow variable (agency budgets) in a single VAR. Combining stocks and flows in this manner may be inappropriate and is likely to lead to results that are hard to interpret. Beard et al.'s approach was also criticized in Sinclair and Vesey (2012) for lacking robustness and for failing to present results from granger causality tests and forecast error variance decompositions, which are commonly presented in papers employing VAR techniques.

Unlike with cointegration, which typically uses levels of nonstationary time series data, VARs should be run using stationary variables. This is a key distinction between the VAR estimates and the cointegration estimates in this paper. For the VAR estimates, time series data is converted from I(1) to I(0) using a band-pass filter. Obtaining I(0) data could also be done by taking first differences, and because the data is in log form, this would approximate growth rates. Instead, the Baxter and King (1999) band-pass time series filter is used to separate data into trend and cyclical components, with the cyclical component being what is used in the regression analysis. This has some advantages over taking first differences, such as the fact that first differences can induce phase shifts in the data.

Images of the raw data used in this study and the filtered trend line can be found in Figure 28 through Figure 33 in Appendix A. Note that several years of data at the beginning and the end of each series are lost when run through the time series filter. One lag is used in each regression, which is generally consistent with the Bayesian information criteria (BIC) approach for selecting lag lengths.[23] An alternative to using BIC to select lag lengths is the Akaike information criteria (AIC), which again usually recommends one lag for these data but sometimes recommends two lags. Using two lags was experimented with but did not alter any core results.

Both monetary and nonmonetary shocks likely influence prices in the short run, so each are controlled for; the nominal M2 money supply is used to control for monetary shocks and the nominal price of oil is used to control for non-monetary shocks other than regulation. Because one variables in each VAR is always a measure of the price level, this accounts for real changes in controls, i.e., implicitly changes in real money balances and the real price of oil are controlled for because the price level is also controlled for in the regression. Additional monetary and nonmonetary controls were experimented with, including stock prices, government spending, and interest rates and spreads of various kinds. These controls had little explanatory power and so were left out of the final

---

[23] See Lütkepohl (2005), chapter 4, for details on selecting lag lengths in VARs.

analysis.

3.4 Data

The measure of federal regulation used comes from Dawson and Seater (2013), which uses the number of pages in the US Code of Federal Regulations (CFR) as a proxy for the level of the total accumulated stock of US federal regulations. This is annual data and is adjusted from raw page-counts to account for changes in font size and other layout differences that occur over time. Chapters of the CFR that contain information not considered regulatory in nature are also excluded from the page-count.[24] The original dataset from Dawson and Seater (2013) ran from 1949-2005, however the data in this study has been updated through the year 2015.

[24] For more information about the dataset, as well as information on adjustments made to the raw CFR page-count data, see Appendix A of Dawson and Seater (2013).

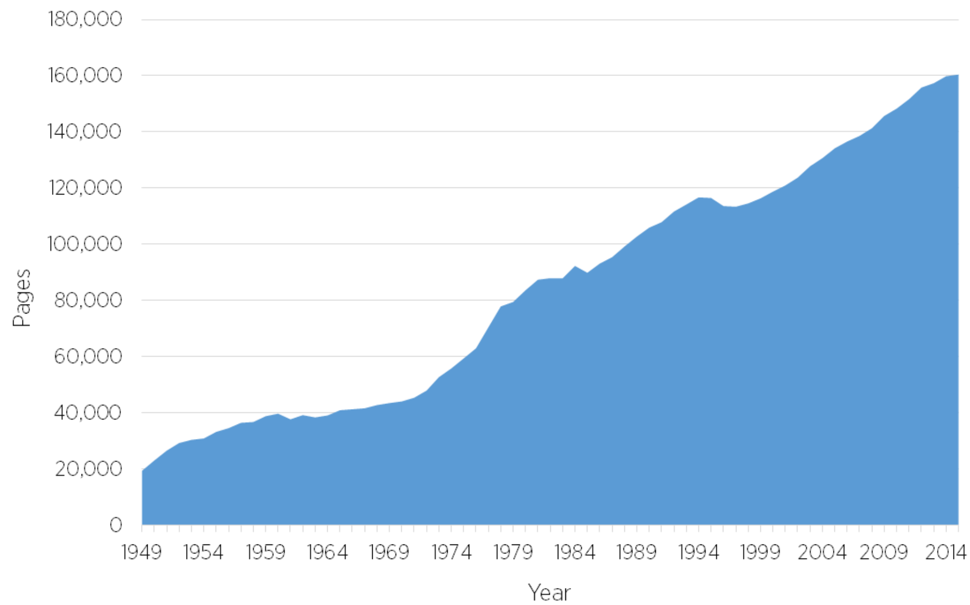


Figure 17: Adjusted Page-Count of the US Code of Federal Regulations, 1949-2015
Source: Dawson and Seater 2013.

Figure 17 plots the adjusted CFR page-count over time. The data plotted in logarithmic form can be found in Figure 31 in Appendix A. In 1949, when the series begins, the adjusted US Code was 19,335 pages in length compared to over 160,000 pages in 2015. This represents a total increase of over 700 percent during the sample period. While there have been occasional years when the level of federal regulation declined, those periods were always fully offset by increases in regulation in subsequent years. The persistent trend over time is for the level of regulation to increase, suggesting that regulation may be a nonstationary process.

Whereas the trend in the level of regulation is fairly consistent, much more variation arises in the growth rate of federal regulation, which is plotted in Figure 18. Here first differences of the logged adjusted CFR page-count data are plotted, which can be interpreted as the annual growth rate of federal regulation. Whereas the level of regulation grows steadily, the growth rate of regulation is more volatile. The most prominent features of Figure 18 are the rapid, but slowing, growth of federal regulation in the early 1950s, the highly volatile growth rates of regulation throughout the 1960s and 70s, and the amelioration of volatility in more recent decades during the so-called great moderation period into the period during and after the Great Recession.

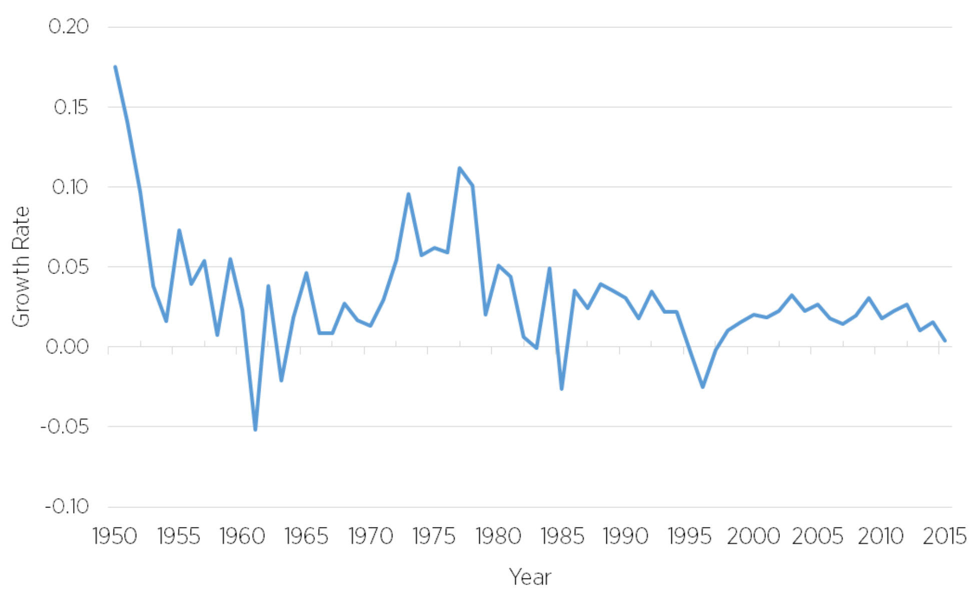


Figure 18: Growth Rate of Federal Regulation, 1949-2015
Source: Dawson and Seater 2013.

Three price indices are used in this analysis: the Personal Consumption Expenditures (PCE) Chain-Type Price Index published by the US Bureau of Economic Analysis (USBEA); the Producer Price Index (PPI) for all commodities, which comes from the US Bureau of Labor Statistics (USBLS); and the Consumer Price Index for All Urban Consumers (CPI), which also comes from the USBLS. Figure 19 plots the growth rate of regulation against the growth of the CPI, again using first differences of logs to approximate growth rates. From a superficial standpoint, there appears to be some co-movement between the two series. Regulation sometimes lags consumer prices, and at other times regulation appears to be a leading indicator.

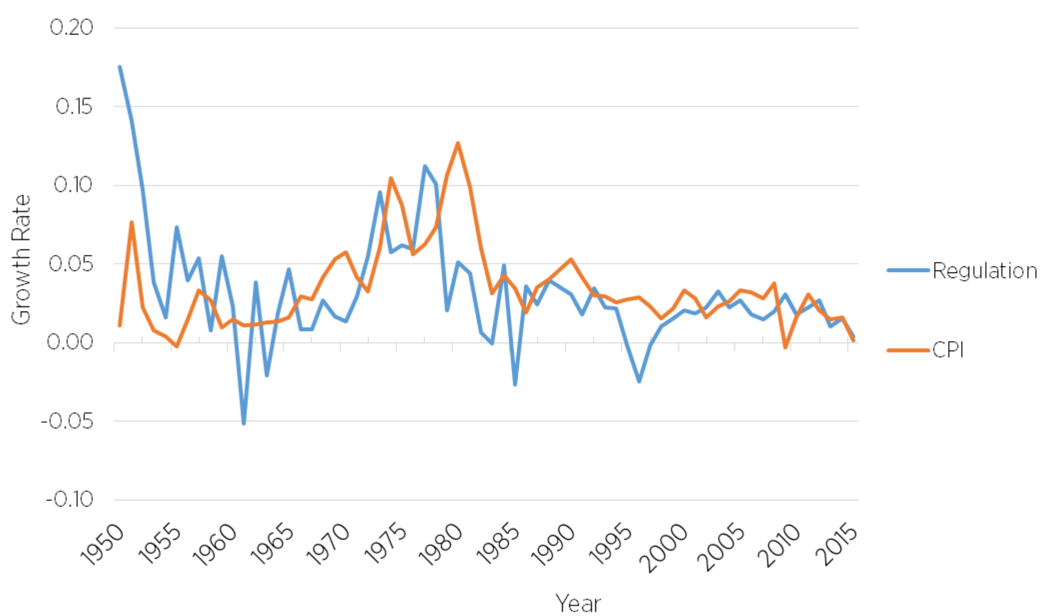


Figure 19: Producer Price Inflation and Growth Rate of Regulation, 1949-2015
Source: US Bureau of Labor Statistics and Dawson and Seater 2013.

As mentioned, in the short-run analysis, monetary and non-monetary factors likely to produce macroeconomic shocks are controlled for. On the non-monetary side, two time series on oil prices are merged. From 1949 to 2012, the Wall Street Journal West Texas Intermediate Spot Oil Price is used, which was discontinued in mid-2013. From 2013 through 2015, the U.S. Department of Energy's West Texas Intermediate Crude Oil Prices dataset is used. These two series are virtually identical for the periods where they overlap, so merging these data is not likely to raise issues.

To control for monetary factors, the Federal Reserve's M2 measure of the money supply is used. This series runs from 1959 onwards, so for the first decade of the sample these data are merged with supplemental M2 monetary data from the St. Louis Fed, as described in Rasche (1990). A graph of logged M2 is available in Figure *32* in Appendix A. Note that M2 data was used in part because of the dramatic increase in the monetary base that shows up in M1 in the post-2008 period. This choice appears to be reasonable in part because results are consistent with other findings in the monetary literature. For example, the common "price puzzle" finding in the monetary VAR literature, i.e., that monetary shocks are followed by reductions in the price level, is confirmed here. Given the unusual increase in the monetary base in the post-2008 period, analysis of the period from 1949-2007 was also conducted and the core results do not significantly change.

3.5 Results

3.5.1 Long-Run Cointegration

This study is interested in whether the level of regulation, as measured by pages in the CFR, is cointegrated with three measures of the price level: the PCE index, the PPI index, and the CPI index for all Urban Consumers. Recall that cointegration estimates here rely on logged levels rather than filtered data. As such, evidence of cointegration would establish a long-run relationship between *the level* of the cumulative stock of federal regulation and *the level* of prices.

The first step towards establishing cointegration, as discussed in the methodology section, is to determine whether these series are I(1). Thus, Dickey-Fuller (DF) tests are performed, the null hypothesis being that a series has a unit root. Coefficients and standard errors for the lagged term in the OLS regression for the Dickey-Fuller tests are presented in Table 5, along with the relevant test statistics. Note that the critical values for a Dickey-Fuller distribution differ from the critical values for a standard t-distribution. The relevant test statistics are -1.610 at the ten percent level, -1.950 at the five percent level, and -2.614 at the one percent level.

Table 5: Dickey-Fuller Test Results for Inflation Indices
Source: Author's calculations.

Dep. Var.	D.ln*cfr* (w/ trend)	D.ln*pce*	D.ln*ppi*	D.ln*cpi*
Coefficient on lagged term	-.0906896	-.0042764	-.0046424	-.002644
Std. error	.03418	.003926	.0080594	.0041279
DF Test Statistic	-2.653	-1.089	-0.576	-0.641

For each price index, the null hypothesis of a unit root fails to be rejected.[25] For the regulation measure, initially the null hypothesis of a unit root is rejected. However, when a trend term is added to the regression, the presence of a unit root cannot be ruled out. It may be reasonable to suspect there is an underlying trend component to regulatory output. For example, each year regulators finalize roughly the same number of regulations, somewhere between three and five thousand. Thus, there appears to be a certain degree of regulatory inertia that occurs because of institutional features of the regulatory process.

A coefficient near zero for the lagged term in the DF test implies the presence a unit root for the following reason. The Dickey-Fuller test begins from the assumption that a time series follows a random walk process, such as

$$y_t = \alpha + \delta y_{t-1} + \varepsilon_t. \tag{28}$$

To avoid problems related to autocorrelation, first differences of Equation (28) are estimated instead, which can be written as,

$$y_t - y_{t-1} = \alpha + (\delta - 1)y_{t-1} + \varepsilon_t. \tag{29}$$

Thus, the OLS regression actually estimated is

$$\Delta y_t = \alpha + \beta y_{t-1} + \varepsilon_t. \tag{30}$$

The null hypothesis that a unit root is present (i.e. $\delta = 1$ in Equation (28)) is identical to testing that $\beta = 0$ in Equation (30). Not surprisingly then, the coefficient is approximately equal to zero in each case in Table 5.

---

[25] Test statistics and critical values for the unit root test are not presented here but can be made available upon request.

Next, first differences of these series are taken and Dickey-Fuller tests are performed again. These results are presented in Table 6 below. The null hypothesis of a unit root is rejected for the first difference of the regulation measure and the PPI index at the 1% level, for the CPI index at the 5% level, and for the PCE index at the 10% level.

Table 6: Dickey-Fuller Test Results for First Differences of Dependent Variables
Source: Author's calcalations.

Dep. Variable	D2.ln*cfr* (w/ trend)	D2.ln*pce*	D2.ln*ppi*	D2.ln*cpi*
Coefficient on first lag	-.5737	-.2092	-.5752	-.2436
Std. error	.0994	.0784	.1203	.0836
Test Statistic	-5.77	-2.67	-4.78	-2.91

Taken together, there is considerable evidence that the three price indices are I(1) processes because the logged levels of each series display evidence of a unit root, while first differences display strong evidence of stationary processes. There is slightly weaker, but still solid evidence that regulation is an I(1) process as well.

Next, following Johansen (1995), tests for the rank of the pairs of variables are performed. Pairings are regulation-PCE index, regulation-PPI index, and regulation-CPI index. Testing for the rank of a pair of time series variables generates a trace statistic and critical values, which in turn allows for testing the number of cointegrating equations for each pair. Trace statistics for each pairing are presented in Table 7.

Table 7: Trace Statistics and Critical Values for Regulation/Price Index Pairings
Source: Author's calculations.

Rank	ln*cfr*/ln*pce* Trace Statistic	ln*cfr*/ln*ppi* Trace Statistic	ln*cfr*/ln*cpi* Trace Statistic	5% Critical Value
0	23.86	29.56	24.33	15.41
1	1.10*	1.45*	1.12*	3.76
2	-	-	-	-

Each test suggests there is one cointegrating equation between the measure of the price level and the measure of regulation. If there were zero cointegrating equations, then the analysis would stop. However, given the evidence of a single cointegrating equation between each pair, next the coefficients of the cointegrating equation are estimated. Recall that Equation (26) above is the cointegrating equation being estimated. Normalizing the coefficient on the regulation measure $lncfr$ to 1 yields

$$8.37 = lncfr - .76lnpce, \tag{31}$$

for the ln*cfr*/ln*pce* index pairing,

$$7.92 = lncfr - .77lnppi, \tag{32}$$

for the ln*cfr*/ln*ppi* index pairing, and

$$8.30 = lncfr - .67lncpi, \tag{33}$$

for the ln*cfr*/ln*cpi* index pairing.

The coefficient on the price index is statistically significant in all cases.

Furthermore, the magnitudes are similar across equations. The direction of the sign suggests that a rise in regulation is associated with a rise in the price level, and vice versa.

A final test of cointegration is to confirm that the error term in the cointegrating equations are stationary. This can be done by testing the stability of the cointegrating equations and by plotting the predicted values of the cointegrating equations to see if the equations appear to be mean reverting. Each of the cointegrating equations passes these stability tests. An interesting feature of these tests is that the cointegrating relation appears to begin around 1960.[26]

3.5.2 Short-Run Structural Vector Autoregressions

Next, VARs are run to estimate how shocks to the flow of regulatory output predict changes in prices in the short-run. Following the recommendations of Stock and Watson (2001), Granger causality tests, impulse response functions, and forecast error variance decompositions are presented for each VAR. Recall that these regressions use stationary data with the time trend removed. Thus, data used represents deviations from a trend for all variables. Note also that these regressions use nominal data. By including the relevant price index as a dependent variable in each regression, changes in the real factors are implicitly controlled for.

i. Granger Causality Tests

Granger causality tests are a special case of Wald F-tests, which are joint hypothesis tests. The Granger test looks at whether the coefficients on a particular

[26] Data and Stata Do-Files are available upon request for anyone who wishes to reproduce these stability tests.

independent variable and its lags are jointly significantly different from zero in an OLS regression. The null hypothesis is that the coefficients on each of these independent variables and its lags are equal to zero. Thus, a low p-value recommends rejecting the null hypothesis that these coefficients are equal to zero, i.e. the variable and its lags have significant explanatory power in the regression. Granger causality tests help to establish whether past and contemporaneous changes in independent variables have explanatory power over future and contemporaneous changes in a dependent variable. A failure to reject the null hypothesis would appear to rule out causation since changes in an independent variable are not followed temporally by identifiable changes in a dependent variable.

Three SVARs were run, the orderings of which are presented in Table 8. Recall that the order in an SVAR is important in that variables appearing later in the regression (a higher number in columns in Table 8) are assumed *not* to cause contemporaneous changes in variables appearing earlier in the regression (a lower number in columns in Table 8). The main dependent variable of interest is entered last in the regression. Each other variable is assumed to impact the main dependent variable contemporaneously, but not vice versa.

Table 8: Structural Vector Autoregressions, and Variable Orderings

SVAR #	Ind. Variable (1)	Ind. Variable (2)	Ind. Variable (3)	Main Dep. Variable (4)
1	*reg*	*m2*	*oilprice*	*pce*
2	*reg*	*m2*	*oilprice*	*ppi*

3	*reg*	*m2*	*oilprice*	*cpi*

The rationale behind the orderings in Table 8 is fairly straightforward. Regulation is the primary independent variables of interest and is assumed to be exogenous. This is a reasonable assumption since it is difficult to imagine a causal mechanism whereby oil price, the money supply, or the price level, influence regulatory output immediately, especially since regulators typically begin working on rules years before the rules are finalized. Furthermore, it seems realistic that if regulators do respond to inflation by reducing the flow of regulation, that this occurs only after some time has passed.

The money supply is assumed to be able to influence oil prices contemporaneously, but not vice versa, which also seems reasonable. The relevant price index, which appears last in each regression, is assumed to be impacted contemporaneously by each of the other variables: money, oil and regulation, which also seems reasonable.

Results of Granger Causality tests are presented in Table 9. In all three cases, the null hypothesis that regulation *does not* Granger cause the relevant price index is rejected. The possibility of reverse causation is also tested for, and there is no evidence that changes in the price indices Granger cause changes in regulation. Importantly, the direction of the sign of these changes is not yet known, but the Granger test results appear to be consistent across price indices and imply the possibility of a causal relationship.

Table 9: Results of Granger Causality Tests
Source: Author's calculations.

SVAR #	Main Dependent Variable	P-Value for test that *reg does not* Granger cause Dependent Variable	P-Value for test that Dependent Variable *does not* Granger cause *reg*
1	ln*pce*	0.000	0.763
2	ln*ppi*	0.023	0.703
3	ln*cpi*	0.000	0.630

ii. Impulse Response Functions

Impulse Response Functions (IRFs) plot how unexplained shocks in one variable predict future changes in other variables. Usually, these are presented in graphical form. For the sake of space only how shocks to regulation predict future changes in each of the three price indices are presented. Graphical versions of the full set of IRFs for each SVAR set can be found in Figure 34 through Figure 36 in Appendix B.

The x-axis in Figure 20 through Figure 21: SVAR #2 - impulse(reg), response(ppi)

represents years after a shock has occurred. The y-axis plots the deviation from the trend component of the log of the response variable (in this case the price index). This can be interpreted as the percentage change in the response variable's unexpected component.

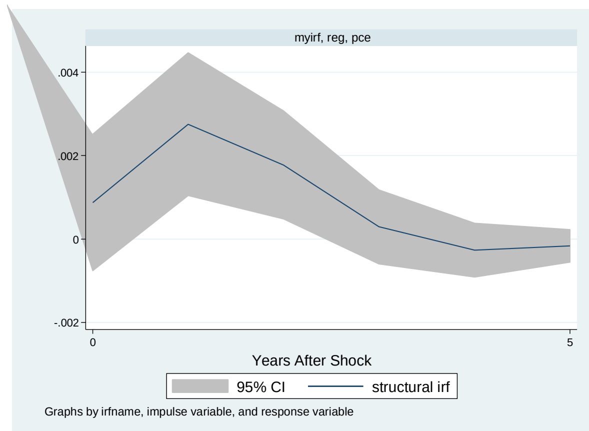


Figure 20: SVAR #1 - impulse(*reg*), response(*pce*)

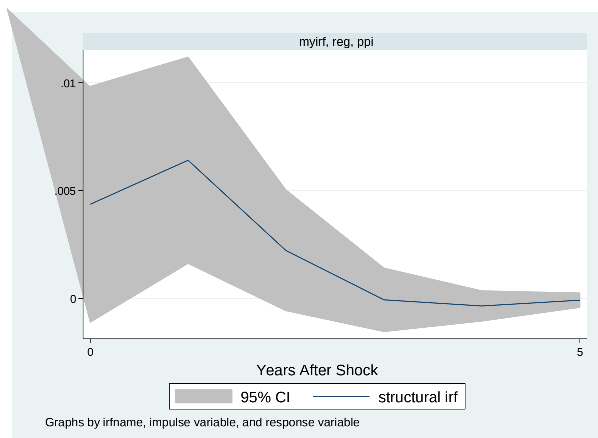


Figure 21: SVAR #2 - impulse(*reg*), response(*ppi*)

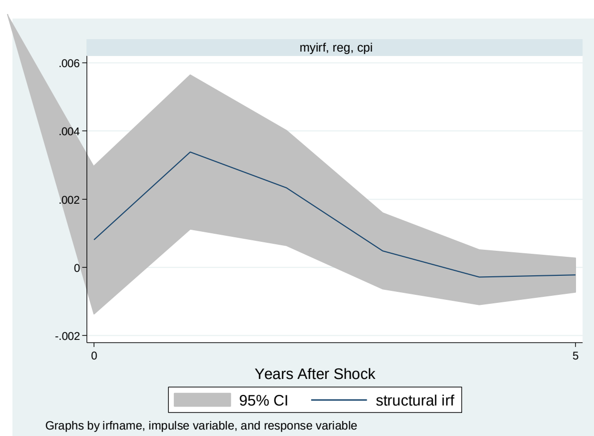


Figure 22: SVAR #3 - impulse(*reg*), response(*cpi*)

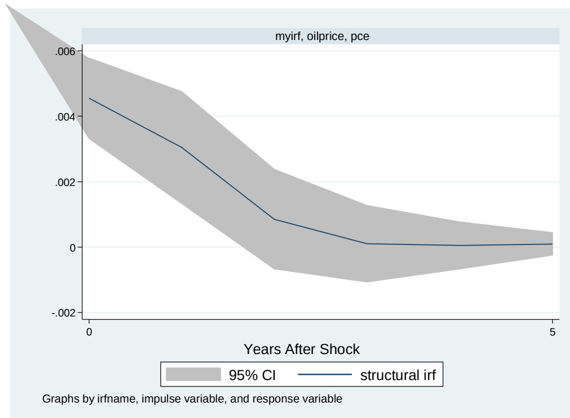


Figure 23: SVAR #1 - impulse(*oilprice*), response(*pce*)

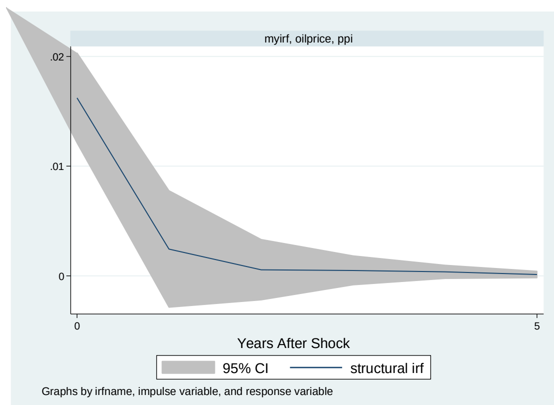


Figure 24: SVAR #5 - impulse(*oilprice*), response(*ppi*)

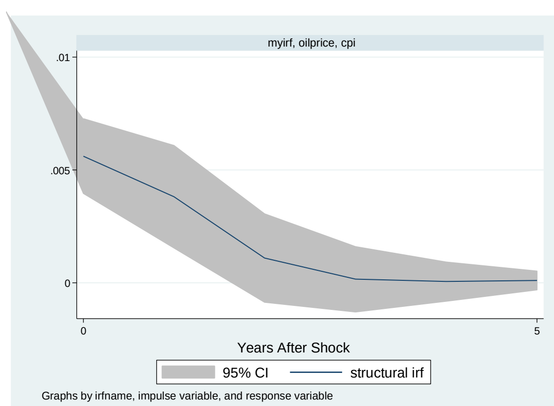


Figure 25: SVAR #6 - impulse(*oilprice*), response(*cpi*)

Regulatory shocks predict significant rises in all three price indices in subsequent years. For the PCE and CPI indices, the rise is statistically significant at the 95% confidence level in years one and two following the shock. For the PPI index, it is statistically significant in year 2. The magnitude of the effect is comparable across price indices but may be slightly larger for producer prices than for consumer prices.

For comparison, IRFs for oil shocks impulses and price level responses are presented in Figures 23-25 for each of the three SVARs. Again, the x-axis plots time and the y-axis plots the percentage change in the deviation from the trend component of the response variable (i.e. the price index). Nominal oil price shocks predict significant increases in the all three price indices in years zero and also in year one following the shock for the CPI and PCE price indices. This suggests oil shocks may work their way through the economy faster than do regulatory shocks. The magnitude of the price level response is also larger for oil shocks compared to regulatory shocks.

iii. Forecast Error Variance Decompositions

Results of forecast error variance decompositions (FEVDs) are presented for each SVAR in Table 8. FEVDs present the fraction of the variance of the error made in forecasting a variable that can be explained by previous shocks to other variables in the regression. The FEVDs for the price index variable are presented in Table 10 through Table 12.

Regulation is statistically significant at the 10 percent level and explains about 17 percent of the forecast error variance in the PCE index. By contrast, oil prices explain a little under half of the variance and the M2 money supply explains about 18 percent.

Results for the producer price index look similar. Oil price explains about half of the forecast error variance, while regulation explains about 12 percent (just failing to meet significance levels) and the money supply about 15 percent. Similarly, regulatory shocks explain about 15 percent of the forecast error variance in the CPI, which is statistically significant at the 10 percent level.

Across the three indices, regulation appears to explain approximately 12 to 17 percent of the forecast error variance in the price level in the years following a regulatory shock. These magnitudes are quite large, especially relative to monetary factors, though the significance levels vary across price indices.

Table 10: Forecast Error Variance Decomposition and Standard Errors for impulse: *reg*, *m2*, *oilprice*, *pce* and response: *pce*[27]
Source: Author's Calculations.

Forecast Horizon	*reg*	S.E.	*m2*	S.E.	*oilprice*	S.E.	*pce*	S.E.
1	.02	.03	.17	.09	.49	.09	.31	.07
2	.13	.08	.18	.09	.47	.10	.22	.05
3	.17	.09	.17	.09	.45	.11	.21	.05
4	.17	.09	.18	.09	.45	.11	.21	.05
5	.17	.09	.18	.09	.44	.11	.20	.05

Table 11: Forecast Error Variance Decomposition and Standard Errors for impulse: *reg*, *m2*, *oilprice*, *ppi* and response: *ppi*
Source: Author's Calculations.

[27] Decimals in these tables should sum to one, unless due to rounding.

Forecast Horizon	*reg*	S.E.	*m2*	S.E.	*oilprice*	S.E.	*ppi*	S.E.
1	.04	.05	.13	.08	.55	.09	.28	.06
2	.11	.07	.14	.08	.49	.09	.25	.06
3	.12	.08	.15	.08	.49	.09	.25	.06
4	.12	.08	.15	.08	.49	.09	.25	.06
5	.12	.08	.15	.08	.49	.09	.25	.06

Table 12: Forecast Error Variance Decomposition and Standard Errors for impulse: *reg*, *m2*, *oilprice*, *cpi* and response: *cpi*
Source: Author's Calculations.

Forecast Horizon	*reg*	S.E.	*m2*	S.E.	*oilprice*	S.E.	*cpi*	S.E.
1	.01	.02	.20	.09	.43	.09	.36	.07
2	.10	.07	.22	.10	.41	.10	.26	.06
3	.15	.09	.21	.09	.40	.10	.24	.06
4	.15	.09	.21	.09	.40	.10	.24	.06
5	.15	.09	.22	.09	.39	.10	.24	.06

3.5.3 Robustness Checks

Several robustness checks were performed to determine the strength of the relationships found in this paper. First, the ordering of the variables were experimented with. For example, the ordering of *m2* and *oilprice* were switched in these regressions and this did not significantly alter the results. Data from 2008 onwards was dropped, since the link between the monetary base and the price level in the post-2008 period was weakened. Again, there was no significant change in the core results. A second lag length was added and the results remained similar. Additional control variables were added, including stock market data and interest rates and spreads of various kinds.[28] These variables had little explanatory power in the model and so were excluded from the final analysis.

Notably, in the short-run SVAR estimates significant declines in the price level were found following monetary shocks. While this finding is somewhat counterintuitive, the idea that contractionary monetary policy will initially increases prices (and inversely that expansionary monetary policy will reduce prices) is consistent with the literature on the "price puzzle".[29] This finding provides further evidence of the robustness of the results and helps to validate the decision to employ M2 data, rather than M1 or M0, in the analysis.

3.6 Discussion

A key question that follows from the findings above is whether the relationship between the quantity of regulation and the price level is simply a statistical anomaly or instead represents a causal relationship. If the relationship is not an anomaly, a theory is needed to help explain how this has come about.

One possible channel by which regulation and the price level may be linked relates to political economy. In the complete set of IRFs appearing in Figure 34 through Figure 36 in Appendix B, one can see that there appears to be a relationship between regulation and the money supply in the short-run. These two time series (de-trended so as

---

[28] Examples of interest rates used include treasury yields and the Moody's seasoned Aaa and Baa corporate bond yields, as well as the spreads between these rates.
[29] For examples of studies that refer to the price puzzle, see Eichenbaum (1992) and Hanson (2004), among others.

to represent short-run shocks only) significantly predict changes in one another, and these variables generally granger cause one another as well.

One possible explanation for this is that there could be a link between activist regulatory policy and activist monetary policy. Perhaps administrations that are interventionary are likely to be interventionary along a number of different policy dimensions. Then regulation may be picking up some of the effect of other variables on inflation. Similarly, just as some administrations may be more activist, others might be more conservative, employing tighter monetary policy and issuing fewer regulations. To test this hypothesis, measures of federal government spending were included in SVARs to determine whether fiscal activism was also a significant predictor of inflation—it was not. Additionally, the *levels* of M2 and regulation do not appear to be cointegrated, suggesting any relationship between money and regulation is short-run only.

Another political explanation for the link between regulation and the price level may be that when inflation rises, the government feels a need to respond by "doing something." It is politically painful to reduce the money supply, so perhaps a relatively ineffective, but politically symbolic, measure taken by policymakers is to reduce regulation. This appears to be in line with the experience of the Council on Wage and Price Stability in the 1970s.

On the other hand, regulation may actually be influencing the price level. In the literature on inflation, there are a number of studies that examine the micro-level price dynamics in the data that underlie price indices.[30] This approach is imitated here by looking at per capita personal consumption expenditures on durable goods, which forms part of the construction of the PCE index.

Table *13* below presents the ordering of two SVARs involving durable goods spending. The first uses per capita durable goods spending in general. The second looks at a subcomponent of durable goods, spending on motor vehicles and parts.

Table 13: Durable Goods and Motor Vehicles and Parts

SVAR #	Ind. Variable (1)	Ind. Variable (2)	Ind. Variable (3)	Ind. Variable (4)	Main Dep. Variable (5)
4	*reg*	*m2*	*oilprice*	*pce*	*durable*
5	*reg*	*m2*	*oilprice*	*pce*	*cars*

As was the case in the earlier SVARs, the ordering of the variables here forms assumptions about which variables causally influence the other variables contemporaneously. All variables are again in nominal form, since the price level is controlled for.

[30] For examples of research that takes this approach using the CPI, see Bils and Klenow (2004), Klenow and Kryvtsov (2008), and Nakamura and Steinsson (2008), among others.

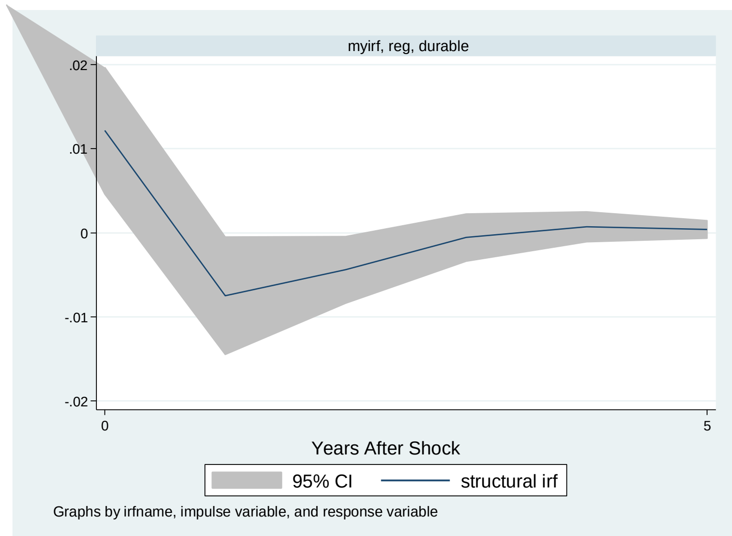


Figure 26: SVAR #4 - impulse(*reg*), response(*durable*)

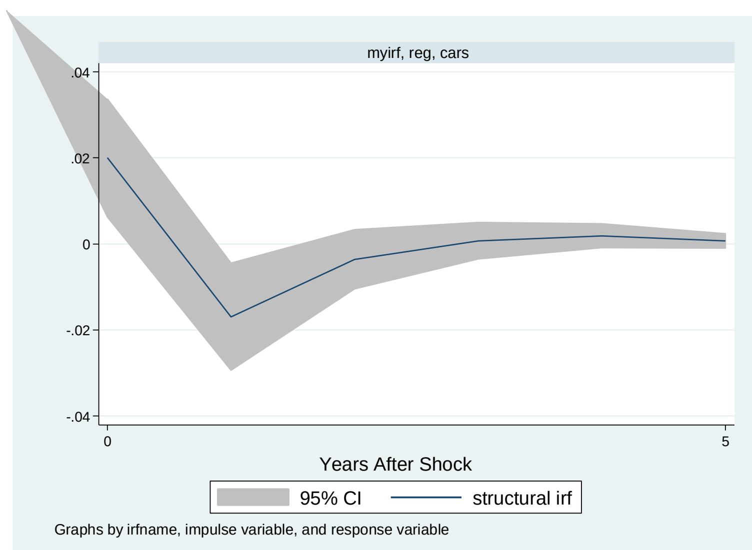


Figure 27: SVAR #5 - impulse(*reg*), response(*cars*)

In year zero (i.e. in the same period as a regulatory shock), there is a significant increase in spending on durable goods. This is an interesting finding because there is no significant increase in the price level in year zero under any of the three price indices.

The increase in spending on durable goods is followed by a just-significant (at the 5% level) decline in spending in years 1 and 2 following the shock. The decline in spending is even clearer when looking at per capita spending on motor vehicles and parts.

A decline in spending is consistent with the idea that regulation is a supply-side technology shock. Regulation increases prices, generally in years one and two following regulatory shocks, and these price increases are associated with declines in certain spending during those same periods.

Interestingly, however, spending on durable goods rises in the short run before falling. One explanation for this might be that some kind of intertemporal substitution is taking place. Perhaps consumers anticipate that regulations are going to drive up the prices of certain durable goods, such as automobiles. Consumers purchase products today and reduce spending one or two years out in the future when prices rise.

An alternative explanation is that regulations drive up the prices of products fairly quickly (perhaps due to rational expectations about rising prices in the future). Consumers spend more in response to the price change, perhaps because spending is relatively inelastic in the short-run. Thus, regulation may actually increase aggregate demand in the short run. Over time, however, consumers find substitutes or simply spend less due to their reduced real incomes. Spending falls as a result. In this case, the price increases of durables would have to be offset by declines elsewhere since there is no evidence of an aggregate increase in prices so quickly.

Notably, Granger causality tests and forecast error variance decompositions also support the case that regulations predict changes in per capita spending on durable goods and motor vehicles and parts. Regulation predicts about 20 percent of the forecast error variance in durable goods in the years after a shock. This is significant at the 98% confidence level. Similarly, regulation predicts about 19 percent of the forecast error variance in per capita spending on motor vehicles and parts, also significant at the 98% confidence level. By comparison the money supply predicts about 14-16 percent of these changes.

Notably, when one looks at per capita spending on nondurable goods, the subsequent decline in spending is not observed. Spending rises initially in the first year following a shock and then returns to normal. This suggests nondurable spending may be more inelastic than durable spending. It also means that regulation may not be a pure technology shock in the sense that it may increase demand for some goods and services while decreasing demand for others.

From the standpoint of economic theory, increases in regulation may increase aggregate demand while decreasing aggregate supply. In terms of Equation (25), V would rise. The effect on real output, y, is ambiguous however. It rises for some goods and falls for others. The price level, P, unambiguously rises because both positive demand shocks and negative supply shocks put upward pressure on the price level. This may be why the price level effect of regulation is appears quite robust in the data, while changes in spending are more ambiguous.

A more subtle point that flows from this analysis is that if price increases are occurring due to regulation, then adjustments should be made to indices of core inflation to account for these price changes. Presumably, core inflation is intended to account only for those price changes due to changes in the value of the monetary unit. Regulation may be an important overlooked source of bias in measures of inflation if it is a significant source of price increases in a manner similar to food and energy prices.

3.7 Conclusion

This paper explored how federal regulation impacts aggregate measures of the price level. The long-run impacts of the cumulative stock of all regulations and the short-run impacts of shocks to the flow of new regulations finalized each year are distinguished from one another. Using cointegration estimates, there is evidence that the stock of regulation is cointegrated with three measures of the price level. Using structural vector autoregression, analysis suggests shocks to the short-run flow of regulation significantly predict changes in the same three measures of the price level. Regulation predicts roughly 12-17 percent of the forecast error variance in the price level in the years following a regulatory shock, a magnitude that is only slightly smaller than what is explained by monetary shocks. The level of significance of these estimates varies by price index.

This relationship may be partly explained by micro-level spending changes on durable goods subcomponents, such as motor vehicles and parts. Regulation predicts about 19 to 20 percent of the forecast error variance in per capita spending on durable goods and subcomponents of durable goods, like motor vehicles and parts. Regulation may act like a technology shock in that positive regulatory shocks may raise prices and reduce spending on future goods. However, regulation may also act like a demand-shock in that it may cause intertemporal substation that increases spending in the short run, and may also increase spending on relatively price-inelastic goods, such as nondurables, when prices rise.

More research is needed to determine whether regulation is an important overlooked contributor to inflation. If the relationship holds, a theory of regulation and the price level should be further developed, and regulation should be considered in models, such as the various incarnations of the Fisher equation, that seek to explain inflation. Price changes due to regulation should also be excluded from measures of core inflation that seek to identify price changes only resulting from changes in the value of the monetary unit.

Appendix A: Raw Data vs. Smoothed Data with Time Trend Removed

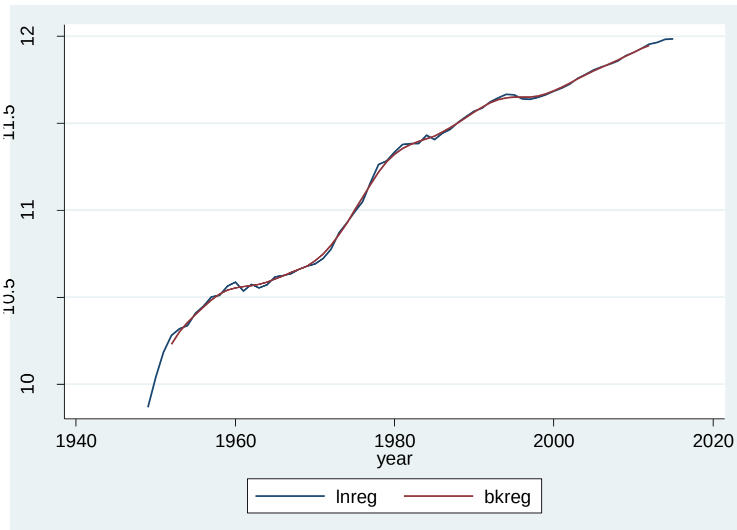


Figure 28: Log Regulation Level and Smoothed Trend

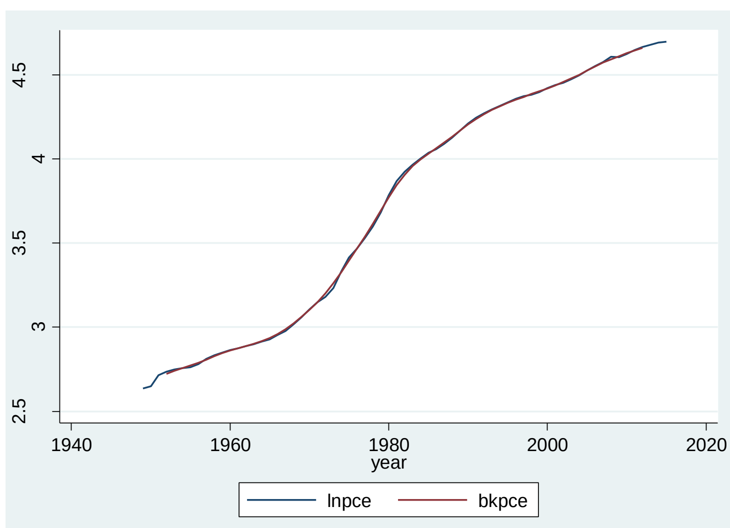


Figure 29: Log PCE Price Level and Smoothed Trend

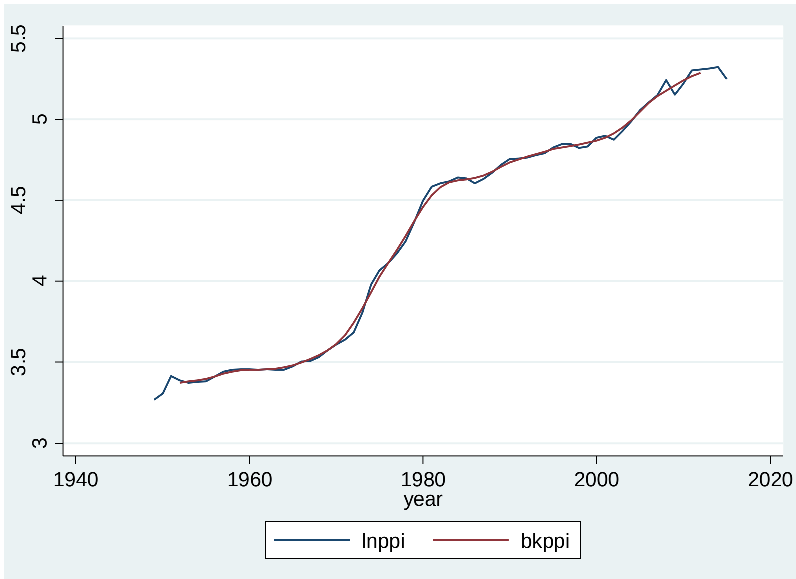


Figure 30: Log Producer Price Level and Smoothed Trend

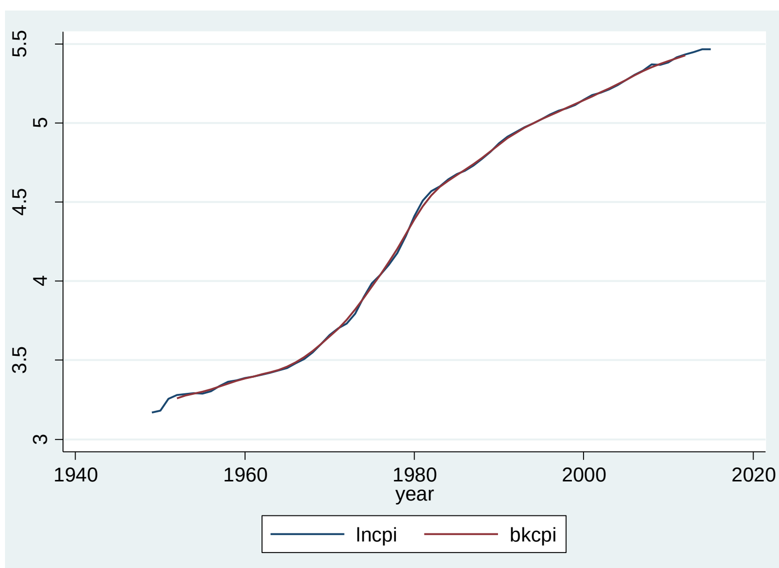


Figure 31: Log Consumer Price Index and Smoothed Trend

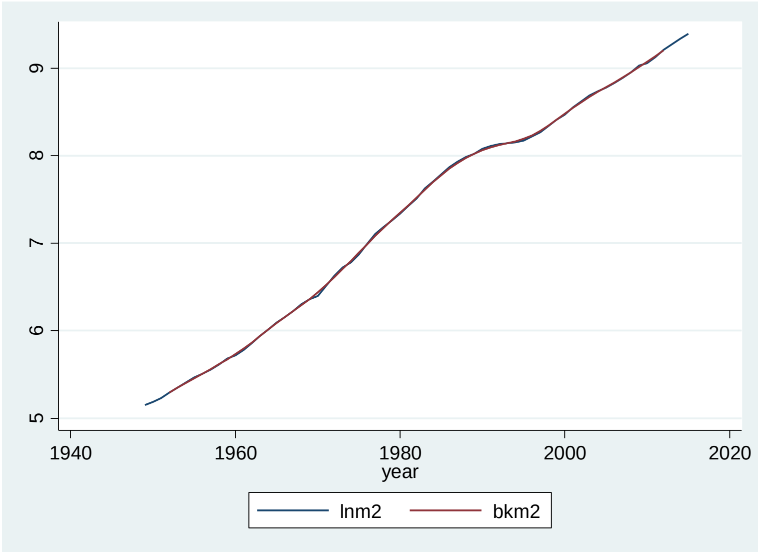


Figure 32: Log M2 Money Supply and Smoothed Trend

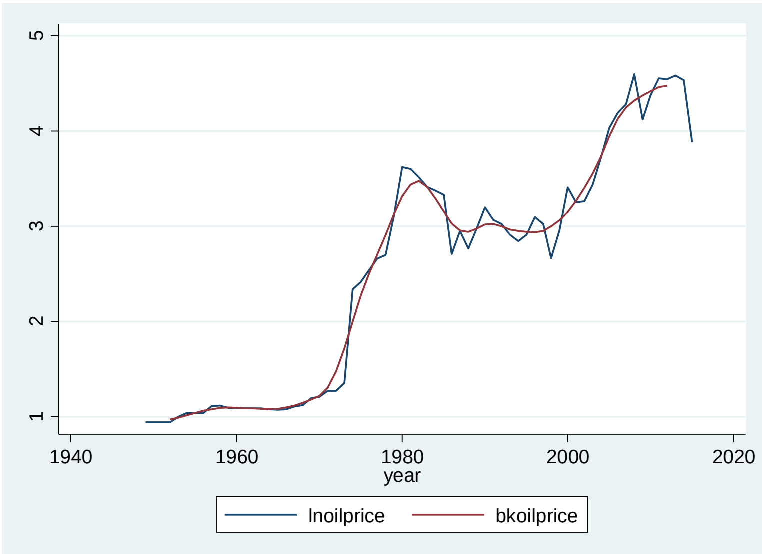


Figure 33: Log Oil Price and Smoothed Trend

Appendix B: Impulse-Response Functions

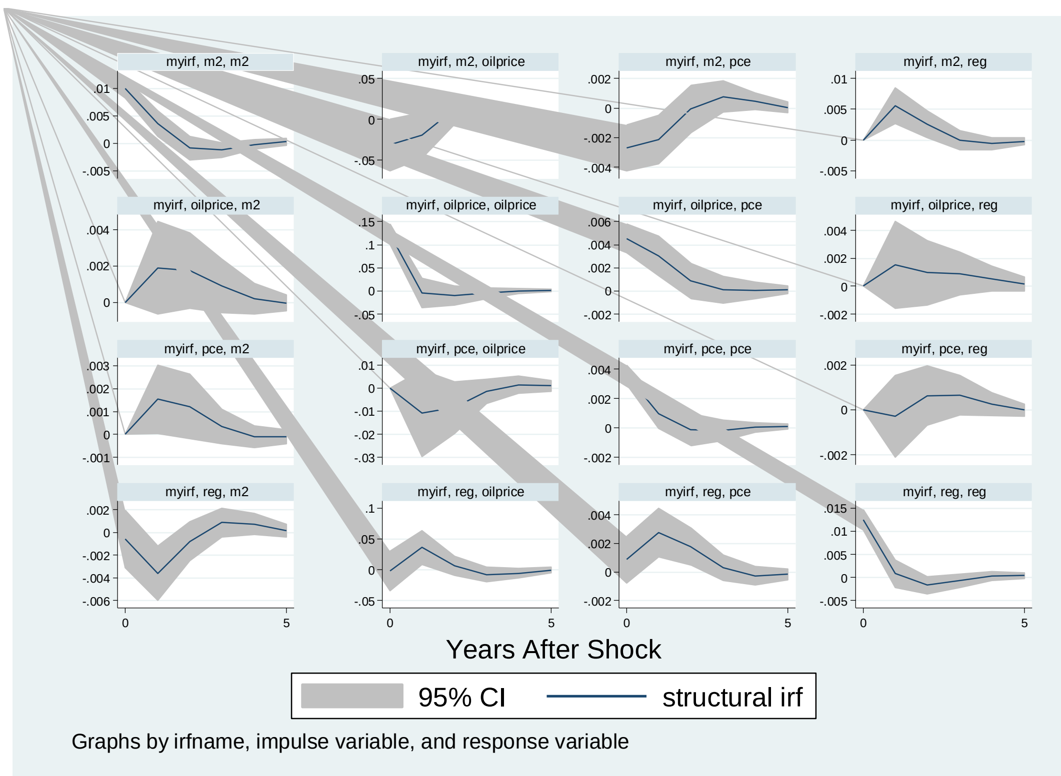


Figure 34: Impulse Response Functions for SVAR#1 ordered *reg*, *m2*, *oilprice*, *pce*

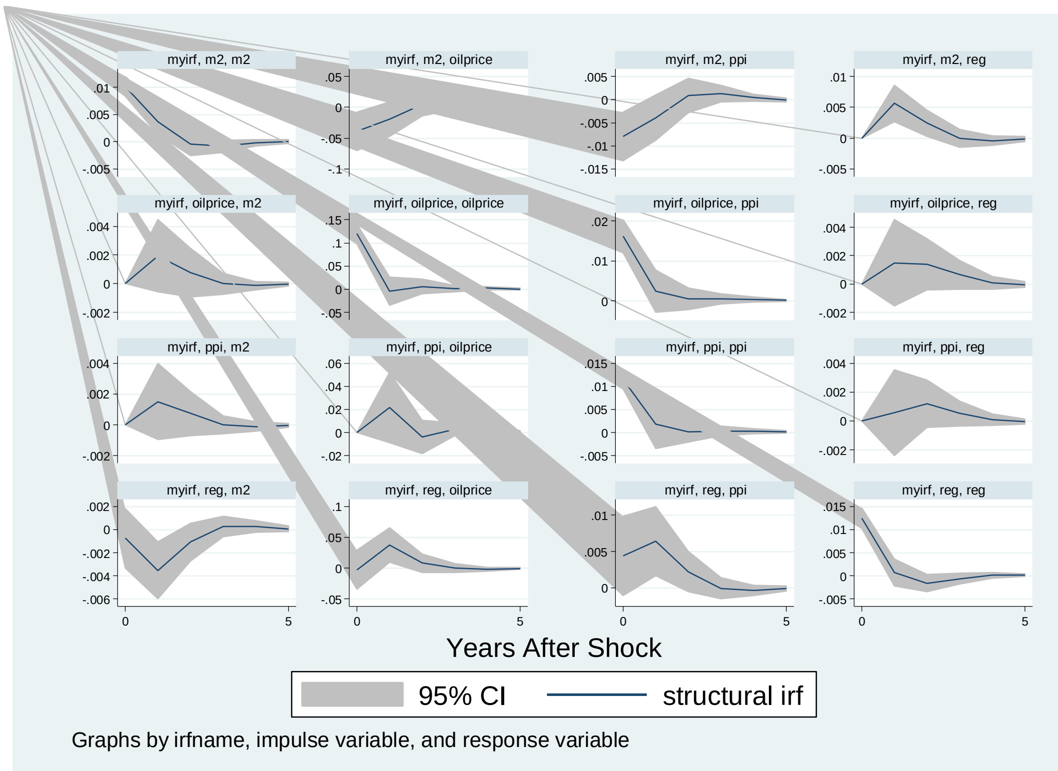


Figure 35: Impulse Response Functions for SVAR#2 ordered *reg*, *m2*, *oilprice*, *ppi*

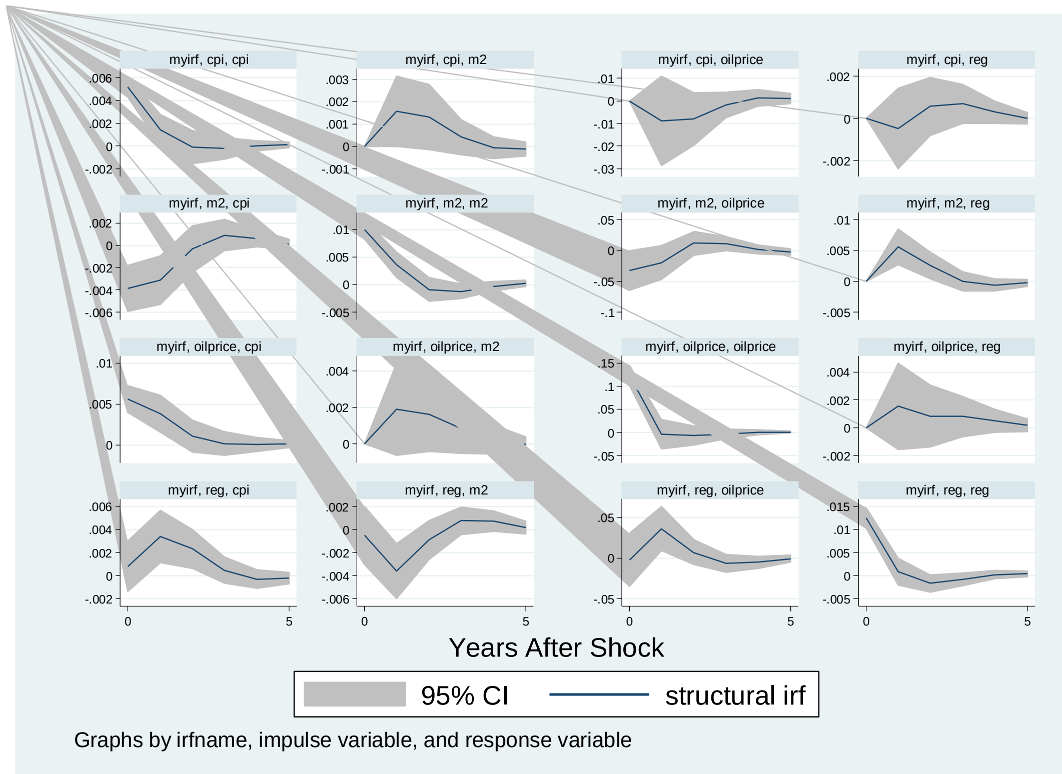


Figure 36: Impulse Response Functions for SVAR#3 ordered *reg*, *m2*, *oilprice*, *cpi*

References

Al-Ubaydli, O., & McLaughlin, P. A. (2015). RegData: A Numerical Database on Industry-Specific Regulations for all United States Industries and Federal Regulations, 1997–2012. *Regulation & Governance*.

Baxter, M., & King, R. G. (1999). Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series. *The Review of Economics and Statistics*, *81*(4), 575–593.

Beard, T. R., Ford, G., Kim, H., & Spiwak, L. (2011). *Regulatory Expenditures, Economic Growth, and Jobs: An Empirical Study* (Phoenix Center Policy Bulletin No. 28). Washington, D.C.: Phoenix Center for Advanced Legal & Economic Public Policy Studies.

Bernanke, B., Gertler, M., & Watson, M. (1997). Systematic Monetary Policy and the Effects of Oil Price Shocks | Brookings Institution. *Brookings Papers on Economic Activity*, *1997*(1). Retrieved from https://www.brookings.edu/bpea-articles/systematic-monetary-policy-and-the-effects-of-oil-price-shocks/

Bils, M., & Klenow, P. J. (2004). Some Evidence on the Importance of Sticky Prices. *Journal of Political Economy*, *112*(5), 947–985.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Chapter 2 Monetary policy shocks: What have we learned and to what end? In *Handbook of Macroeconomics* (Vol. Part A, pp. 65–148). Elsevier.

Coffey, B., McLaughlin, P., & Peretto, P. (2016). *The Cumulative Cost of Regulations* (Mercatus Working Paper). Arlington, VA: Mercatus Center at George Mason University.

Crafts, N. (2006). Regulation and Productivity Performance. *Oxford Review of Economic Policy*, *22*(2), 186–202.

Crews, W. (2016). *Ten Thousand Commandments 2016: An Annual Snapshot of the Federal Regulatory State*. Washington, D.C.: Competitive Enterprise Institute.

Dawson, J. (2006). Regulation, Investment, and Growth Across Countries. *Cato Journal*, *26*(3), 489–509.

Dawson, J. W. (2007). Regulation and the Macroeconomy. *Kyklos*, *60*(1), 15–36.

Dawson, J. W., & Seater, J. J. (2013). Federal Regulation and Aggregate Economic Growth. *Journal of Economic Growth*, *18*(2), 137–177.

de Rugy, V., & Davies, A. (2009). Midnight regulations and the Cinderella effect. *The Journal of Socio-Economics*, *38*(6), 886–890.

Djankov, S., McLiesh, C., & Ramalho, R. M. (2006). Regulation and Growth. *Economics Letters*, *92*(3), 395–401.

Dudley, S., & Warren, M. (2016). *Regulators' Budget from Eisenhower to Obama* (Regulators' Budget). Washington, D.C. and St. Louis: George Washington University Regulatory Studies Center and the Washington University of St. Louis Weidenbaum Center on the Economy, Government, and Public Policy.

Égert, B. (2016). Regulation, Institutions, and Productivity: New Macroeconomic Evidence from OECD Countries. *American Economic Review*, *106*(5), 109–113.

Eichenbaum, M. (1992). "Interpreting the macroeconomic time series facts: The effects of monetary policy." *European Economic Review*, *36*(5), 1001–1011.

Engle, R. F., & Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, *55*(2), 251–276.

Granger, C. W. J. (1981). Some Properties of Time Series Data and Their use in Econometric Model Specification. *Journal of Econometrics*, *16*(1), 121–130.

Hamilton, J. D. (1983). Oil and the Macroeconomy since World War II. *Journal of Political Economy*, *91*(2), 228–248.

Hamilton, J. D. (2003). What is an Oil Shock? *Journal of Econometrics*, *113*(2), 363–398.

Hanson, M. S. (2004). The "price puzzle" reconsidered. *Journal of Monetary Economics*, *51*(7), 1385–1413.

Hopkins, T., & Stanley, L. (2015). The Council on Wage and Price Stability: A Retrospective. *Journal of Benefit-Cost Analysis*, *6*(02), 400–431. https://doi.org/10.1017/bca.2015.41

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2), 231–254.

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, *59*(6), 1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Klenow, P. J., & Kryvtsov, O. (2008). State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation? *The Quarterly Journal of Economics*, *123*(3), 863–904.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Michaels, R. J., & DeVany, A. S. (1995). Market-Based Rates for Interstate Gas Pipelines: The Relevant Market and the Real Market. *Energy Law Journal*, *16*, 299–346.

Miller, J. C. (2011). The Early Days of Reagan Regulatory Relief and Suggestions for OIRA’S Future. *Administrative Law Review*, *63*, 93–101.

Nakamura, E., & Steinsson, J. (2008). Five Facts about Prices: A Reevaluation of Menu Cost Models. *The Quarterly Journal of Economics*, *123*(4), 1415–1464.

Nicoletti, G., Scarpetta, S., & Lane, P. R. (2003). Regulation, Productivity and Growth: OECD Evidence. *Economic Policy*, *18*(36), 9–72.

Office of Management and Budget. (2016). *2015 Report to Congress on the Benefits and Costs of Federal Regulations and Agency Compliance with the Unfunded Mandates Reform Act*. Washington, D.C.

Rasche, R. (1990). Demand Functions for Measures of U. S. Money and Debt. In *Financial Sectors in Open Economies: Empirical Analysis and Policy Issues*. Board of Governors of the Federal Reserve System.

Selgin, G. (1997). *Less than Zero: The Case for a Falling Price Level in a Growing Economy*. London: The Institute of Economic Affairs.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48.

Sinclair, T., & Vesey, K. (2012). *Regulation, Jobs, and Economic Growth: An Empirical Analysis* (Working Paper). Washington, D.C.: George Washington University Regulatory Studies Center.

Stock, J. H., & Watson, M. W. (2001). Vector Autoregressions. *The Journal of Economic Perspectives*, *15*(4), 101–115.

Biography

James Broughel grew up in Newton, Massachusetts. In 2011, he graduated summa cum laude with a BA and MA in economics from Hunter College of the City University of New York. He earned his PhD in economics from George Mason University in 2017. James is currently a Research Fellow for the State and Local Policy Project at the Mercatus Center at George Mason University. He specializes in the economic analysis of regulations and regulatory processes. James has authored numerous policy briefs and reports on regulatory issues. His writing has appeared in outlets such as U.S. News & World Report, Real Clear Policy, the Hill, and the Washington Examiner. He has published in scholarly journals, including the Harvard Journal of Law & Public Policy: Federalist Edition and the European Journal of Risk Regulation.