FEDERAL RESERVE BANK OF ST. LOUIS

REVIEW

THIRD QUARTER 2018
VOLUME 100 | NUMBER 3

REVIEW

Volume 100 • Number 3

FEDERAL RESERVE BANK *of* ST. LOUIS

CENTRAL TO AMERICA'S ECONOMY®

Review
Review is published four times per year by the Research Division of the Federal Reserve Bank of St. Louis. Complimentary print subscriptions are available to U.S. addresses only. Full online access is available to all, free of charge.

Online Access to Current and Past Issues

The current issue and past issues dating back to 1967 may be accessed through our Research Division website: **http://research.stlouisfed.org/publications/review**. All nonproprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* also are available to our readers on this website.

Review articles published before 1967 may be accessed through our digital archive, FRASER: **http://fraser.stlouisfed.org/publication/?pid=820**.

Review is indexed in Fed in Print, the catalog of Federal Reserve publications (**http://www.fedinprint.org/**), and in IDEAS/RePEc, the free online bibliography hosted by the Research Division (**http://ideas.repec.org/**).

Authorship and Disclaimer

The majority of research published in *Review* is authored by economists on staff at the Federal Reserve Bank of St. Louis. Visiting scholars and others affiliated with the St. Louis Fed or the Federal Reserve System occasionally provide content as well. *Review* does not accept unsolicited manuscripts for publication.

The views expressed in *Review* are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Subscriptions and Alerts

Single-copy subscriptions (U.S. addresses only) are available free of charge. Subscribe here: **https://research.stlouisfed.org/publications/review/subscribe/**.

Our monthly email newsletter keeps you informed when new issues of *Review*, *Economic Synopses*, *Regional Economist*, and other publications become available; it also alerts you to new or enhanced data and information services provided by the St. Louis Fed. Subscribe to the newsletter here: **http://research.stlouisfed.org/newsletter-subscribe.html**.

Copyright and Permissions

Economic Data

General economic data can be obtained through FRED® (Federal Reserve Economic Data), our free database with over 500,000 national, international, and regional data series, including data for our own Eighth Federal Reserve District. You may access FRED through our website: **https://fred.stlouisfed.org**.

ISSN 0014-9187

How Well Does Agency Theory Explain Executive Compensation?

George-Levi Gayle, Chen Li, and Robert A. Miller

As the share of all income going to the top 1 percent has risen over the past four decades, so has the share of top incomes coming from labor income relative to capital income. The rise in labor income is mainly due to the explosion in executive compensation over the same period—mostly because of the increase in executives being paid with stocks, options, and bonuses. The principal-agent model explains the reason for such compensation instead of a flat salary. Yet hundreds of papers in economics, finance, accounting, and management have reached no consensus on whether executive compensation is efficient or whether empirically it conforms to the prediction of the principal-agent theory. In this article, we argue that this lack of consensus is due to two issues: The first is a measurement issue, and the second is that the exact prediction of the principal-agent model depends on many objects unobservable to the econometrician. We illustrate how using theory-based estimation together with a model-motivated measure of total compensation can help overcome these issues. Finally, using a model-consistent measure of compensation and theory-based estimation, we conclude that executive compensation broadly conforms to the principal-agent theory; however, each situation and the variables used have to be carefully modeled, identified, and estimated. (JEL D82, L25, M12, M52)

Federal Reserve Bank of St. Louis *Review*, Third Quarter 2018, *100*(3), pp. 201-36.
https://doi.org/10.20955/r.2018.201-36

1 INTRODUCTION

Over the past three to four decades, income inequality in the United States has substantially increased. One measure of inequality used in academic research and the popular press is the share of all income that goes to the top income earners (especially the top 1 percent). In the United States, top income shares dropped dramatically from 1929 to 1950 but have increased dramatically since 1980.[1] In the early part of the twentieth century, top incomes were made up of primarily capital income; however, today's top incomes are divided 50/50 between labor and capital income. The increase from labor income is primarily from the

George-Levi Gayle is an associate professor at Washington University in St. Louis and a research fellow at the Federal Reserve Bank of St. Louis. Chen Li is an assistant professor at the Zicklin School of Business, Baruch College, CUNY. Robert A. Miller is the Richard M. Cyert and Morris DeGroot Professor of Economics and Statistics at the Tepper School of Business, Carnegie Mellon University.

explosion of executive compensation since 1980, paid mainly with firm-denominated securities, that is, stocks, options, and bonuses.[2]

Why are executives paid in firm-denominated securities? The principal-agent model is the main theoretical underpinning for why managers are compensated with stocks, options, and bonuses instead of a flat salary. The model captures the economic interactions of an uninformed party (the principal) who delegates tasks to the informed party (the agent) whose private action can affect both parties' benefits and whose interest is not perfectly aligned with the uninformed party. Modern firms are characterized by a dispersed ownership structure; the shareholders of a firm delegate the business operation to professional managers. Unlike the input of physical capital that can be easily measured, the input of managerial effort is hardly measurable and cannot be directly traded. A principal-agent problem, called moral hazard, arises when self-interested managers intend to secretly choose an effort level different from what would maximize the benefits of shareholders. To align interests, shareholders have to base executive compensation on the output of managerial effort, for example, the stock price. The unobservability of managerial effort is the main reason why executive compensation is largely based on firm equity rather than a flat salary.

Principal-agent models use techniques that characterize the optimal incentive mechanism for aligning the principal's and the agent's interests by simplifying assumptions about their preferences, technologies, and information structures. However, some uncertainties in the economy may also affect output, which risk-averse managers want to be insured against. Shareholders have to pay an extra amount as a risk premium to managers while balancing between incentives and insurance. An efficient compensation contract provides insurance at a sufficient amount that can guarantee the manager makes the effort that shareholders desire. In addition to the information asymmetry on effort, a manager may take advantage of his private information regarding the firm's state, which shareholders do not have access to. The optimal contract also has to provide the incentive for the manager to truthfully reveal the private information on the firm's state at an extra cost to the firm.

In contrast to the sophisticated, complex compensation schemes in the real world, there is a question of whether abstract principal-agent models can provide good explanations and predictions on executive compensation. Empirical research on managerial compensation seeks to examine whether the observed compensation schemes conform to an optimal contract supported by the principal-agent model. Furthermore, this line of research identifies and quantifies the effects of asymmetric information and assesses its impact on welfare, competition, and policy. Ultimately, if these models are a good description of the complex real-world compensation practices, they can be used to understand one aspect of the increase in inequality over the past four decades.

Empirically, there are three main ways of evaluating the output of a model. The first is to test a major prediction of the model while leaving unspecified the main structure of the model, for example, the positive-correlation test of agency theory. For examples of this approach for insurance markets, see Chiappori and Salanie (2000), Cardon and Hendel, (2001), and Cohen (2005), among others.[3] For examples for performance-pay settings, see Jensen and Murphy (1990), Hall and Liebman (1998), and Aggarwal and Samwick (1999), among others.[4] The second way of evaluating the output of a model is to specify the complete structure of the model and then derive over-identifying restrictions. A model that has empirical content imposes restrictions on the relationship between variables observed in the data. These restrictions can be used to recover the parameters of the model. If the number of independent restrictions that the model imposes on the observables is more than the number of parameters that need to be estimated, then the additional restrictions are called over-identifying restrictions. These over-identifying restrictions can be used to test the validity of the model. Without such restrictions, the model could be rationalized by any data. The final way is to specify the complete structure and perform out-of-sample validation. For example, say a researcher uses data from before 1980 to estimate a principal-agent model of executive compensation and then uses the estimated model to predict compensation after 1980. A test of the model would be to see if the model can predict the post-1980 rapid increase in executive compensation. The last two ways are the focus of theoretical-based estimation. Theoretical-based estimation is normally called for by the need to move beyond testing a model and to quantify welfare, efficiency, and the potential impact of policy reforms.

There are hundreds of papers in economics, finance, accounting, and management on whether executive compensation is efficient[5] and whether empirically it conforms to the prediction of the principal-agent model. Most of this research is based on testing one of the major predictions of the principal-agent model while leaving unspecified the main structure of the model. As of yet there is little conclusive evidence from this approach as to whether executive compensation packages are correctly structured or conform to the principal-agent model. In this article we argue that this is due to two issues: The first is a measurement issue, and the second is that the exact prediction of the principal-agent model depends on many objects unobservable to the econometrician. We illustrate how using theory-based estimation together with a model-motivated measure of total compensation can help overcome these issues. We conclude that using a model-consistent measure of compensation and theory-based estimation shows that executive compensation broadly conforms to the principal-agent theory; however, each situation and the variables used have to be carefully modeled, identified, and estimated.

The rest of the article is organized as follows. Section 2 takes up the issue of the measurement of total compensation, which is consistent with the principal-agent model and shows how this measurement differs significantly from what is used in most of the literature. It then applies this measurement concept to two distinct datasets covering over 60 years and documents how each component has changed. Section 3 takes up the issue of the unobservability of important aspects of the basic moral hazard model (called a pure moral hazard model from now on) and shows how theory-based estimation can be used to obtain measures of these important concepts. The pure moral hazard model is then estimated and used to answer this question: Why has executive compensation risen 10 times as fast as the pay of the average worker over the past 60 years? Section 4 illustrates that some aspects of executive compensation that seem to contradict the basic model can be easily reconciled with the more-general theory and hence the issue of marginal versus joint distribution of variables needs to be considered more carefully when choosing a model. Section 5 concludes and gives some direction for future research.

2 HOW IS EXECUTIVE COMPENSATION MEASURED?

The cost to shareholders of employing a manager, called direct compensation, is the sum of salary and bonuses, the value of restricted stocks and options granted, and the value of retirement and long-term compensation schemes. The discounted sum of these direct-compensation items measures the reduction in the firm's value from outlays to management. Total compensation to a manager is defined as direct compensation plus changes in wealth from holding firm options and changes in wealth from holding firm stock. To compute the remaining two components in total compensation, one must address where managers would place this wealth if it were not held in their firms' financial securities. We assume that managers would hold a well-diversified portfolio instead, an implication of our model. When forming their portfolio of real and financial assets, managers recognize that part of the return from their firm-denominated securities should be attributed to aggregate factors, so they reduce their holdings of other stocks to neutralize those factors. Hence, the change in wealth from holding their firms' stock is the value of the stock at the beginning of the period multiplied by the abnormal return.

The principal-agent model implies that changes in wealth from holding firm options and changes in wealth from holding firm stock both have mean zero. An efficient contractual arrangement would not induce a risk-averse agent to hold more risk than is absolutely necessary, because any additional risk held by the risk-adverse agent would have to be compensated for by a risk premium. Therefore, all risk beyond the agent's control should be netted out of the compensation. This can be done by allowing the agent to hold a well-diversified market portfolio, which is equivalent to netting out the market portfolio and any predictable component of the firm's securities. Therefore, both from the manager's and firm's perspectives, the netting out of the return on the market portfolio is desirable and, hence, the expected values of the change in wealth from holding firm options and change in wealth from holding firm stock are zero. This implies that direct and total compensation have the same expected value. Therefore, whether risk-neutral shareholders minimize expected total compensation or expected direct compensation is moot. However, changes in wealth from holding firm stock and options reflect the costs a manager incurs from not being able to fully diversify his wealth portfolio because of restrictions on stock and option sales. Consequently, managers care about total compensation, not direct compensation, because the former determines how their wealth changes from period to period when they optimally smooth their consumption over the life cycle and make optimal portfolio choices.

A third measure of compensation, called constrained compensation, is the sum of cash, bonuses, and the value of restricted stock and option grants, plus the change in the value of restricted stock and grant holdings. Constrained compensation exposes the manager to aggregate risk to the degree that the firm's share price fluctuates with the market. Rational managers would neutralize their market risk by reducing their holdings of the market portfolio to compensate for the additional market risk that holding restricted stock entails. Suppose managers held no diversified stock after receiving their compensation and were prevented from selling futures in the market portfolio (maturing when their firm-specific securities can be redeemed through sales). Then we might conclude compensation is based on market returns if cash and bonus payments were not sufficiently countercyclical to offset the manager's aggregate risk of holding a portfolio of his firm's financial securities. We are unaware of any evidence showing that the wealth portfolio of a manager is constrained by his own shareholders to hold more market risk than he voluntarily chooses. This explains why the measure of compensation most consistent with the principal-agent model is total compensation rather than constrained compensation.

2.1 The Income-Equivalent Measure of Total Compensation

This section presents techniques for estimating a current income-equivalent measure of total compensation, which follow Antle and Smith (1985, 1986), Hall and Liebman (1998), and Margiotta and Miller (2000). The current income equivalent is defined to be the amount of before-tax dollars that an executive would require to offset exactly the value of the compensation package received in a given year. The term "compensation package" refers to the before-tax value of salaries, short-term bonuses, deferred-to-retirement bonuses, stockholdings, stock bonuses, stock options, dividend units, phantom shares, pension benefits, savings-plan contributions, long-term performance plans, and any other special items (such as a loan to the executive made at a below-market rate). In the absence of evidence to the contrary, the following assumptions underlie the estimation procedures: (i) executives remain with their firms until retirement at age 80; (ii) all non-contingent, deferred compensation is sure to be received; (iii) salary levels are not expected to fall; and (iv) an executive does not possess inside information regarding future stock prices or the probability that he or she will die in any given year.

2.1.1 Data Construction Details. In this article, we use data from two sources. The first dataset covers the years 1944-78, and the details of how it is constructed can be found in Antle and Smith (1985, 1986) and Margiotta and Miller (2000). The second dataset covers the years 1993-2009. Below we provide some essential details on its construction.

Firm type is defined as a combination of the industrial sector and firm characteristics for each firm in each year. The data used to measure firm characteristics are from Compustat. First, we classify the whole sample into three industrial sectors according to the Global Industry Classification Standard (GICS) code. The primary sector includes firms in the energy (GICS code 1010), materials (GICS code 1510), industrials (GICS codes 2010, 2020, 2030), and utilities (GICS code 5510) sectors. The consumer goods sector includes firms in the consumer discretionary (GICS codes 2510, 2520, 2530, 2540, 2550) and consumer staples (GICS codes 3010, 3020, 3030) sectors. The services sector includes firms in the health care (GICS codes 3510, 3520), financial (GICS codes 4010, 4020, 4030, 4040), and information technology and telecommunication services (GICS codes 4510, 4520, 5010) sectors.

We use raw stock prices and adjustment factors from the Compustat PDE dataset. For each firm in the sample, we calculate monthly compounded returns adjusted for splitting and repurchasing for each fiscal year; we then subtract the return to a value-weighted market portfolio (NYSE/NASDAQ/AMEX) from this raw return to determine the net excess return for the firm's corresponding fiscal year. We drop firm-year observations if the firm changed its fiscal year end such that all compensation and stock returns are based on 12 months and consequently comparable with each other. The excess return is obtained by adding the total compensation in the fiscal year (scaled by the firm's value at the beginning of the fiscal year) to the net excess returns in the same firm year.

2.1.1.1 Compensation. In addition to the total compensation included in Compustat ExecuComp data, we also calculate the holding value of firm-specific equities. Due to data limitations, we cannot observe for each sample year all the inputs of the Black-Scholes formula for grants carried from before 1993, the beginning year of our sample. Compustat ExecuComp provides the valuation information only for those options newly granted after 1993, including the number of underlying stock shares, exercised prices, expiration dates, and issue dates. However, we need to know these Black-Scholes inputs for options granted before 1993 to completely value the wealth change of CEOs by estimating the value of unexercised options and updating them each year. To facilitate the calculation, we assume that (i) all options are exercised on their expiration dates, (ii) stock options granted before 1993 are exercised in a first-in first-out fashion, and (iii) each CEO holds his own stock options granted before 1993 for a period of the average length of the holding period across all years when he is in the sample. Consequently, we can back out the issue dates and exercised prices for options granted before 1993 for each CEO. The same routine applies to nonzero options granted before the CEO entered our sample. We then apply the dividend-adjusted Black-Scholes formula to reevaluate the call options for each CEO in each year. The dividend-adjusted Black-Scholes formula used is as follows: Let c denote the call option value, K the exercise price, T_m the time to maturity (in years), S the underlying security price, q the dividend yield, r the risk-free rate, and σ the implied volatility. Let $N(\cdot)$ denote the standard normal cumulative distribution function. Then the call option value is given by

$$c = Se^{-qT_m} N(d_1) - Ke^{-rT_m} N(d_2), \tag{1}$$

$$d_1 = \frac{\ln(S/K) + (r - q + \sigma^2/2)T_m}{\sigma\sqrt{T_m}}, \tag{2}$$

and

$$d_2 = d_1 - \sigma\sqrt{T_m}. \tag{3}$$

Following the concept of income-equivalent total compensation defined above, we construct the total compensation by adding the change in wealth from options held and stock held to the other components of compensation included in ExecuComp.

2.1.2 Documentation of the Changes in Components of Three Different Samples. Table 1 summarizes and compares the distribution of the five main compensation components among three samples. The components include salary and bonuses, the value of options granted, the value of restricted stock granted, the change in wealth from options held, and the change in wealth from stock held. The remaining unlisted components include retirement and long-term compensation. The "Old" sample covers the years 1944-78. The other two

Table 1

Cross-Sectional Information on Components of Compensation in Thousands of US$ (2000)

Variable	Rank	Compensation Old	New Restricted	New All
Salary and bonuses	All	151 (68)	672 (576)	707 (1,036)
	CEO	151 (67)	1,199 (833)	1,176 (1,674)
	Non-CEO	146 (75)	530 (373)	584 (739)
Value of options granted	All	29 (104)	2,170 (7,184)	2886 (12,198)
	CEO	29 (105)	5,015 (12,432)	5,967 (18,263)
	Non-CEO	29 (93)	1,402 (4,593)	2,079 (9,861)
Value of restricted stock granted	All	0.0078 (0.0679)	242 (720)	306 (1,622)
	CEO	0.0085 (0.0708)	551 (1,310)	637 (2,097)
	Non-CEO	0.0001 (0.0006)	159 (404)	219 (1,460)
Change in wealth from options held	All	10 (284)	141 (6,131)	–235 (13,040)
	CEO	12 (286)	414 (10,503)	–479 (21,028)
	Non-CEO	–17 (257)	68 (4,239)	–171 (9,937)
Change in wealth from stock held	All	12 (896)	211 (12,144)	21 (20,170)
	CEO	0.7 (826)	632 (21,741)	109 (34,720)
	Non-CEO	142 (1,484)	98 (7,733)	–3 (14,055)

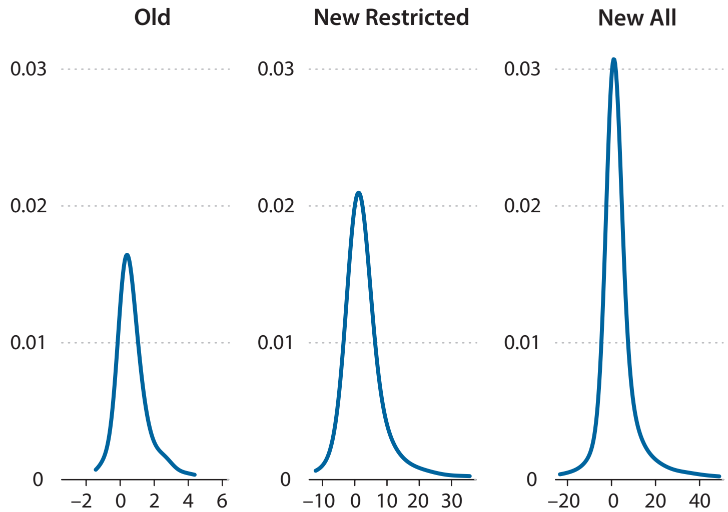
NOTE: Standard deviations are in parentheses.

samples cover the years 1993-2009. The Old sample and the "New Restricted" sample include the same industries, that is, aerospace, chemicals, and electronics. The "New All" sample includes all industries in the dataset from merged ExecComp, COMPUSTAT Fundamentals Annual, and CRSP[6] monthly data.

Two persistent patterns arise from the comparison. Regardless of the time period and executive rank, the level of total compensation mainly depends on the first three components explicitly specified in compensation contracts, including salary and bonuses, the value of options granted, and the value of restricted stock granted. The wealth change, either from

Figure 1

Kernal Estimates of the Density of Total Compensation



holding options or from holding stock, contributes less. However, the variation in total compensation is mainly driven by the last two components, which are based on wealth changes from holding firm-specific equity.

In addition, some time-series variations across the three samples are worth noting. Both the absolute level and the relative level of the compensation components change over time. First, all the components increase from the old period to the new period. For the Old and New Restricted samples, which cover the same three industries, Table 1 shows that the increase in total compensation is dominated by the equity-based components. Salary and bonuses increase almost four and half times in the three industries, which is the smallest increase among the five components. Second, the relative weights of these components change over time as well. We observe a dramatic increase in the importance of equity-based compensation. In the Old sample, cash-based compensation (salary and bonuses) is almost three times the size of equity-based compensation (the sum of the value of options granted and the value of stock granted) in the three industries, but it becomes only about one-quarter of the latter in the New Restricted sample. Thus, total compensation has increased much faster than salary and bonuses in the three industries. The component contributing the most to this dramatic shift is the options granted to managers, valued using the Black-Scholes formula. In both the restricted and unrestricted samples, the value of options granted is the biggest component of managerial compensation. In addition, the growth of stock compensation outperforms that of options compensation, even though it accounts for a smaller portion of total compensation. The value of options granted increases by more than 170 times for CEOs and by about 50 times for non-CEOs, and the value of restricted stock granted increases from almost nothing for all executives to $551,000 for CEOs and $159,000 for non-CEOs on average.

The third pattern is about the change in dispersion. The five components become more dispersed from the old period to the new period. Holding financial securities in their own firms rather than a well-diversified market portfolio exposes managers to considerable uncertainty. Table 1 shows that changes in wealth from holding stock and changes in wealth from holding options are more dispersed than any other component. The standard deviation is higher than for cash and bonuses, options granted, and stock granted. Note that the standard deviations of these components have dramatically increased—wealth changes in stocks and options by more than 100 fold. The two components account for a considerable amount of the increase in the volatility of total compensation. The value of options granted also contributes to a significant degree to the volatility of total compensation in the new period.

Figure 1 illustrates the distributional differences among the three samples. First, the dispersion of total compensation increases over time. The standard deviation of total compensation in the new period is several times as much as that in the old period. What's more, a significant portion of CEOs have negative compensation, even though the distribution presents a longer right tail. The negative compensation mainly stems from the change in wealth from stock held and the change in wealth from options held. To summarize, managerial compensation has substantially increased in real terms and become more dispersed. This has been accomplished by a dramatic increase in stock option grants.

3 WHAT IS THE RELATIONSHIP BETWEEN FIRM SIZE AND EXECUTIVE COMPENSATION?

The dramatic increase in both the level of CEO compensation and its sensitivity to firm performance over the past 50 years is widely documented.[7] These studies show that, of all the components making up executive pay—including cash, bonuses, stock grants, and retirement benefits—the biggest increases have been in option grants. Thus, much of the increase in managerial compensation is attributable to increases in asset grants whose value is explicitly tied to the value of the firm. Since moral hazard explains why managerial compensation and firm performance should be connected, it is tempting to suggest that changes in the nature of moral hazard might have triggered these trends.

The theory of moral hazard provides a plausible transmission mechanism for connecting the compensation paid to a firm's executives with the returns on their firm's assets. There are two channels for inducing secular changes in managerial compensation within the principal-agent paradigm. First, contracts reflect heterogeneity across firms, such as their size, their capital-to-labor ratios, the sectors they belong to, and the dispersion of their financial returns. Consequently, changing the heterogeneity across firms induces changes in the aggregate level and variability of compensation. Second, the optimal contract is a function of the preferences and risk attitudes of managers. Changing those preferences also affects the probability distribution of compensation across executives. This section summarizes the results from Gayle and Miller (2009a), who estimate a model of moral hazard with data spanning a 60-year period in order to investigate how well these two channels explain secular changes in managerial compensation and to assess their relative importance. We then contrast their findings with others papers in the literature. In this section, we demonstrate, using a simple moral hazard model and the data above, that the change in firm size is responsible for most of these changes observed over time.

3.1 The Relationship Between Firm Size and the Different Components of Total Compensation

The positive relationship between firm size and pay for ordinary workers is one of the most robust empirical finding in labor economics (Idson and Oi, 1999). As documented by Gayle, Golan, and Miller (2015), this is also true in the executive labor market. However, executive compensation has many more components than the pay of ordinary workers. But which component of executive compensation is responsible for the positive correlation between total compensation and firm size? Can the increase in firm size over time explain the increase in compensation over time?[8]

Table 2 presents the results of measures of firm size, total assets, and the number of employees on three of the basic components of total compensation—salary and bonuses, the value of options granted, and the value of restricted stock granted. The results are presented for three different samples and for CEOs and non-CEOs separately. For the New All sample, we observe positive and significant relationships between both measures of firm size and all three components of compensation for both CEOs and non-CEOs. When the sample is restricted to the industries in the Old sample, the same positive and significant relationships are observed, with the exception of the relationship between the number of employees and the value of restricted shares granted to CEOs, which is positive but not statistically significant. However, the positive and statistically significant relationships between firm size and the components of total compensation are not ubiquitously present in the Old sample. This gives reason to pause when considering the conclusion that the increases in the level of executive compensation over these periods are driven by an increase in firm size over time.

The most fundamental prediction of the principal-agent model for executive compensation is that in order to align shareholders' interests with the interests of the executive, the executive's compensation should be tied to the output of the firm. In practice, the change in wealth of the manager from holding firm-denominated securities is the main instrument for achieving this goal. Therefore, Table 3 presents the results of some basic regressions of the empirical measures of the change in the wealth of executives from holding options and restricted stocks of the firm on excess returns on the firm's stocks, firm size measures, and interactions of excess returns and these measures of firm size. The results show that the basic prediction of the principal-agent model is borne out by the data, that is, that there is a positive relationship between firm performance and the change in executive wealth from holding firm-denominated securities. However, like the results in Table 2, the positive relationship between firm size and the sensitivity of executive wealth to firm performance is only robust in the New All sample. This could be for a number of reasons. First, it could be that the other sample sizes are just too small to draw any conclusion. Second, it could be that the regression does not properly control for all the elements the theory predicts. Below we will use a fully specified principal-agent model to see whether we can provide more conclusive evidence from the samples we have.

Table 2

Regression of Compensation Components on Firm Characteristics

A: CEO

	New All			New Restricted			Old		
	Salary and bonuses	Value of options granted	Value of restricted stock granted	Salary and bonuses	Value of options granted	Value of restricted stock granted	Salary and bonuses	Value of options granted	Value of restricted stock granted
Total assets	0.006	0.037	0.008	0.018	0.396	0.035	–0.002	0.00286	3.28e-10
	(0.0002)	(0.003)	(0.0003)	(0.005)	(0.077)	(0.009)	(0.004)	(0.008)	(4.7e-09)
Number of employees	4.236	37.79	4.46	9.953	70.03	1.208	1,623	223.8	–4.0e-04
	(0.239)	(2.63)	(0.29)	(1.163)	(19.39)	(2.287)	(110.6)	(240.8)	(1.0e-04)
Observations	19,599	19,599	19,599	1,000	1,000	1,000	753	753	753

B: Non-CEO

	New All			New Restricted			Old		
	Salary and bonuses	Value of options granted	Value of restricted stock granted	Salary and bonuses	Value of options granted	Value of restricted stock granted	Salary and bonuses	Value of options granted	Value of restricted stock granted
Total assets	0.004	0.022	0.004	0.014	0.090	0.009	–0.016	–0.005	5.71e-11
	(5.1e-05)	(0.001)	(1.1e-04)	(0.001)	(0.017)	(0.002)	(0.012)	(0.02)	(1.76e-11)
Number of employees	1.685	12.15	1.872	2.330	30.32	1.561	2,235	922	–9.1e-06
	(0.052)	(0.764)	(0.116)	(0.292)	(4.28)	(0.371)	(571)	(958)	(8.4e-06)
Observations	75,379	75,379	75,379	3,693	3,693	3,693	68	68	68

NOTE: Standard errors are in parentheses. All regressions are controlled for the industry, year fixed effect, and debt-to-equity ratio.

Table 3

Regression of Wealth Change on Firm Characteristics and Excess Returns

	New All		New Restricted		Old	
	Change in wealth from options held	Change in wealth from stock held	Change in wealth from options held	Change in wealth from stock held	Change in wealth from options held	Change in wealth from stock held
A. CEO						
Total assets	0.013	0.004	0.184	0.017	1.81e-06	5.03e-06
	(0.003)	(0.005)	(0.084)	(0.142)	(2.2e-05)	(6.0e-05)
Number of employees	14.55	–1.474	–1.932	–7.699	–1.122	–1.214
	(3.06)	(4.837)	(21.15)	(35.43)	(0.642)	(1.758)
Excess return	12,427	21,280	8,877	16,243	438	1,893
	(335)	(530)	(1,181)	(1,979)	(79.5)	(218)
Excess return sq.	–817	–1161	–3448	9,419	–285	–467
	(45.17)	(71.42)	(1,581)	(2,650)	(105)	(292)
Excess return × total assets	0.082	0.038	0.487	–0.671	7.6e-06	–4.3e-04
	(0.011)	(0.017)	(0.337)	(0.565)	(8.6e-05)	(2.4e-05)
Excess return sq. × total assets	0.025	0.004	0.153	–2.453	–4.1e-05	1.89e-04
	(0.011)	(0.017)	(0.813)	(1.363)	(1.0e-04)	(2.8e-04)
Excess return × No. emp.	158	128	167	143	1.143	13.82
	(10.17)	(16.08)	(76.73)	(128.6)	(2.46)	(6.75)
Excess return sq. × No. emp.	–26.08	–8.696	19.02	605	17.19	5.676
	(5.43)	(8.581)	(186)	(312)	(2.951)	(8.084)
Observations	19,599	19,599	1,000	1,000	753	753
B. Non-CEO						
Total assets	0.006	0.0001	0.003	0.016	–3.9e-5	6.12e-5
	(0.000756)	(0.00113)	(0.0174)	(0.0322)	(0.000116)	(0.000314)
Number of employees	3.700	–0.370	9.958	–7.072	0.891	3.770
	(0.726)	(1.083)	(4.446)	(8.202)	(4.687)	(12.72)
Excess return	4,289	5,017	2,593	2,559	182	511
	(78.85)	(117.5)	(237.0)	(437.1)	(505.1)	(1,370.6)
Excess return sq.	–339	–148	–976	1,363	–302	1,715
	(13.37)	(19.92)	(320)	(591)	(1,149)	(3,118)
Excess return × Total assets	0.040	0.027	0.427	–0.041	3.1e-04	0.003
	(0.003)	(0.004)	(0.068)	(0.126)	(3.2e-04)	(8.7e-04)
Excess return sq. × Total assets	–0.002	–0.002	0.300	–0.535	4.4e-04	0.005
	(2.0e-04)	(3.0e-04)	(0.164)	(0.302)	(0.002)	(0.004)
Excess return × No. emp.	41.47	29.47	–55.03	72.72	–12.59	–51.33
	(2.36)	(3.52)	(16.31)	(30.08)	(15.47)	(41.98)
Excess return sq. × No. emp.	–1.936	1.094	–76.01	235.3	1.020	–87.70
	(0.761)	(1.134)	(39.86)	(73.53)	(69.66)	(189)
Observations	75,379	75,379	3,693	3,693	68	68

NOTE: No. emp., number of employees. sq., squared. Standard errors are in parentheses. The regressions also include the debt-to-equity ratio interacted with total assets and the number of employees.

3.2 A Basic Model for Inference

The importance of moral hazard can be characterized three ways: the gross loss shareholders would incur (before accounting for managerial compensation) from the manager tending his own interests, the benefits accrued to the manager from tending his own interests instead of shareholder interests, and how much the shareholders are willing to pay to eliminate the problem of moral hazard altogether.

The first measure, denoted τ_1, is the expected gross-output loss to the firm from switching from the distribution of abnormal returns for the diligent work to the distribution for shirking, that is, the difference between the expected firm output from the manager pursuing the firm's goals versus his own before netting out expected managerial compensation. Let v denote the value of the firm at the beginning of the period, and let x denote the firm's abnormal returns realized at the end of the period. Following the convention in the economic literature, we describe a manager who pursues the interests of the firm as "working" and a manager who pursues his own interests, when compensation is independent of firm performance, as "shirking." Then

$$\begin{aligned}\tau_1 &= E\left[x \mid manager\ works\right]v - E\left[x \mid manager\ shirks\right]v \\ &= -E\left[x \mid manager\ shirks\right]v,\end{aligned}$$

where the second equality exploits the identity that the expected value of abnormal returns is zero when the manager is working (pursuing the interests of the firm).

The second measure, τ_2, is the nonpecuniary benefits to the manager from shirking, that is, pursuing his own goals within the firm. Let w_2 denote the manager's reservation wage to work under perfect monitoring or if there were no moral hazard problem, and let w_1 denote the manager's reservation wage from shirking. Then τ_2, the compensating differential for these two activities, can be expressed as the difference

$$\tau_2 = w_2 - w_1.$$

We also estimate the maximum amount shareholders are willing pay to eliminate the moral hazard problem—the value of a perfect monitor. Absent moral hazard, the firm would pay the manager the fixed wage, w_2, instead of according to compensation, $w(x)$. The firm's willingness to pay for eliminating the moral hazard problem, denoted τ_3, is accordingly defined as

$$\tau_3 = E\left[w(x)\right] - w_2. \tag{4}$$

This measure is actually a lower bound on the shareholders' willingness to pay for a perfect monitor because it is based on asking the manager to perform the same tasks. If, however, the manager's actions could be monitored perfectly, it is plausible that shareholders would modify the manager's job description to better exploit the monitoring technology for the benefit of the firm, an issue analyzed in Prendergast (2002).

Against the output reduction from shirking, τ_1, is the savings in managerial compensation coming from two terms: the shadow value of a perfect monitor and the cost of inducing the manager to work diligently when a perfect monitor is removed. Subtracting from τ_1 the sum of τ_2 and τ_3, we obtain the net income loss a firm would sustain from signing a shirking contract with a manager. This net amount represents the value of preventing the manager from undoing contracts that align his incentives with the firm's by dealing with a lender who does not recognize the folly of allowing the manager to insure himself against poor firm performance and is unaware of public disclosure laws that require the manager to report his holdings of firm-related securities.

3.2.1 A Model of Pure Moral Hazard. This section lays out a theoretical principal-agent framework on which our empirical analysis is based. At each time period *t*, there are three activities in which a person can be engaged: working as the firm manager in the shareholders' interests, being employed as a manager at the firm but pursuing interests different from the shareholders', or not being engaged by the firm. Let $l_t \equiv (l_{0t}, l_{1t}, l_{2t})$ denote the three possible activities, where $l_{jt} \in \{0,1\}$ is an indicator for choice $j \in \{0,1,2\}$ and

$$\sum_{j=0}^{j=2} l_{jt} = 1.$$

The indicator $l_{0t} = 1$ denotes that the manager is not employed by the firm, $l_{1t} = 1$ denotes shirking, and $l_{2t} = 1$ denotes working diligently. While l_{0t} is common knowledge, the values of (l_{1t}, l_{2t}) are hidden from the shareholders. Apart from choosing his activity, the manager also chooses his consumption for the period. Let c_t denote the manager's consumption in period *t*. We assume that preferences over consumption and work are parameterized by a utility function exhibiting absolute risk aversion that is additively separable over periods and multiplicatively separable with respect to consumption and work activity within periods. In the model we estimate, lifetime utility can be expressed as

$$-\sum_{t=0}^{\infty} \sum_{j=0}^{3} \alpha_j \beta^t l_{tj} \exp(-\rho c_t),$$

where β is the constant subjective discount factor, α_j are utility parameters associated with setting $l_{jnt} = 1$, and ρ is the constant absolute level of risk aversion. We set $\alpha_0 = 1$ as a normalization, since behavior is invariant to linear transformation of the utility function under the independence axiom. We assume that $\alpha_2 > \alpha_1$, or that diligence is more distasteful than shirking. This assumption is the vehicle by which the manager's preferences are not aligned with shareholders' interests. We are not suggesting that managers are inherently lazy, merely that their personal goals do not motivate them to maximize the value of the firm if their compensation is independent of the firm's performance. Finally, we require $\alpha_1 > 0$ to ensure utility is increasing in consumption.

In the optimal contract, shareholders induce their manager to bear risk on only that part of the return whose probability distribution is affected by his actions. Since managers are risk averse (an assumption we test empirically), his certainty equivalent for a risk-bearing security is less than the expected value of the security, so shareholders would diversify among themselves every firm security whose returns are independent of the manager's activities, rather than use it to pay the manager. We define the abnormal returns of the firm as the residual component of returns that cannot be priced by aggregate factors the manager does not control.

In an optimal contract, compensation to the manager might depend on this residual in order to provide him with appropriate incentives, but it should not depend on changes in stochastic factors that originate outside the firm, which in any event can be neutralized by adjustments within his wealth portfolio through the other stocks and bonds he holds.

More specifically, let w_t denote the overall compensation received by the manager at the end of period t as compensation for work done during the period and v_t the value of the firm at that point in time. Then the gross abnormal returns attributable to the manager's actions is the residual

$$x_t \equiv \frac{v_t + w_t - v_{t-1}}{v_{t-1}} - \pi_t - z_t\gamma,$$

where π_t is the difference between the return on the market portfolio in period t and the return on the firm's stock, and $z_t\gamma$ is a linear combination of some risk factors, denoted z_t, that lead to systematic deviations between the expected return on the firm's shares and the market portfolio. This study assumes that x_t is a random variable that depends on the manager's activity choice in the previous period but, conditional on (l_{1t}, l_{2t}), is independently and identically distributed across both firms and periods. Given $l_{jt} = 1$, for $j \in \{1,2\}$ we denote the probability density function of x_t by $f_j(x_t)$.

The measures of moral hazard described in the previous section can be derived as functions of the parameters defining this framework. The expected loss per period to the firm from the manager pursuing his own interests rather than value maximization is

$$\tau_1 = -v\int x f_1(x)dx,$$

where v is the value of the firm in the previous period. The compensating differential to the manager from pursuing his own interests within the firm compared with working diligently is derived directly from the manager's utility function:

$$\tau_2 = \rho^{-1}\log\left(\frac{\alpha_2}{\alpha_1}\right).$$

In contrast to the other two measures, the welfare cost of moral hazard depends on the optimal contract. It is the expected value of managerial compensation, less its certainty equivalent:

$$\tau_3 = \int w(x) f_2(x)dx - \rho^{-1}\log\left(\frac{\alpha_2}{\alpha_0}\right).$$

The value of being able to offer a contract that creates the manager's incentive to work, as opposed to paying him a fixed wage, is thus

$$\tau_1 - \tau_2 - \tau_3 = \rho^{-1}\log\left(\frac{\alpha_1}{\alpha_0}\right) - v\int x f_1(x)dx - \int w(x) f_2(x)dx.$$

Within this model there are five parameters that might account for differences in executive compensation, that is, apart from the firm's abnormal returns: (i) the probability distribution of abnormal returns conditional on working, (ii) the probability distribution of abnormal returns conditional on shirking, (iii) the risk-aversion parameter, (iv) the nonpecuniary benefit from shirking versus working, and (v) the nonpecuniary benefit of working versus retiring or accepting employment outside the firm. The first two production parameters, $f_2(x)$ and $f_1(x)$, determine τ_1; three of the taste parameters, ρ and α_2/α_1, are used to define τ_2; and as our brief discussion of the optimal contract shows below, all the parameters affect τ_3. Our empirical analysis allows each parameter to differ across firm type and executive position. We also consider the possibility that the five parameters have changed over time and that they depend on underlying factors whose values have changed. In this way we seek to discover why managerial compensation has increased and become more diffuse over the past 60 years.

3.2.2 Estimation. All three measures of moral hazard require us to compute a counterfactual. In the case of τ_1, we must impute the firm's value before compensation is paid if the manager shirks. The manager's utility from shirking is required for τ_2, and in the case of τ_3, what the firm would have paid if there were no moral hazard problem. To identify the parameters of the model, we make the behavioral assumption that shareholders contract with the manager to minimize his expected compensation subject to two weak inequality constraints that induce the manager (i) not to quit the firm (participation) and (ii) to pursue the shareholders' interests rather than his own (incentive compatibility).

The two constraints are satisfied by the optimal contract with strict equality. In our framework, the participation constraint is

$$\alpha_2^{1/(1-b_t)} = E\left[\exp\left(\frac{-\rho w_t}{b_{t+1}}\right)\right],$$

where b_t is the price of a bond in period t that pays a unit of consumption per period forever. The incentive-compatibility constraint is

$$E\left\{\exp\left(\frac{-\rho w_t}{b_{t+1}}\right)\left[g(x_t)-\left(\frac{\alpha_2}{\alpha_1}\right)^{1/(b_t-1)}\right]\right\}=0,$$

where

$$g(x_t) \equiv \frac{f_1(x_t)}{f_2(x_t)}$$

is the ratio of the two probability density functions for shirking and working, respectively. Notice the range of $g(x_t)$ is nonnegative and that its expectation under $f_2(x_t)$ is 1. We interpret $g(x_t)$ as the signal shareholders receive about the manager's effort choice. If the realized value of the signal is zero, they conclude that the manger must have worked diligently, but the greater the realized value of the signal, the less confident they are.

The optimal cost-minimizing contract that implements diligent behavior in this setting can be written as

$$w_t = \frac{b_{t+1}}{\rho(b_t - 1)} \ln(\alpha_2) + \frac{b_{t+1}}{\rho} \ln\left[1 + \eta_t \left(\frac{\alpha_2}{\alpha_1}\right)^{1/(b_t - 1)} - \eta_t g(x_t)\right],$$

where η_t is the unique strictly positive solution to the equation

$$\int \left[\eta(\alpha_2/\alpha_1)^{1/(b_t-1)} - \eta g(x_t) + 1\right]^{-1} f_2(x)dx = 1.$$

Optimal compensation is the sum of two pieces. The second expression determines how compensation varies with abnormal returns through the slope of the signal function, $g(x_t)$. If moral hazard was not a factor because managerial effort could be monitored, then a manager would be paid the flat rate $w_2 = \frac{b_{t+1}}{\rho(b_t - 1)} \ln(\alpha_2)$. The expected value of the other expression is τ_3, the shadow value of moral hazard. Tracing out the contract as a function of abnormal returns, x_t, we recover the signal function, $g(x_t)$, up to a normalization. By definition $f_1(x_t) = g(x_t)f_2(x_t)$, and the probability density function for abnormal returns is identified from data on abnormal returns. Therefore we can estimate $f_1(x_t)$, the density abnormal returns in the absence of appropriate incentives, from a nonlinear regression of w_t on x_t.

To accommodate other factors that might affect compensation but are not included in our model of moral hazard, we assume that our observation on compensation, denoted $\tilde{w}_t$, is the sum of true compensation, denoted w_t, plus an independently distributed error ε_t, assumed orthogonal to the other variables of interest:

(5) $$\tilde{w}_t = w_t + \varepsilon_t.$$

These four equations form the basis for the estimation.

Gayle and Miller (2015) provide regularity conditions for identifying and estimating, from cross-sectional or time-series data on (w_t, x_t, r_t, p_t), the production functions $f_1(x)$ and $f_2(x)$ along with taste parameters $(\rho, \alpha_2, \alpha_1)$. In this analysis, we parameterize $f_1(x)$ and $f_2(x)$, the distributions of abnormal returns under shirking and working, respectively, as truncated normal with support bounded below by ψ, setting

(6) $$f_j(x) = \left[\Phi\left(\frac{\mu_j - \psi}{\sigma}\right)\sigma\sqrt{2\pi}\right]^{-1} \exp\left[\frac{-(x-\mu_j)^2}{2\sigma^2}\right],$$

where $j \in \{1,2\}$ denotes shirking and working, respectively, Φ is the standard normal distribution function, and (μ_j, σ^2) denotes the mean and variance of the parent normal distribution.

As indicated in the previous section, we cannot reject the null hypothesis of restricting the mean of abnormal returns conditional on working to zero conditional on the data. We impose this restriction in the estimation of the parameter μ_2, which implies that μ_2 is determined as an implicit function of the parameters of the truncated normal distribution under work. Denoting by ϕ the standard normal probability density function, the implicit function for μ_2 is given by

(7) $$0 = E(x_t \mid l_{2t} = 1) = \mu_2 + \frac{\sigma\phi[(\psi - \mu_2)/\sigma]}{1 - \Phi[(\psi - \mu_2)/\sigma]}.$$

This leaves the following to be estimated: the bankruptcy return, ψ; the mean of the parent normal distribution under shirking, μ_1; the common variance of the parent normal, σ; the risk aversion parameter, ρ; the ratio of nonpecuniary benefits from working to shirking, α_2/α_1; and the ratio of nonpecuniary benefits from working to quitting, α_2/α_0.

The parameters of the distribution of returns are estimated separately for each sector. For each sector, the production parameters μ_1 and σ^2 are specified as functions of the number of employees in the firm, the firm's assets-to-equity ratio, and an aggregate economic condition—annual gross domestic product. Denoting the controls for observed heterogeneity by z_{1t}, we assume

$$\mu_1 = u_1' z_{1t}$$

and

$$\sigma^2 = \exp(s' z_{1t}).$$

The taste parameters α_2/α_1 and α_2 are specified as linear mappings of executive rank, firm sector, the number of employees in the firm, and the total assets of the firm. Denoting this vector of controls by z_{2t}, we assume

(8) $$\alpha_2/\alpha_1 = a_1' z_{2t}$$

and

(9) $$\alpha_2 = a_2' z_{2t}.$$

The parameter estimates and their asymptotic standard are obtained in three steps. In the first step, maximum-likelihood estimates of the parameter vector determining the distribution of abnormal returns, (ψ, s) are obtained using data on abnormal returns over time and across companies. In the second step, we used data on the abnormal returns and managerial compensation to form a generalized methods-of-moments estimator from the participation constraint, the incentive-compatibility constraint, and the managerial compensation schedule and thus the remaining parameter (ρ, u_1, a_1, a_2). The third step corrects the estimated standard errors in the second step to account for the pre-estimation in the first step (see Gayle and Miller, 2009a, for more details).

3.2.3 Results from the Estimated Model. Table 4 presents the estimated average loss over all firms (i.e., before compensation) from inducing the manager to shirk, both per year and as a net present-value calculation, by sector and for the two samples. The implied average losses have increased more than tenfold in the aerospace and electronics sectors and by a factor of about five in the chemicals sector. In aerospace and electronics, the mean return to

Table 4

Gross Losses to Firms from Shirking in Millions of US$ (2000)

Industry	Parameter τ_1 Old	New
Aerospace		
Per year	13.751	180.212
	(29.522)	(261.294)
Present value	81.065	1,261.484
	(177.132)	(1,829.058)
Chemicals		
Per year	33.392	160.038
	(73.537)	(240.970)
Present value	200.352	1,120.266
	(441.222)	(1,686.79)
Electronics		
Per year	16.650	230.566
	(49.182)	(600.607)
Present value	99.907	1,613.962
	(894.492)	(4,204.249)

NOTE: Standard deviations are in parentheses.

firms from the manager shirking has fallen and the size of the firms has increased. Both factors contribute to the larger expected losses. In the chemicals sector, the mean return from shirking, while negative, has increased and partly offsets the greater loss due to the fact that chemical firms are larger. Comparing the present value of the losses as a ratio of the total assets and the equity value of the firm, we see two measures of how much claimants on the firm, and in the latter case shareholders, would lose from not providing an incentive to managers. Controlling for sector, as a ratio of total assets, the implied losses are of the same order of magnitude in the two datasets, roughly one-ninth in aerospace, just under one-half in chemicals, and about two-thirds in electronics. As a fraction of assets, the losses that would be incurred by not providing an incentive to managers appear relatively stable in these three sectors. Since firms are more leveraged than before, the loss has increased as a fraction of equity value. This is most noticeable in two of the sectors (electronics and chemicals), where the average estimated present value of losses exceeds the average equity value in the new data but not the in old.

The dominant role of firm size in explaining the large increase in the cost of ignoring moral hazard is evident from expressing τ_1 as the negative of the product of firm size v and the expected value of the signal $g(x)$ when the manager works diligently. Differencing the estimates obtained for the two regimes, we obtain the decomposition

$$(10) \qquad -\Delta\tau_1 = (\Delta v)\int xg(x)f_2(x)dx + v\int x\left[\Delta g(x)\right]f_2(x)dx + v\int xg(x)\left[\Delta f_2(x)\right]dx.$$

Table 5

Nonpecuniary Benefits of Shirking in Thousands of US$ (2000)

	Parameter τ_2	
Industry	Old	New
Aerospace		
CEO	2,380	4,000
	(43)	(92)
Non-CEO	1,500	3,400
	(72)	(78)
Chemicals		
CEO	920	3,800
	(274)	(209)
Non-CEO	812	600
	(321)	(451)
Electronics		
CEO	747	3,048
	(432)	(387)
Non-CEO	436	2,070
	(515)	(366)

NOTE: Standard deviations are in parentheses.

The first of the three expressions on the right-hand side, the change in the cost of moral hazard due to the increasing size of firms, is unambiguously positive. The second expression arises because of changes in $g(x)$. In two of the sectors, the signal has weakened, reducing the gap between $f_1(x)$ and $f_2(x)$ and thus mitigating the losses that would be incurred from encouraging the manager to pursue his own goals instead of expected-value maximization. The third expression captures the effects of the change in the distribution of abnormal returns. Noting that $f_2(x)$ has undergone a mean-preserving spread in two sectors and that $g(x)$ is a convex decreasing function, it follows that the third expression is positive for these sectors, thus reducing the loss incurred. In summary, the growth of firms increased the losses from shirking so much that it dominates the other two effects.

The two remaining measures of moral hazard, τ_2 and τ_3, can now be computed from the estimated parameters. The nonpecuniary value of deviating from the incentive-based contract depends only on the preferences of the manager, not the distribution of the abnormal returns. For each observation, we compute a consistent estimator for τ_2:

$$\tau_2 = \left[\rho\left(b_t - 1\right)\right]^{-1} b_{t+1} \ln\left(\alpha_2/\alpha_1\right). \quad (11)$$

Table 5 reports, by sector and executive position, the average of the consistent estimators and consistent estimates of their respective standard deviations. The firm averages for each executive type by sector have increased in five of the six categories, by a factor of more than three for CEOs in two sectors. As a proportion of total compensation averaged over observa-

Table 6

Welfare Cost of Moral Hazard in Thousands of US$ (2000)

	Parameter τ_3	
Industry	**Old**	**New**
Aerospace		
CEO	500 (1,316)	10,350 (15,473)
Non-CEO	330 (1,413)	1,280 (10,501)
Chemicals		
CEO	490 (1,437)	2,973 (5,087)
Non-CEO	299 (206)	301 (1,678)
Electronics		
CEO	278 (1,257)	4,873 (17,285)
Non-CEO	67 (188)	1,206 (11,159)

NOTE: Standard deviations are in parentheses.

tions for each executive type by sector, the compensating differential to managers for pursuing their own interests has fallen in all six categories. A key factor contributing to this measure of importance, τ_2/w, is that changes in the supply and demand for managerial services has roughly doubled the compensation of managers of a firm with any given set of characteristics.

In both samples, the average τ_2 is tiny compared with the expected losses a firm would incur; our model predicts there are enormous gains from having managers act in the interest of shareholders. From the manager's perspective, however, τ_2 is quite substantial, and for a sizeable proportion of the sample population, exceeds actual and even expected compensation. This paradox is resolved by noting that the manager would be harshly penalized if the firm does poorly, which is of course more likely if he shirks. Perhaps the most striking feature of these results is how they compare with estimates of α_1' as defined in equation (8). Table 5 averages the predicted α_2/α_1 from equation (8) over firms within each sector after taking logarithms and scaling by $\rho - 1$. Since ρ has not changed much and the estimated changes in α_1' are for the most part insignificant or negative, we attribute the dramatic differences between the tables to the changing composition of firms within each sector. More specifically, the effects of the average growth in firm assets dominate the decline in employment and are largely responsible for the increased compensating differential to work for shareholders versus pursuing some other agenda.

The last measure of moral hazard, τ_3, is the welfare cost of the moral hazard—the willingness of a firm to pay for a perfect monitor—thus eliminating moral hazard. From the definition of τ_3 and the solution for the optimal contract, it follows that the welfare cost may be expressed as

$$(12) \qquad \tau_3 = b_{t+1}\rho^{-1}\int ln\left[1+\eta_t\left(\alpha_2/\alpha_1\right)^{\frac{1}{(b_t-1)}} - \eta_t g\left(x_t\right)\right] f_2(x)dx.$$

Table 6 presents consistent estimates of the average of τ_3 in the two samples and three sectors, along with the consistent estimates of the standard deviations. The table shows that the increase in managerial compensation presented in Table 3 is mirrored in the increased cost of moral hazard. From the formula above and the formula for η_t, changes in τ_3 are ultimately attributable to changes in α_2/α_1, $f_1(x)$, and $f_2(x)$ only. After adjusting for the general rise in living standards, the estimated model attributes practically all the increase in managerial compensation to moral hazard and hardly any of it to changes in the supply and demand for managers, as reflected in the participation condition and hence α_2/α_0.

To further investigate the sharply increased cost of moral hazard, we first note that changes in b_{t+1}/ρ between the two samples are minimal and decompose $\Delta\tau_3$ into changes stemming from changes in $f_2(x)$ and changes in the integrand. Since $g(x_t)$ is a convex decreasing function,

$$(13) \qquad \ln\left[1+\eta_t\left(\alpha_2/\alpha_1\right)^{1/(b_t-1)} - \eta_t g\left(x_t\right)\right]$$

is a concave increasing function. Noting again that $\Delta f_2(x)$ is a mean-preserving spread in chemicals and engineering but not in aerospace, it therefore follows that

$$(14) \qquad b_{t+1}\rho^{-1}\int \ln\left[1+\eta_t\left(\alpha_2/\alpha_1\right)^{1/(b_t-1)} - \eta_t g\left(x_t\right)\right]\Delta f_2(x)dx$$

is positive in chemicals and engineering but negative in aerospace. Thus, changes in the distribution of abnormal returns cannot explain why the welfare cost of moral hazard increased in aerospace, the sector where the biggest increases occurred. The remaining component to explain $\Delta\tau_3$ is

$$b_{t+1}\rho^{-1}\int \left\{\Delta \ln\left[1+\eta_t\left(\alpha_2/\alpha_1\right)^{1/(b_t-1)} - \eta_t g\left(x_t\right)\right]\right\} f_2(x)dx.$$

The predominant change is due to a sharp increase in α_2/α_1 averaged over firms.[9] This component is the most important factor responsible for the increase in τ_3. To recapitulate, increased firm assets exacerbated the conflict between managers and shareholders by creating new opportunities for managers to act against shareholder interests. These were resolved through the compensation schedule by placing greater weight on penalizing poor firm performance and rewarding superior abnormal firm returns, thus subjecting risk-averse managers to the vagaries of greater insider wealth and causing their expected compensation to rise at a rate much greater than that of national income per capita.

3.3 Summary

The welfare cost of moral hazard is a compensating differential paid to risk-averse managers to hold insider wealth and accept nondiversifiable risk that aligns their incentives to those of the stockholders, who do not price risk from an individual firm's abnormal returns because of their portfolio choices. Tables 5 and 6 show that the welfare cost of the moral hazard associated with employing CEOs has increased by an estimated factor of more than 20 times in the aerospace and electronics sectors and sixfold in the chemicals sector. Subtracting the welfare costs of the moral hazard displayed in Table 6 from the expected compensation paid to top executives reported in Table 1, we obtain, for each of the six categories, the average certainty equivalent wage, which equates the supply and demand for managerial services for a given firm. The overall increase in the 60-year period is 2.3, the same as the increase in national income per capita. Therefore, our results attribute all the difference between the rate of increase in managerial compensation and the rate of increase in national income per capita to the rising welfare cost of moral hazard.

The cost of moral hazard depends on the preferences of managers, what shareholders observe about their behavior, the distribution of abnormal returns accruing to firms, and the characteristics of the firms managed. We do not attribute the steep increase in the welfare cost to changing tastes. Gayle and Miller (2009a) show that, if anything, the conflict between a firm with a given set of characteristics and its executives has declined. As documented in Table 1, there have been changes in the probability distributions of abnormal returns, but not all in the same direction. Gayle and Miller (2009a) show that managerial preferences for risk have remained stable in an economic sense and that the compensating differential of deviating from the goal of maximizing the expected value of the firm with a given set of characteristics has not increased. Nevertheless, Table 4 shows that if managers were paid a flat wage to prevent skimming, and if our model of moral hazard is correctly specified, then conflict between managerial and shareholder objectives would remain unresolved and the ensuing losses incurred by firms would be catastrophic and would have grown substantially over the past 60 years.

4 WHY ARE ACCOUNTING RETURNS RELATED TO EXECUTIVE COMPENSATION?

As an empirical matter, managerial compensation varies significantly with abnormal financial returns.[10] The theory of pure moral hazard postulates that risk-averse managers should receive compensation that fluctuates with signals (most notably abnormal returns) that risk-neutral shareholders observe based on decisions their managers make. That is, when managers' nonpecuniary goals differ from maximizing shareholder wealth and the actions and decision of management are not monitored, managers need to be incentivized to align their goals with those of the shareholders. Although the dominant paradigm, this explanation for executive compensation has been challenged on several fronts. First, several empirical studies find that trading by corporate insiders appears profitable,[11] but in models of pure moral hazard, managers do not have private information about the firm's future prospects. Second, as we show below, managerial compensation depends on not only the financial returns of the firm, but also its accounting returns. In models of pure moral hazard, shareholders might use signals other than financial returns to determine optimal compensation, but the reporting of accounting income is subject to considerable discretion by the manager. In qualitative terms, these three anomalies for the pure moral hazard model can be rationalized by the hybrid model. In this section, we see how fast the hybrid principal-agent model can go in rationalizing these anomalies.

4.1 Accounting Return and Executive Compensation

For the study in this section, we use binary variables based on firm size and capital structure (the debt-to-equity ratio) to categorize firms into four types. Firm size is measured by the total assets on a firm's balance sheet (AT; variable names in parentheses hereafter) at the end of period *t*. The capital structure is reflected by the debt-to-equity ratio. The numerator of the ratio is the total liabilities (LT), and the denominator is the total common equity (CEQ). The book values of assets, liabilities, and equity are deflated to the base year 2006. We classify each firm by (i) whether its total assets averaged over years were less than or greater than the median of the average total assets of firms in the same sector and (ii) whether its average debt-to-equity ratio was less than or greater than the median of the average debt-to-equity ratio of firms in that sector. Therefore, firm type is measured by the coordinate pair (A, C), with each corresponding to whether that element is above (L) or below (S) the medians of the industry. For example, (S,L) denotes lower total assets and a higher debt-to-equity ratio than the median debt-to-equity ratio for firms in that sector. By doing so, one firm stays in the same firm category and sector for the entire sample period.

In the model presented earlier, after accepting the contractual arrangement, CEOs collect and convey their private information on the firm's prospects. We construct an empirical measure of the report by equity return evaluated at book value, which is consistent with the concept of comprehensive income in accounting practice. Accounting numbers feature the private state in the theoretical framework because many of the estimations are used to generate accounting numbers. For example, accrual (defined as the difference between realized cash flow and reported earnings) is one of the typical accounting features used as an information system. The smoothing-over periods require information about the state of the firm, which may be unknown to shareholders, especially in modern firms where the control rights and ownership are separated. Based on estimation, the accounting numbers can convey private information to shareholders about prospects.

Specifically, we define the binary private state (good or bad), denoted as S_{nt}, conditional on the accounting return to equity that is measured by book value. The accounting return is denoted as r_{nt} and calculated as

$$r_{nt} = \frac{Asset_{nt} - Debt_{nt} + Dividend_{nt}}{Asset_{n,t-1} - Debt_{n,t-1}}, \quad (15)$$

where for firm *n* in year *t*, *Asset* is the total assets (AT) at the end of year *t*, *Debt* is the total liability (LT) minus minority interest (MIB), *Dividend* is the dividend to common stock (DVC) plus the dividend to preferred stock (DVP). All variables are deflated to the base year 2006 before calculating the accounting return.

4.1.1 The Relationship Between Accounting Returns and Executive Compensation. In Table 7, we summarize and contrast total compensation between the bad state and the good state for each type of firm in each sector. If shareholders did not exploit a proper compensation contract designed to solicit CEOs' private information, which we assume is embedded in accounting returns , it is unlikely that the distribution of total compensation would present a

Table 7

Total Compensation by Accounting Returns

	Total Compensation		
	Overall	Bad state	Good state
Primary			
(S,S)	2,576 (12,787)	737 (9,331)	4,716 (15,625)
(S,L)	1,965 (8,759)	428 (7,113)	3,995 (10,206)
(L,S)	5,462 (12,957)	4,104 (10,997)	7,172 (14,903)
(L,L)	5,320 (12,734)	3,981 (11,429)	6,957 (13,997)
Consumer goods			
(S,S)	2,479 (20,991)	–1,285 (15,058)	7,351 (25,998)
(S,L)	1,858 (13,639)	–477 (10,663)	4,501 (15,974)
(L,S)	6,896 (31,409)	1,693 (23,671)	12,711 (37,427)
(L,L)	8,234 (27,373)	4,152 (22,382)	13,744 (32,131)
Services			
(S,S)	3,580 (20,116)	480 (14,521)	7,159 (24,591)
(S,L)	3,627 (16,985)	2,000 (13,124)	5,460 (20,342)
(L,S)	11,070 (37,636)	5,386 (30,669)	18,285 (43,933)
(L,L)	10,003 (26,144)	6,733 (22,103)	14,772 (30,497)

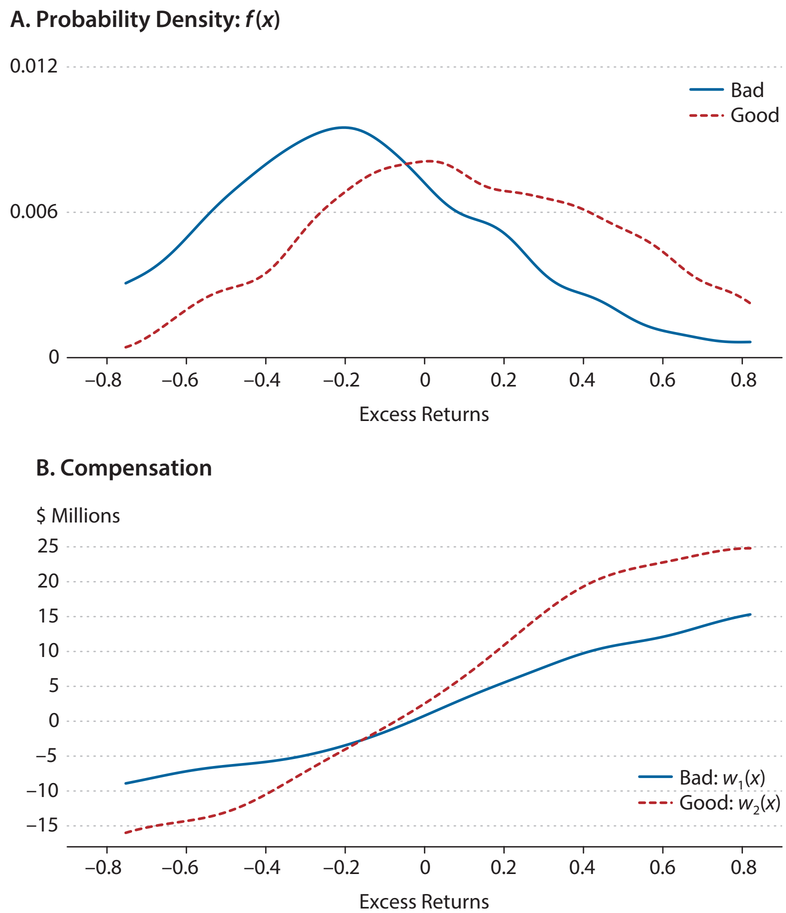
NOTE: Standard deviations are in parentheses. Firm type is measured by the coordinate-pair (A, C), where A is assets and C is the debt-to-equity ratio, with each corresponding to whether that element is above (L) or below (S) its industry median. Accounting returns are classified as "good (bad)" if they are greater (less) than the industry average. Assets (compensation) is measured in millions (thousands) of US$.

systematic distinction between the two states. However, Table 7 presents the opposite. It reports the mean and standard deviations of total compensation conditional on firm size, capital structure, and industrial sector. The universal pattern is that total compensation always shows a lower level and smaller standard deviation in the bad state than in the good state. CEOs are paid more in the good state, but their pay is less concentrated.

In addition, total compensation is, not surprisingly, smaller in small firms than in large firms, regardless of the state. The relationship between capital structure and total compensation

Figure 2

Empirical Excess-Return Densities and the Total Compensation Schedule



NOTE: The panels present the non-parametrically estimated density of excess returns and the optimal compensation of firms with large size and high leverage in the primary sector. The compensation of both periods is anchored at bond prices equal to 16.5 (b_t) and 16.4 (b_{t+1}).

behaves differently between the two states. In the bad state, firms with a high debt-to-equity ratio are more likely to have higher compensation, except the firms in the primary sector. However, in the good state, this happens only in large firms in the consumer goods sector.

After some simple calculations, the smallest difference in total compensation between the two states is about $3 million, for large firms in the primary sector with a high debt-to-equity ratio, and the largest difference is nearly $12 million, for large firms in the services sector with a low debt-to-equity ratio. The between-state difference tends to be larger in larger firms than in smaller firms in the consumer goods sector and the services sector, but smaller in the primary sector. The difference is always smaller in firms with a high debt-to-equity ratio, regardless of size or sector.

Figure 2 graphically compares the distribution of total compensation and excess returns between the two states, taking large firms in the primary sector as an example. Panel A of Figure 2 presents the kernel density of excess returns for each of the two states. Excess returns are lower on average in the bad state than in the good state, indicating the lower compensation in the bad state, as Table 7 reports, which may reflect punishment of inferior performance too. Thus we need a more structured research design to separate the effect of productive performance and that of information rent on the level of total compensation.

In Panel B of Figure 2, the non-parametrically estimated compensation schedule is compared between the two states. The curve of the optimal contract in the good state is steeper than that in the bad state, indicating that in the good state, compensation is more sensitive to performance. The empirically estimated compensation schedule increases with excess returns and flattens at very high rates of excess returns. These features illustrate the agency problem and suggest that hidden information, not just hidden actions, may be a part of the agency problem.

4.2 A Hybrid Principal-Agent Model

To this end we now lay out a dynamic principal-agent model of optimal contracting between risk-neutral shareholders and a risk-averse CEO, based on Gayle and Miller (2015), in which the CEO has hidden information and also takes actions that cannot be directly observed by shareholders. An important feature of this model is that it treats accounting information as a non-verifiable statement by the CEO, whose credibility depends on the incentives that determine his payoff as a function of what he reports.

At the beginning of period t, the CEO is paid compensation, denoted by w_t, for his work the previous period, denominated in terms of period-t consumption units. He makes his consumption choice, a positive real number denoted by c_t, and the board proposes a new contract. The board announces how CEO compensation will be determined as a function of what he will disclose about the firm's prospects, denoted by $r_t \in \{1,2\}$[12], and its subsequent performance, measured by excess returns x_{t+1}, revealed at the beginning of the next period. We denote this mapping by $w_{rt}(x)$, with the subscript t designating that the optimal compensation schedule may depend on current economic conditions, such as bond prices. Then the CEO chooses whether to be engaged by the firm. Denote this decision by the indicator $l_{t0} \in \{0,1\}$, where $l_{t0} = 1$ if the CEO chooses to be engaged outside the firm and $l_{t0} = 0$ if he chooses to be engaged inside the firm.

If the CEO accepts employment with the firm, $l_{t0} = 0$, the prospects of the firm are now fully revealed to the CEO but partially hidden from the shareholders. There are two states, $s_t \in \{1,2\}$, and we denote the probability that state s_t occurs by $\varphi_{st} \in (0,1)$. We assume that CEOs privately observe the true state, $s_t \in \{1,2\}$, in period t, gaining information that affects the distribution of the firm's next-period excess returns, and reports r_t to the board. If the CEO discloses the second state, meaning $r_t = 2$, then the board can independently confirm or refute it; thus, if $s_t = 1$, he reports $r_t = 1$. If $s_t = 2$, the CEO then truthfully declares or lies about the firm's prospects by announcing $r_t \in \{1,2\}$, effectively selecting one of two schedules, $w_{1t}(x)$ or $w_{2t}(x)$.

The CEO then makes his unobserved labor-effort choice, denoted by $l_{stj} \in \{0,1\}$ for $j \in \{1,2\}$ for period t, which may depend on his private information about the state of the firm. There are two possibilities: to work and diligently pursue the shareholders' objectives of value maximization, thus setting $l_{st2} = 1$, or to shirk and accept employment with the firm but follow the objectives he would pursue if he were paid a fixed wage by setting $l_{st1} = 1$. Let $l_{st} \equiv (l_{t0}, l_{st1}, l_{st2})$. Since leaving the firm, working, and shirking are mutually exclusive activities, $l_{t0} + l_{st1} + l_{st2} = 1$.

At the beginning of period $t + 1$, excess returns for the firm, x_{t+1}, are drawn from a probability distribution that depends on the true state, s_t, and the CEO's action, l_{st}, in period t. We denote the probability density function for excess returns when the CEO works diligently and the state is s by $f_{st}(x)$. Similarly, let $f_{st}(x)g_{st}(x)$ denote the probability density function for excess returns in period t when the CEO shirks. Thus, for both states $s_t \in \{1,2\}$:

(16) $$\int x f_{st}(x) g_{st}(x) dx \equiv E_{st}\left[x g_{st}(x)\right] < E_{st}\left[x\right] \equiv \int x f_{st}(x) \mathrm{d}x,$$

with the inequality reflecting the shareholders' preference for diligent work over shirking. Since $f_{st}(x)g_{st}(x)$ is a density, $g_{st}(x)$ is positive and integrating $f_{st}(x)g_{st}(x)$ with respect to x demonstrates $E_{st}[g_{st}(x)] = 1$. We assume the likelihood of shirking declines to zero as excess returns increase without bound:

(17) $$\lim_{x\to\infty}\left[g_{st}(x)\right] = 0$$

for each $s \in \{1,2\}$. We assume the weighted-likelihood ratio of the second state occurring relative to the first given any observed value of excess returns, $x \in R$, converges to an upper finite limit as x increases such that

(18) $$\lim_{x\to\infty}\left[\varphi_{2t} f_{2t}(x)/\varphi_{1t} f_{1t}(x)\right] \equiv \lim_{x\to\infty}\left[h_t(x)\right] = \sup_{x\in R}\left[h_t(x)\right] \equiv \overline{h}_t < \infty.$$

The CEO's wealth is endogenously determined by his consumption and compensation. We assume a complete set of markets for all publicly disclosed events effectively attributes all deviations from the law of one price to the particular market imperfections under consideration. Let b_t denote the price of a bond that pays a unit of consumption each period from period t onward, relative to the price of a unit of consumption in period t; to simplify the exposition, we assume b_{t+1} is known at period t. Preferences over consumption and work are parameterized by a utility function exhibiting absolute risk aversion that is additively separable over periods and multiplicatively separable with respect to consumption and work activity within periods. In the model we estimate, lifetime utility can be expressed as

(19) $$-\sum_{t=0}^{\infty}\sum_{j=0}^{2}\beta^t \alpha_{jt} l_{tj} \exp\left(-\gamma_t c_t\right),$$

where β is the constant subjective discount factor, γ_t is the constant absolute level of risk aversion, and α_{jt} is a utility parameter that measures the distaste from working at level $j \in \{0,1,2\}$. We assume working is more distasteful than shirking, meaning $\alpha_{2t} > \alpha_{1t}$, and normalize $\alpha_{0t} = 1$.

In this framework, there are no gains from a long-term arrangement between shareholders and the CEO: The optimal long-term contract between shareholders and the CEO decentralizes to a sequence of short-term one-period contracts. Therefore, the model can be solved in two steps. First we solve for the optimal consumption and savings plan for a CEO about to retire. It can be proved in this model that given the CEO's reporting about the state of the firm and the true state of the firm, his employment and effort choices depend on his preference parameters $(\alpha_{1t}, \alpha_{2t}, \gamma_t)$, the distribution of excess returns when he shirks $f_{st}(x)g_{st}(x)$ and when he works $f_{st}(x)$, and aggregate economic conditions as reflected in bond prices (b_t, b_{t+1}). However, the employment and effort choices do not depend on his current (outside) wealth. Let $r_t(s)$ denote the CEO's disclosure rule about the state when the true state is $s_t \in \{1,2\}$.

If the CEO, offered a contract of $w_{rt}(x)$ for announcing r, retires in period t or $t+1$ by setting $(1 - l_{t0})(1 - l_{t+1,0}) = 0$, upon observing the state s and reporting $r_t(s)$, he optimally chooses $l_{st} \equiv (l_{t0}, l_{st1}, l_{st2})$ to minimize

$$\sum_{s=1}^{2}\varphi_{st}\left\{l_{t0}+\left(\alpha_{1t}l_{st1}+\alpha_{2t}l_{st2}\right)^{\frac{1}{(b_t-1)}}E_{st}\left[\exp\left(-\frac{\gamma_t w_{r_t(s)t}(x)}{b_{t+1}}\right)\left[g_{st}(x)l_{st1}+l_{st2}\right]\right]\right\}. \tag{20}$$

The optimal short-term contract for shareholders is found by minimizing the expected compensation subject to four constraints the CEO prefers: (i) working for a period rather than leaving the firm, (ii) being truthful rather than lying, (iii) working instead of shirking, or (iv) being truthful and working diligently rather than lying and shirking. Suppressing for expositional convenience the bond price b_{t+1} and recalling our assumption that b_{t+1} is known at period t, we now let $v_{st}(x)$ measure how (the negative of) utility is scaled up by $w_{st}(x)$:

$$v_{st}(x)\equiv\exp\left(-\frac{\gamma_t w_{st}(x)}{b_{t+1}}\right). \tag{21}$$

First, to induce an honest, diligent CEO to participate, his expected utility from employment must exceed the utility he would obtain from retirement. Setting $(l_{t2}, r_t) = (1, s_t)$ in (20) and substituting in $v_{st}(x)$, the participation constraint is thus

$$\sum_{s=1}^{2}\int\varphi_{st}v_{st}(x)f_{st}(x)\mathrm{d}x\le\alpha_{2t}^{-\frac{1}{(b_t-1)}}. \tag{22}$$

Second, given his decision to stay with the firm one more period and to truthfully reveal the state, the incentive-compatibility constraint induces the CEO to prefer working to shirking for $s_t \in \{1,2\}$. Substituting the definition of $v_{st}(x)$ into (20) and comparing the expected utility obtained from setting $l_{t1} = 1$ with the expected utility obtained from setting $l_{t2} = 1$ for any given state, we obtain the incentive-compatibility constraint for work:

$$0\le\int\left(g_{st}(x)-\left(\frac{\alpha_{2t}}{\alpha_{1t}}\right)^{\frac{1}{(b_t-1)}}\right)v_{st}(x)f_{st}(x)\mathrm{d}x. \tag{23}$$

Information hidden from shareholders further restricts the set of contracts that can be implemented. Comparing the expected value from lying about the second state and working diligently with the expected utility from reporting honestly in the second state and working diligently, we obtain the truth-telling constraint:

$$0 \le \int \left[v_{1t}(x) - v_{2t}(x) \right] f_{2t}(x) \mathrm{d}x. \tag{24}$$

An optimal contract also induces the CEO not to understate and shirk in the second state, behavior we describe as sincere. Comparing the CEO's expected utility from lying and shirking with the utility from reporting honestly and working diligently, the sincerity condition reduces to

$$0 \le \int \left(\left(\frac{\alpha_{1t}}{\alpha_{2t}} \right)^{\frac{1}{(b_t - 1)}} v_{1t}(x) g_{2t}(x) - v_{2t}(x) \right) f_{2t}(x) \mathrm{d}x, \tag{25}$$

where $(\alpha_{1t}/\alpha_{2t})^{1/(b_{t-1})} v_{1t}(x)$ is proportional to the utility obtained from shirking and announcing the first state and $f_{2t}(x)g_{2t}(x)$ is the probability density function associated with shirking when the second state occurs. Minimizing expected compensation amounts to choosing $v_{st}(x)$ that maximizes

$$\sum_{s=1}^{2} \int \varphi_{st} \ln \left[v_{st}(x) \right] f_{st}(x) \mathrm{d}x. \tag{26}$$

Noting $\ln v_{st}$ is concave, increasing in v_{st}, the expectation operator preserves concavity, so the objective function is concave in $v_{st}(x)$ for each x. Each constraint is a convex set and its intersection is too. Therefore, we can appeal to the Kuhn-Tucker theorem, which guarantees there is a unique positive solution to the equation system formed from the first-order conditions augmented by the complementary-slackness conditions.

4.3 Comparing the Pure and Hybrid Model Contracts

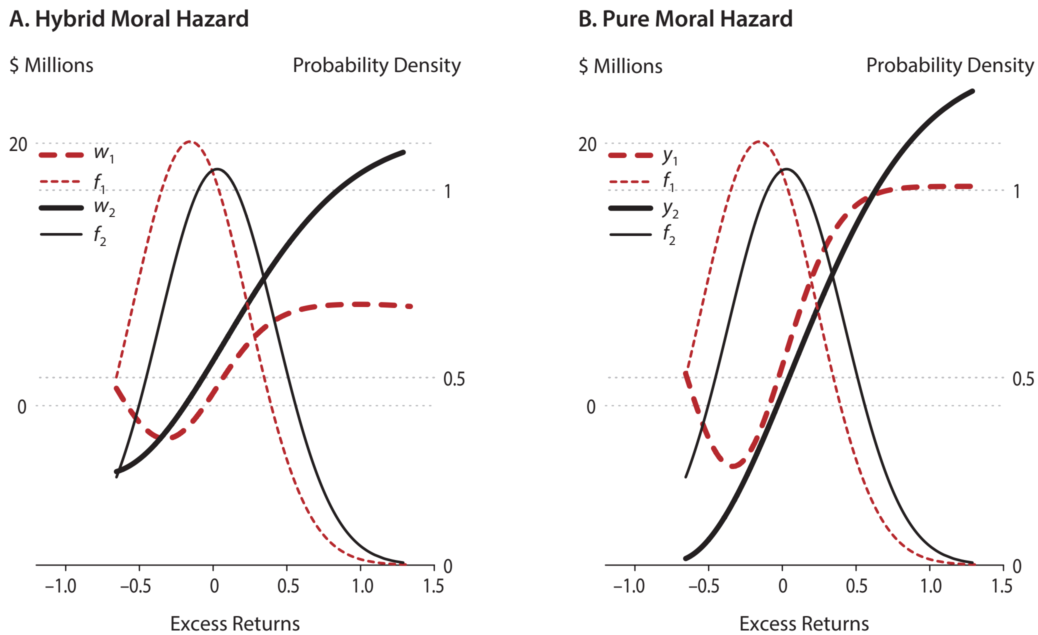
The optimal contract for a parameterization of the hybrid model is plotted in Panel A of Figure 3. This parameterization follows Margiotta and Miller (2000) in assuming that excess returns are drawn from a truncated distribution, with a common lower bound for all states and independent of the effort level.[13] For comparison purposes, we also plot in Panel B optimal compensation for the analogous two-state pure moral hazard model (where there are hidden actions but the state is known), by $y_{st}(x)$.

To derive $y_{st}(x)$, the optimal compensation in the analogous two-state pure moral hazard model, we drop the truth-telling and sincerity constraints, replace the single participation constraint with one for each state, retain both incentive-compatibility constraints, minimize the modified objective function, use the participation constraints to substitute out their associated Kuhn-Tucker multiplier, and rearrange the first-order conditions to obtain

$$y_{st}(x) = \gamma^{-1} \frac{b_{t+1}}{b_t - 1} \ln \alpha_2 + \gamma^{-1} b_{t+1} \ln \left[1 + \eta_{st}^{p} \left(\alpha_2 / \alpha_1 \right)^{\frac{1}{b_t - 1}} - \eta_{st}^{p} g_{st}(x) \right], \tag{27}$$

Figure 3

Optimal Compensation Schedules



NOTE: The plots use the return and optimal compensation of firms with small size and low leverage in the primary sector in the 1993-2001 period. The risk aversion parameter γ is equal to 0.08. The effort cost coefficient of shirking (α_1) equals 0.96, and the effort cost coefficient of working (α_2) equals 1.20. Bond prices are 16.5 (b_t) and 16.4 (b_{t+1}). The excess return is approximated by one-side truncated normal distribution $TN(a,\mu,\sigma)$ with truncated points on the left (a), mean (μ), and standard deviation (σ) as follows: working in the bad state: $TN(-0.66,-0.16,0.39)$; working in the good state: $TN(-0.66,0.03,0.39)$; shirking in bad state: $TN(-0.66,-0.25,0.27)$; and shirking in good state: $TN(-0.66,-0.11,0.36)$. The probability of the bad state is 0.54, and the probability of the good state is 0.46.

where η_{st}^{p} is the unique positive solution to

$$\int_{\bar{x}}^{\infty} \frac{g_{st}(x)-\left[\dfrac{\alpha_2}{\alpha_1}\right]^{\frac{1}{b_t-1}}}{1+\eta_{st}^{p}\left[\dfrac{\alpha_2}{\alpha_1}\right]^{\frac{1}{b_t-1}}-\eta_{st}^{p} g_{st}(x)} f_s(x)dx=0. \tag{28}$$

We approximate the integral (28), accounting for the singularity problem that occurs when the denominator of the integrand is either zero or ∞. First, we perform a grid search to detect the singularity points in the range of x. These singularity points divide the entire range of x into a number of subintervals. The integral (28) is approximated for a given η_{st}^{p} by first being approximated on each subinterval and then summed over the entire range. Then we numerically solve for the optimal value of η_{st}^{p} that satisfies (28) based on this approximated integral.

Figure 3 illustrates four important features. Since compensation in both models is a function of the likelihood ratio between the densities of the excess returns for working and shirking, not excess returns alone, the wage contract is not necessarily monotonically increasing in excess returns. For example, in the bad states of both models of the illustrated parameterization, pay optimally declines with marginal increments to excess returns when they are less than –0.5. The same explanation applies to compensation leveling out at high levels of excess returns; the likelihood ratio converges to a constant, zero, under the assumption of a truncated normal distribution.

The other two noteworthy features relate to differences between the pure and hybrid contracts. The slope of the hybrid compensation schedule is everywhere greater in the good state than in the bad state, whereas in the pure moral hazard model the slope in the bad state is greater than in the good state over the intermediate range, where much of the probability mass of both excess distributions lie. Thus the point where the schedules cross is higher in the pure moral hazard model than in the hybrid model. The figure also illustrates two analytical results: In the hybrid model, expected utility of the agent is greater in the good state than the bad state; but in the pure moral hazard model, expected utilities are equalized across states. Intuitively, the argument is that in the hybrid model the principal induces the agent to truthfully reveal the good state by promising (i) more expected utility in the good state and (ii) a flatter compensation profile in the bad state.

Finally, because the constraints in the pure moral hazard optimization problem are not a subset of those in the hybrid model, there is no presumption that the expected compensation in the pure moral hazard case is lower than in the hybrid model. In other words, the principal may find it cheaper not to know the private information if he can optimally spread the utility the agent receives across both states, rather than meet the participation constraint in each state. Indeed, our parameterization illustrates an instance where the agency cost in the pure moral hazard model is greater than in its hybrid counterpart. Finally, a comparison of Figure 3, which is produced from the estimated hybrid model (see Gayle and Miller, 2015, and Gayle, Li and, Miller, 2016, for details of the estimation), and Figure 2, nonparametric estimates from the data, reveals a strict similarity between the optimal contract from the hybrid model and the empirical contract observed in the data.

4.4 Summary

If every piece of information a manager knows about his firm is codified and independently verifiable in a court of law, managers can be compelled to reveal all their private information through the firm's accounting records. In this case, a basic pure moral hazard model would apply conditional on verifiable information. Within the current legal system, however, managers exercise considerable discretion about how much information they release describing the state of their own firms. If the penal code for accounting protocol were augmented by incentives embedded in managerial compensation designed to elicit truthful revelation, a hybrid model of moral hazard would apply. Therefore, the fact that non-verifiable information is used in compensating executives does not reject the principal-agent model.

Gayle and Miller (2015) and Gayle, Li, and Miller (2016) use a large panel dataset measuring compensation of chief executive officers, financial and accounting returns, and size and sector background characteristics of publicly traded firms. They investigate whether the hybrid model can reconcile the fact that executives are paid based on accounting returns. In the pure moral hazard models estimated and tested by Gayle and Miller (2015), managers do not have discretion about how they report accounting returns. In the hybrid model, estimated by Gayle and Miller (2015) and Gayle, Li, and Miller (2016), they interpret data on accounting returns as information reported by the CEO that cannot be fully corroborated by shareholders. Thus our empirical study compares and contrasts the role of these alternative information assumptions about accounting returns within competing models of moral hazard.

The data show that expected compensation for the next period increases with current accounting returns and also that the gradient of compensation in financial returns is higher with greater accounting returns. The hybrid model predicts that the expected utility of the agent is higher in the firm's good state than in its bad state. Moreover, to induce truth telling and report higher earnings when the firm's prospects are good, the principal lowers and flattens the schedule when the agent reports the bad state, reducing expected compensation and making realized compensation less dependent on the outcome. In our application, this permits financial and accounting returns data to play bigger roles in explaining compensation. Relatively high estimated values of the risk parameter, which are consistent with previous work on pure moral hazard models that do not exploit the accounting data, reduce the certainty equivalent of compensation in the good state. These features reconcile the hybrid model to the data even when tastes for working and risk attitudes are not allowed to vary with the firm's accounting state.

In contrast to the hybrid model, the pure moral hazard model equalizes expected utility across states. The heterogeneous pure moral hazard model mitigates the effects of curvature differences in compensation schedules across states, by making the managers appear almost risk neutral and simultaneously attributing to nonpecuniary benefits the differences in expected compensation across accounting states. The risk parameter in the heterogeneous pure moral hazard model is considerably lower than previous findings for pure moral hazard models that do not exploit differences in accounting states. The nonpecuniary benefits from working for the firm in the bad accounting state are so high that the estimated certainty equivalent compensation is negative. But unless work preferences or risk attitudes differ across accounting states, the pure moral hazard framework lacks the degrees of freedom necessary to fit the differently shaped compensation schedules.

5 CONCLUSION

In this article we illustrate how using theory-based estimation together with a model-motivated measure of total compensation can help overcome the problem that the exact prediction of the principal-agent model (as well as most models of imperfect information) depends on many objects unobservable to the econometrician. The article concludes that using a model-consistent measure of compensation and theory-based estimation shows that executive compensation broadly conforms to the principal-agent theory; however, each situation and the variables used have to be carefully modeled, identified, and estimated.

Across all the different specifications of the principal-agent models summarized in this article, two robust facts emerge. First, more than 80 percent of total executive pay is from the risk premium paid to resolve the agency problem (see Gayle and Miller, 2009a,b, and Gayle, Golan, and, Miller, 2015). Second, the size of this risk premium is explained mostly by firm size (see Gayle and Miller, 2009a, and Gayle, Golan, and Miller, 2015). A risk premium, rationalized in the principal-agent model by incentive contracts to deter shirking, accounts for approximately 80 percent of the firm-size pay premium. More specifically, the estimated risk premium is \$1.6 million for small firms, \$2.6 million for medium-sized firms, and \$4.9 million for large firms. These findings are consistent with explanations that suggest large firms pay large efficiency wages to prevent shirking. Therefore, in order to understand the reasons for the increase in executive pay and the increased top income shares over time, researchers need to examine the reasons behind the increase in firm size over time. ■

NOTES

1 See Piketty and Saez (2003) for more details on these trends.

2 See Gayle and Miller (2009a) for more details.

3 For example, in the insurance market, the basic prediction of agency theory is that there is a correlation between insurance coverage and risk. If this prediction is correct, then policyholders who are known to themselves (but not to their insurer) to be high risk will tend to choose higher insurance coverage (lower deductibles); thus, coverage and risk are expected to be positively correlated. This coverage-risk correlation has been the major focus of empirical work in this area, where risk is measured by the likelihood of an accident.

4 In the managerial compensation setting, the positive correlation test boils down to testing the prediction of a positive correlation between executive compensation and the performance of the firm, that is, pay for performance sensitivity.

5 In this context, efficiency means a contractual arrangement with the minimum cost of balancing the incentive provision with the insurance needs of the executives.

6 Calculated based on data from the CRSP US Stock Databases ©2010 Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business.

7 See Hall and Liebman (1998), Murphy (1999), Gayle and Miller (2009a), and the data analysis in Section 2.

8 See Gabaix and Landier (2008), Treviö (2008), and Gayle and Miller (2009a).

9 See Gayle and Miller (2009a) for more details.

10 See Antle and Smith (1985, 1986), Hall and Liebman (1998), and Gayle and Miller (2009a), who find that about half the total variation in compensation can be explained by a nonlinear regression on excess returns, industry effects, and bond prices.

11 See Lorie and Niederhoffer (1968), Jaffe (1974), Finnerty (1976), and Seyhun (1986), who find that insiders tend to buy before an abnormal rise in stock prices and sell before an abnormal decline. Seyhun (1992a,b) presents evidence showing that insiders earn over 5 percent abnormal returns on average and determines that insider trades predict up to 60 percent of the total variation in one-year-ahead returns. Gayle and Miller (2009b) construct a simple self-financing dynamic portfolio strategy based on changes in asset holdings by managers that significantly outperforms the market portfolio, realizing over 90 percent of the gains that could have been achieved with perfect foresight.

12 $r_t = 1$ if the private state is bad, and $r_t = 2$ if it is good.

13 If the lower bound depends on whether the agent works or shirks, a first-best solution is attained by imposing a sufficiently harsh penalty on the agent when abnormal returns can be attained only by shirking or otherwise paying the agent the first-best fixed wage. See Mirrlees (1975).

REFERENCES

Aggarwal, Rajesh K. and Samwick, Andrew A. "Executive Compensation, Strategic Competition, and Relative Performance Evaluation: Theory and Evidence." *Journal of Finance*, December 1999, *54*(6), pp. 1999-2043; https://doi.org/10.1111/0022-1082.00180.

Antle, Rick and Smith, Abbie. "Measuring Executive Compensation: Methods and an Application." *Journal of Accounting Research*, Spring 1985, *23*(1), pp. 296-325; https://doi.org/10.2307/2490920.

Antle, Rick and Smith, Abbie. "An Empirical Investigation of the Relative Performance Evaluation of Corporate Executives." *Journal of Accounting Research*, Spring 1986, *24*(1), pp. 1-39; https://doi.org/10.2307/2490802.

Cardon, James H. and Hendel, Igal. "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey." *RAND Journal of Economics*, February 2001, *32*(3), pp. 408-27; https://doi.org/10.2307/2696362.

Chiappori, Pierre-Andre and Salanie, Bernard. "Testing for Asymmetric Information in Insurance Markets." *Journal of Political Economy*, February 2000, *108*(1), pp. 56-78; https://doi.org/10.1086/262111.

Cohen, Alma. "Asymmetric Information and Learning: Evidence from the Automobile Insurance Market." *Review of Economics and Statistics*, May 2005, *87*(2), pp. 197-207; https://doi.org/10.1162/0034653053970294.

Finnerty, Joseph E. "Insiders and Market Efficiency." *Journal of Finance*, September 1976, *31*(4), pp. 1141-48; https://doi.org/10.1111/j.1540-6261.1976.tb01965.x.

Gabaix, Xavier and Landier, Augustin. "Why Has CEO Pay Increased So Much?" *Quarterly Journal of Economics*, February 2008, *123*(1), pp. 49-100; https://doi.org/10.1162/qjec.2008.123.1.49.

Gayle, George-Levi; Golan, Limor and Miller, Robert A. "Promotion, Turnover, and Compensation in the Executive Labor Market." *Econometrica*, November 2015, *83*(6), pp. 2293-369; https://doi.org/10.3982/ECTA11020.

Gayle, George-Levi; Li, Chen and Miller, Robert A. "Was Sarbanes-Oxley Costly? Evidence from Optimal Contracting on CEO Compensation." Working Paper 2015-017B, Federal Reserve Bank of Saint Louis, 2016; https://research.stlouisfed.org/wp/more/2015-017.

Gayle, George-Levi and Miller, Robert A. "Has Moral Hazard Become a More Important Factor in Managerial Compensation?" *American Economic Review*, 2009a, *99*(5), pp. 1740-769; https://doi.org/10.1257/aer.99.5.1740.

Gayle, George-Levi and Miller, Robert A. "Insider Information and Performance Pay." *CESifo Economic Studies*, 2009b, *55*(3-4), pp. 515-41; https://doi.org/10.1093/cesifo/ifp010.

Gayle, George-Levi and Miller, Robert A. "Identifying and Testing Models of Managerial Compensation." *Review of Economic Studies*, January 2015, *82*(3), pp. 1074-118; https://doi.org/10.1093/restud/rdv004.

Hall, Brian J. and Liebman, Jeffrey B. "Are CEOs Really Paid Like Bureaucrats?" *Quarterly Journal of Economics*, August 1998, *103*(3), pp. 653-91; https://doi.org/10.1162/003355398555702.

Idson, Todd L. and Oi, Walter Y. "Workers Are More Productive in Large Firms." *American Economic Review*, May 1999, *89*(2), pp. 104-08; https://doi.org/10.1257/aer.89.2.104.

Jaffe, Jeffrey F. "Special Information and Insider Trading." *Journal of Business*, July 1974, *47*(3), pp. 410-28; https://doi.org/10.1086/295655.

Jensen, Michael C. and Murphy, Kevin J. "Performance Pay and Top-Management Incentives." *Journal of Political Economy*, April 1990, *98*(2), pp. 225-64; https://doi.org/10.1086/261677.

Lorie, James H. and Niederhoffer, Victor. "Predictive and Statistical Properties of Insider Trading." *Journal of Law and Economics*, April 1968, *11*(1), pp. 35-53; https://doi.org/10.1086/466642.

Margiotta, Mary M. and Miller, Robert A. "Managerial Compensation and the Cost of Moral Hazard." *International Economic Review*, August 2000, *41*(3), pp. 669-719; https://doi.org/10.1111/1468-2354.00080.

Mirrlees, James A. "The Theory of Moral Hazard and Unobservable Behaviour." Unpublished manuscript, October 1975, Nuffield College, Oxford; https://www.jstor.org/stable/2566946?seq=1#page_scan_tab_contents.

Murphy, Kevin J. "Executive Compensation," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*. Volume 3B, Chap. 38. Elsevier, 1999, pp. 2485-563; https://doi.org/10.1016/S1573-4463(99)30024-9.

Piketty, Thomas and Saez, Emmanuel. "Income Inequality in the United States, 1913-1998." *Quarterly Journal of Economics*, February 2003, *118*(1), pp. 1-41; https://doi.org/10.1162/00335530360535135.

Prendergast, Canice. "The Tenuous Trade-Off between Risk and Incentives." *Journal of Political Economy*, October 2002, *110*(5), pp. 1071-102; https://doi.org/10.1086/341874.

Seyhun, H. Nejat. "Insiders' Profits, Costs of Trading, and Market Efficiency." *Journal of Financial Economics*, June 1986, *16*(2), pp. 189-212; https://doi.org/10.1016/0304-405X(86)90060-7.

Seyhun, H. Nejat. "The Effectiveness of the Insider-Trading Sanctions." *Journal of Law and Economics*, April 1992a, *35*(1), pp. 149-82; https://doi.org/10.1086/467248.

Seyhun, H. Nejat. "Why Does Aggregate Insider Trading Predict Future Stock Returns?" *Quarterly Journal of Economics*, November 1992b, *107*(4), pp. 1303-31; https://doi.org/10.2307/2118390.

Terviö, Marko. "The Difference that CEOs Make: An Assignment Model Approach." *American Economic Review*, June 2008, *98*(3), pp. 642-68; https://doi.org/10.1257/aer.98.3.642.

Top Earners: Cross-Country Facts

Alejandro Badel, Moira Daly, Mark Huggett, and Martin Nybom

We provide a common set of life cycle earnings statistics based on administrative data from the United States, Canada, Denmark, and Sweden. We find three qualitative patterns, which are common across countries. First, top-earnings inequality increases over the working lifetime. Second, the extreme right tail of the earnings distribution becomes thicker with age over the working lifetime. Third, top lifetime earners exhibit dramatic earnings growth over their working lifetime. Models of top earners should account for these three patterns and, importantly, for how they quantitatively differ across countries. (JEL D31, D91, H21, J31)

Federal Reserve Bank of St. Louis *Review*, Third Quarter 2018, *100*(3), pp. 237-57.
https://doi.org/10.20955/r.2018.237-57

1 INTRODUCTION

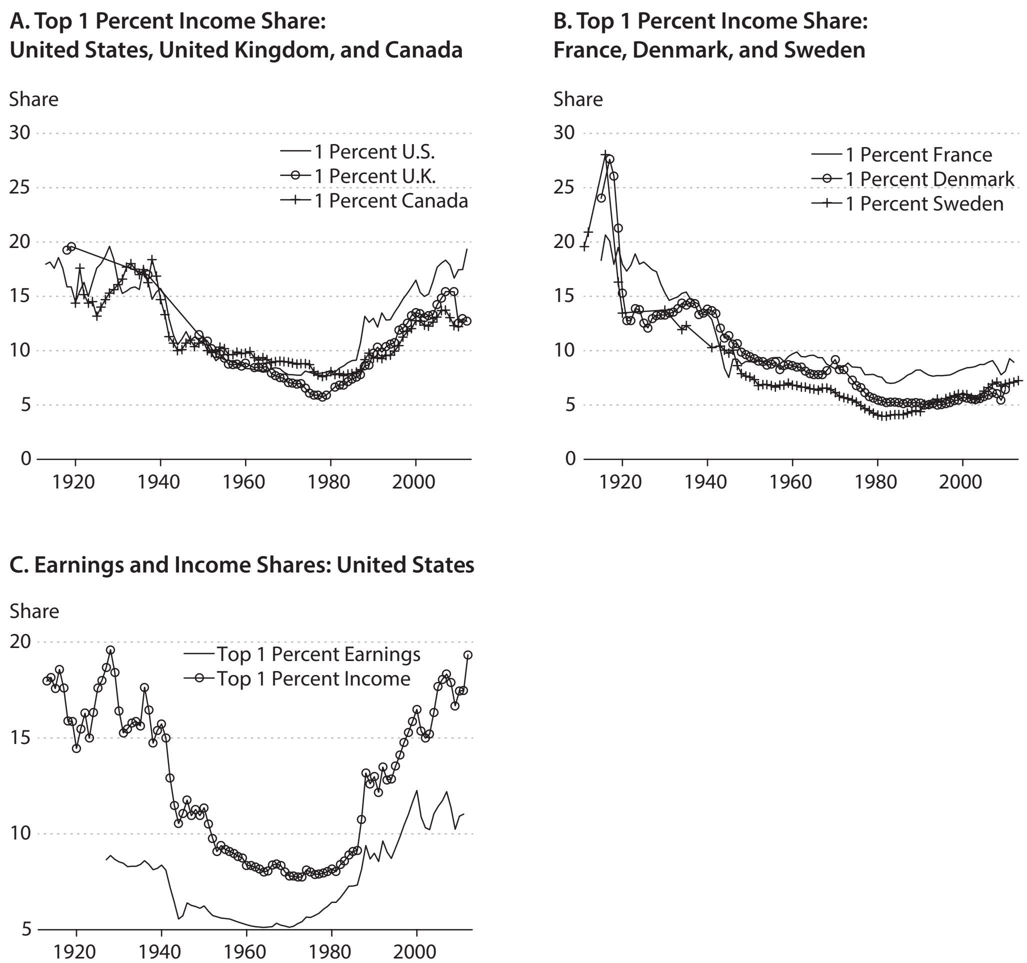
Over the past 100 years, the inequality of top incomes has followed a U-shaped pattern in the United States, the United Kingdom, and Canada. The recent increase in top-income inequality has become an important topic in academic, policy, and media discussions in these countries. In other countries, such as Denmark, France, and Sweden, income inequality also decreased strongly in the first half of the twentieth century but did not rebound strongly afterward. Figure 1 plots the top 1 percent income share for all these countries.[1]

Wage and salary income play a very important role in shaping top-income inequality patterns. First, wage and salary income has been the largest component of top incomes in the United States and Canada in recent decades (see Piketty and Saez, 2003, and Saez and Veall, 2005). Second, income inequality patterns resemble earnings inequality patterns over time. For example, Figure 1 shows that both top income and earnings shares in the United States

Alejandro Badel is an adjunct professor at Georgetown University and was an economist at the Federal Reserve Bank of St. Louis; Moira Daly is a senior advisor at Copenhagen Business School; Mark Huggett is a professor at Georgetown University; and Martin Nybom is an associate professor at the Institute for Evaluation of Labour Market and Education Policy. The authors acknowledge financial support from Danish Social Science Research Council (FSE) grant No. 300279 and support from the Federal Reserve Bank of St. Louis. They also thank Bryan Noeth for research assistance.

Figure 1

Basic Top-End Inequality Facts



NOTE: The income measure excludes capital gains, and the earnings measure is based on wages and salaries. For the United Kingdom, the sampling unit was changed in 1990, and there is a jump in the series in that year.

SOURCE: Income comes from The World Wealth and Income Database. The earnings measure for the United States is from Piketty and Saez (2003 update).

have increased over time starting before 1980. For these reasons, discussions of the determinants of top-income inequality over time and across countries have focused on theories of top-earnings inequality.

The goal of this article is to document a common set of facts concerning the dynamics of the earnings distribution over the working lifetime. We focus on the United States, Canada, Denmark, and Sweden. For these four countries, administrative data on earnings are available to researchers under strict privacy protection arrangements. The datasets we employ have the following four common features: They are large, they do not truncate earnings, they cover several decades, and, importantly, they track individuals over time. These features allow us to document the top of the earnings distribution by age or by birth cohort. They also allow us to observe the annual earnings of individuals for more than 30 years of their working lifetimes.

We find that the life cycle evolution of the earnings distribution for males follows three patterns, which are common across countries. First, top-earnings inequality increases over the working lifetime. Second, the extreme right tail of the earnings distribution becomes thicker with age over the working lifetime. Third, top lifetime earners exhibit dramatic earnings growth between their early and late working years.[2] There are important differences in the magnitudes of these facts across countries.

The patterns that we document provide empirical guidance for the specification and calibration of quantitative theoretical models aimed at understanding the distribution of earnings, income, and wealth within a given country. For many existing models of earnings distributions, these patterns also provide a challenge because these models lack forces generating extremely large earnings growth rates for top lifetime earners. The cross-country facts also provide a new challenge for quantitative theoretical work directed at understanding the underlying sources of cross-country differences in inequality. Ideally, a plausible quantitative theory should be able to account for cross-country differences in cross-sectional inequality and, simultaneously, account for the substantial cross-country differences in the three life cycle earnings facts that we document.

This article is closest to two literatures. First, there is a large literature that documents the life cycle evolution of the distribution of earnings, wages, and consumption.[3] This literature documents how summary measures of dispersion, such as the variance of log earnings, wages, or consumption, vary with age based on survey data, controlling for time or cohort effects. Our work focuses on quantiles of the earnings distribution by age and properties of the top 1 percent by age. Focusing on quantiles is useful because these can fully describe a distribution. Much larger sample sizes and the lack of top coding allow us to address the behavior of the top 1 percent of the distribution by age. The very top of the distribution is critical for optimal tax theory (see Piketty and Saez, 2013, and Badel and Huggett, 2017) as specific statistics of the top of the distribution enter formulae that determine optimal top tax rates. Second, a recent literature uses administrative data to describe the top of the earnings distribution over time. See, for example, Guvenen, Kaplan, and Song (2014) and Guvenen et al. (2015). We differ because we focus on how three life cycle facts differ across countries.

This article is organized into four sections. Section 2 describes basic features of each dataset and provides inequality facts. Section 3 documents three facts that characterize the dynamics of earnings over the working lifetime. Section 4 discusses the ability of existing quantitative models of earnings and labor productivity to produce the three life cycle earnings facts that we document.

2 DATA

This section describes the earnings data, the samples, and some background facts.

2.1 Earnings Data

Our earnings data come from records kept by government agencies for administrative purposes. These datasets are not publicly available and are accessible only under special arrangements that protect personally identifiable information. Except for the United States, we directly access each country's microdata via the relevant statistical agency. For the United States, we lack access to the microdata, so we use the summary tables provided by Guvenen, Ozkan, and Song (2014) and Guvenen et al. (2015).

The U.S. summary tables are based on data from W-2 forms of wage and salary workers held by the Social Security Administration. Their earnings measure includes wages and salary, bonuses, and exercised stock options. The data consist of a 10 percent random sample of males with a Social Security number in the period 1978-2011. The summary tables include minimum, maximum, mean, and various percentiles of the earnings distribution for each year and include percentiles by age and year.

The earnings data for Canada come from the Longitudinal Administrative Databank (LAD) administered by Statistics Canada. LAD is a 20 percent random sample of the Canadian population covering the period 1982-2013. The earnings measure we employ is total earnings from T4 slips plus other employment income. T4 slips are issued by employers to the Canadian Revenue Agency and contain employment income and taxes deducted. T4 slips include wages, salaries and commissions, and exercised stock option benefits. Other employment income includes tips, gratuities, and director's fees not included in T4 slips.

The tax registers for Denmark are provided by Statistics Denmark. The sample period is 1980-2013. Over the sample period, the registers provide panel data on earnings for more than 99.9 percent of Danish residents between ages 15 and 70. We focus on individuals never classified as immigrants in the data. The earnings measure we employ is the sum of two variables in the registers. The first variable measures taxable wage payments and includes fringe benefits, jubilee and termination benefits, and the value of exercised stock options.[4] It excludes contributions to pension plans and to ATP (the Danish Labour Market Supplementary Pension). The second variable is ATP contributions.

Earnings data for Sweden are provided by Statistics Sweden. We have access to earnings data for 1980, 1982, and 1985-2013. The data cover the entire Swedish population with taxable income in a given year. The earnings measure is based on taxable labor market earnings reported by individuals' employers to the national tax authority.[5]

2.2 Sample Selection

Cross-sectional samples are used to produce statistics by year or by age and year. Our cross-sectional samples for Canada, Denmark, and Sweden are designed to mimic the sample selection criteria employed in the U.S. sample. Thus, we employ harmonized samples that allow cross-country comparisons.

The U.S. cross-sectional sample includes an individual earnings observation in a given year t if (i) the individual is a male age 25 to 60, (ii) earnings are greater than a time-varying threshold, denoted $\underline{e}_t^{US}$, and (iii) self-employment income does not account for more than 10 percent of the earnings and does not exceed the $\underline{e}_t^{US}$ threshold. The threshold $\underline{e}_t^{US}$ employed by Guvenen et al. (2014, 2015) is defined as half the minimum hourly wage in year t times 520 hours.

Our cross-sectional samples for Canada, Denmark, and Sweden implement these three criteria: First, each sample includes only males age 25 to 60. Second, an earnings observation is included for a given county if it exceeds a threshold ($\underline{e}_t^{CA}$, $\underline{e}_t^{DK}$, $\underline{e}_t^{SW}$). Third, we implement the self-employment income criteria described above.[6]

We provide a method to obtain harmonized samples across countries. For each country $i \in \{CA, DK, SW\}$ and year t, we calculate the minimum earnings threshold as the product of a common factor at time t, denoted $factor_t$, and median earnings $median_t^i$:

$$\underline{e}_t^i = factor_t \times median_t^i.$$

The common $factor_t$ is based on the U.S. threshold and U.S. median earnings as follows:

$$factor_t = \underline{e}_t^{US} / median_t^{US}.$$

2.3 Background Facts

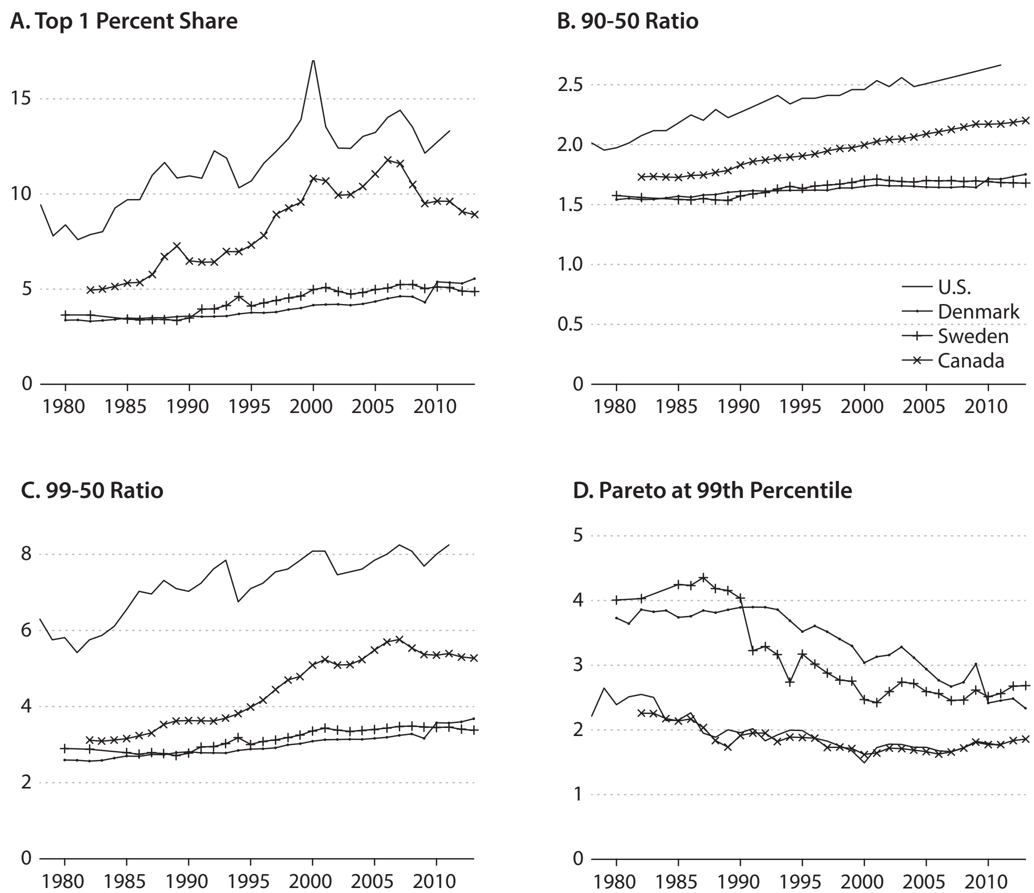
We document a number of earnings facts based on our cross-sectional samples. Figure 2 shows that, over the full sample period, the share of earnings obtained by the top 1 percent is substantially higher in the United States and Canada than in Denmark and Sweden. Furthermore, top-earnings shares trend upward in the United States and Canada over the sample period. Top-earnings shares in Denmark and Sweden also increased over the sample period but by much less than in the United States and Canada.[7] The top income share patterns in Figure 1 resemble the earnings patterns we document.

Figure 2 shows that the earnings distribution above the median in both Denmark and Sweden is more compressed compared with that for the United States. Each of the 90-50 earnings ratios for Denmark and Sweden is about three quarters of the U.S. ratio, whereas each of the 99-50 earnings ratios for Denmark and Sweden is roughly half the U.S. ratio. Thus, compression is stronger above the 90th percentile in these countries. Dividing one-half by three quarters implies that the 99-90 ratios in Denmark and Sweden have been roughly two-thirds of the U.S. 99-90 ratio. Figure 2 also shows that earnings dispersion above the 50th percentile increases in all countries over time. Specifically, over the sample period, the 90-50 and the 99-50 earnings percentile ratios increase for all countries.

Figure 2 documents the evolution of the Pareto statistic of earnings at the 99th percentile over time. This statistic is defined as $\bar{e}_{99}/(\bar{e}_{99} - e_{99})$. That is, mean earnings beyond the 99th percentile, $\bar{e}_{99}$, divided by the difference between $\bar{e}_{99}$ and the 99th percentile, e_{99}. The figure also shows that the Pareto statistic at the 99th percentile has trended downward in all countries over the sample period. A lower value for the Pareto statistic implies a thicker upper tail

Figure 2

Top-End Earnings Inequality Facts



NOTE: For the United States, the top 1 percent share and the Pareto statistic in each year are based on the assumption of a Pareto distribution within the top 1 percent and tabulated values for the 99th and 99.999th percentiles.

SOURCE: Authors' calculations based on the cross-sectional samples for each country.

in the sense that the mean, for observations above the threshold, is a higher multiple of the threshold. The Pareto statistic is particularly important in theories of taxation of top incomes or top earnings. It enters into formulas used to determine welfare- or revenue-maximizing top tax rates (see Piketty and Saez, 2013, and Badel and Huggett, 2017). Lower values of the Pareto statistic imply, other things equal, a higher revenue-maximizing top tax rate.

3 EARNINGS FACTS

We document the evolution of the earnings distribution over the working lifetime with a focus on properties of the upper tail of the distribution.

3.1 Fact 1: Top-Earnings Inequality Increases with Age

We determine how the earnings distribution above the median evolves with age. For example, we calculate the 99-50 earnings percentile ratio $e_{99,j,t}/e_{50,j,t}$ for all ages j and all sample years t. We then estimate the time and age effects (α_t, β_j) or, alternatively, the cohort and age effects (γ_c, β_j) in the regressions below. An individual's birth year (i.e., cohort) is denoted c. Clearly, cohort c, current age j, and current year t are linearly related: $c = t - j$. The cohort-effects regression controls for cohort-specific effects that impact the 99-50 ratio for a cohort at any age, whereas the time-effects regression controls for time-specific effects that impact the 99-50 ratio for all age groups alive at that time. The variables D_j, D_t, D_c are dummy variables that take the value 1 when the observation occurs at age j, year t, or cohort c, respectively. We employ a full set of age, year, and cohort dummy variables:

$$\text{Time Effects: } e_{99,j,t}/e_{50,j,t} = \alpha_t D_t + \beta_j D_j + \varepsilon_{j,t};$$

$$\text{Cohort Effects: } e_{99,j,t}/e_{50,j,t} = \gamma_c D_c + \beta_j D_j + \varepsilon_{j,t}.$$

We use the estimated age effects $\hat{\beta}_j$ to describe how the 99-50 earnings percentile ratio evolves with age. We plot the estimated age coefficients adjusted by a constant $\hat{\beta}_j + k$. The constant k is chosen so that the height of the age profile at age 45 equals the empirical 99-50 ratio for those age 45 years in 2010 for each country.[8]

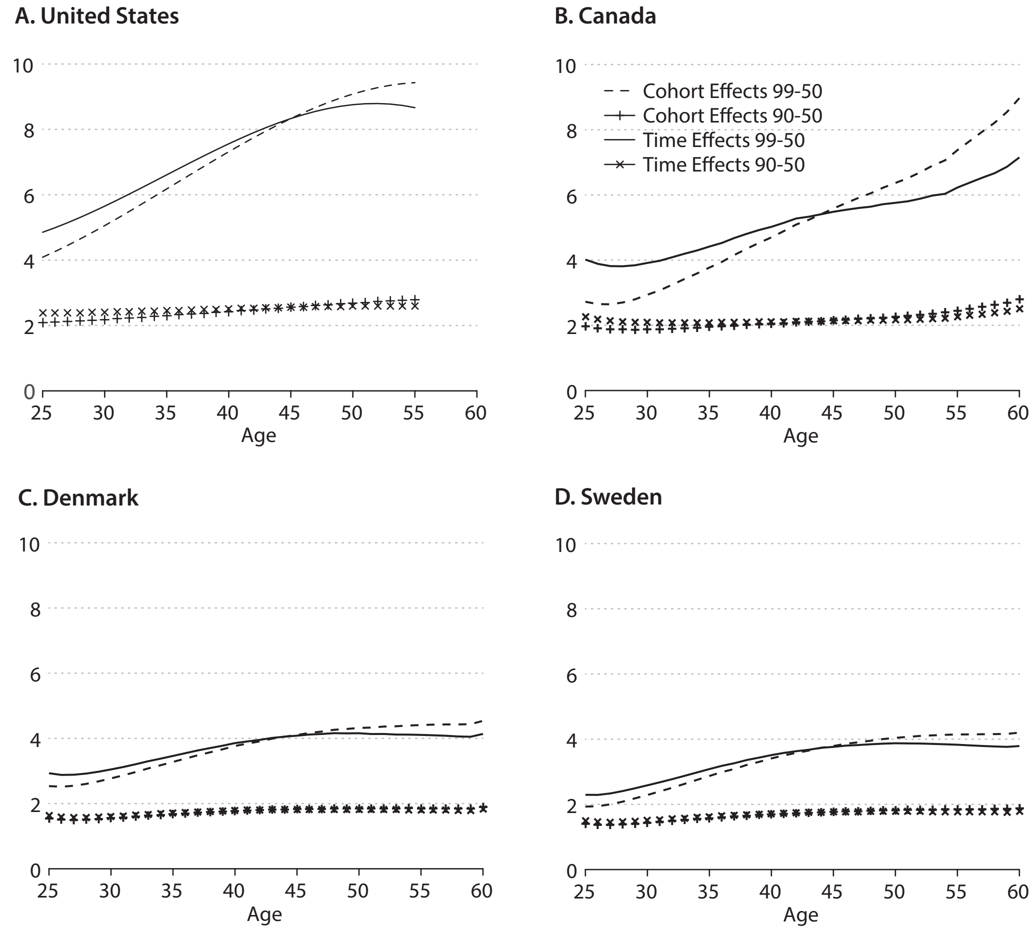
Figure 3 presents the results. The main finding is that the 90-50 and the 99-50 ratios tend to increase with age in all countries. In this sense there is fanning out in the top half of the distribution with respect to the median in all countries. The cohort-effects view produces a more dramatic pattern of fanning out compared with the time-effects view. The most striking pattern occurs for the 99-50 ratio. First, the 99-50 ratio is much larger at any age in the United States and Canada compared with Denmark and Sweden. Second, the 99-50 ratio roughly doubles from age 25 to age 55 in each country under the cohort-effects view. Thus, we conclude that there is growing earnings dispersion with age above the median and that this is driven by earnings beyond the 90th percentile.

Many studies have documented growth in summary measures of earnings or income dispersion with age for individuals or households based on dispersion measures such as the variance of log earnings or the Gini coefficient. The results in Figure 3 indicate that one reason summary measures display growing dispersion with age is because of the behavior of the very top of the distribution compared with the median.

To put these results into perspective, it is useful to characterize how real median earnings evolve with age.[9] Figure 4 provides the results of regressing real median earnings by age and time effects or age and cohort effects. Median earnings display a hump-shaped pattern with

Figure 3

Percentile Ratios: 90-50 and 99-50 Ratios by Age



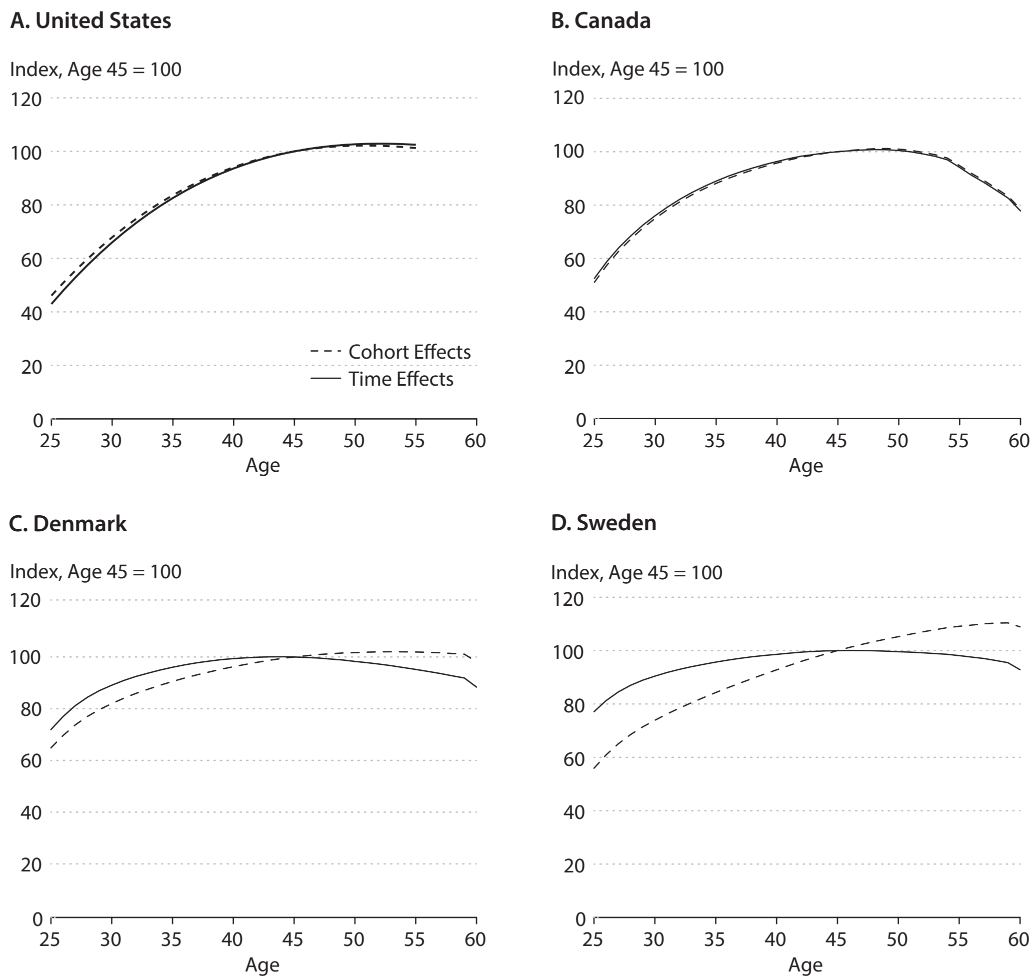
NOTE: The figure plots the estimated age coefficients after adding a vertical shift term, so each figure is normalized to equal the data value of the 99-50 ratio or the 90-50 ratio at age 45 in the year 2010.

age in each country. Many previous studies have documented that male earnings or wage rates by age are hump-shaped over the working life.[10]

Figure 4 shows that median earnings in the United States and Canada approximately double with age from age 25 to 50. This holds regardless of whether one controls for time or for cohort effects. In contrast, for Denmark and Sweden the time-effects view implies that the median earnings profile is flatter, with less than a doubling of median earnings. Focusing on the time-effects view across countries reveals substantial differences in the timing of the peak of the earnings profile. For the United States and Canada, median earnings peak near age 50, whereas for Denmark and Sweden the peak occurs in the early 40s.

Figure 4

Median Earnings by Age



NOTE: The figure plots the estimated age coefficients after adding a vertical shift term, so each figure is normalized to equal 100 at age 45.

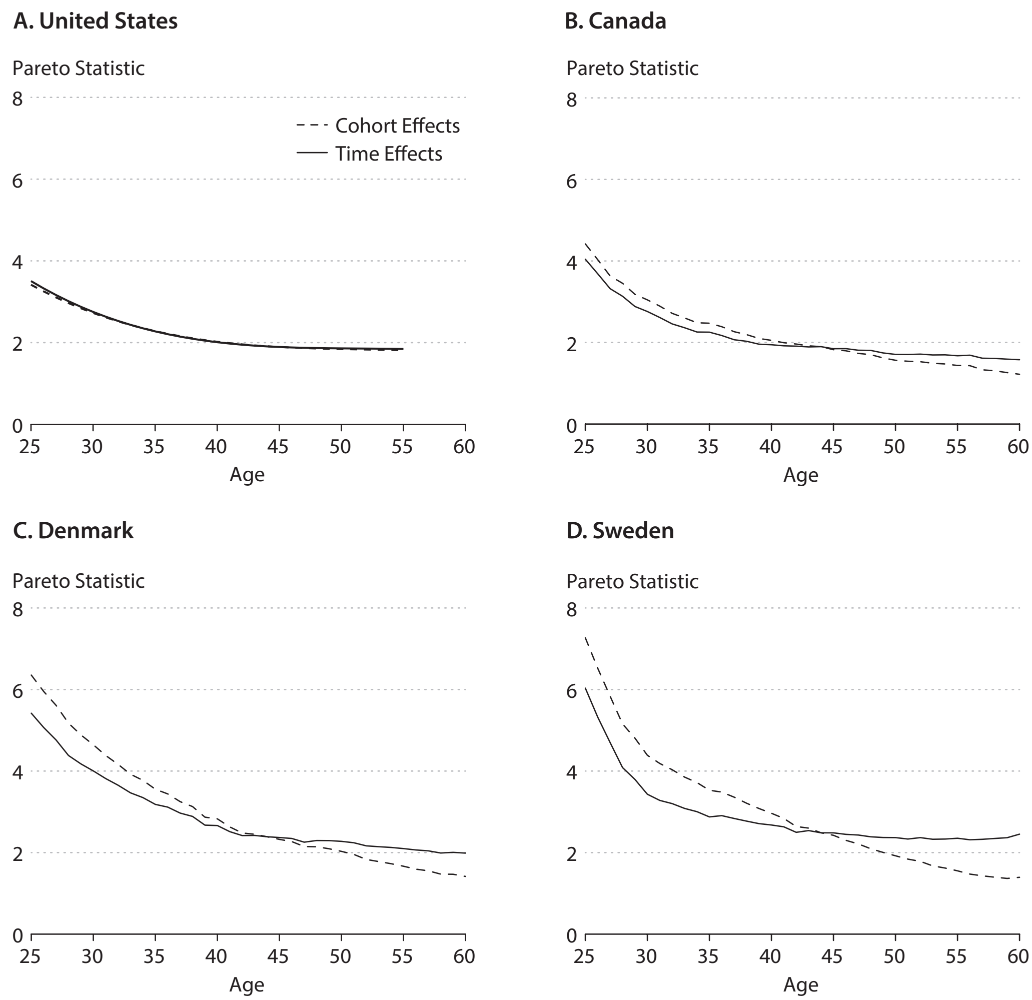
3.2 Fact 2: The Upper Tail Becomes Thicker with Age

Next we analyze how the Pareto statistic at the 99th percentile evolves with age. This is a way to describe how the thickness of the upper tail of the earnings distribution evolves with age. To do so, we run the two basic regressions from the previous section after replacing ratios of earnings percentiles with the Pareto statistic for each age-year pair.

Figure 5 shows that the Pareto statistic declines with age in all countries. This holds in both the time- and cohort-effects regressions. Thus, the upper tail of the earnings distribution becomes thicker with age in each country in the sense that mean earnings beyond this thresh-

Figure 5

Pareto Statistic at the 99th Percentile by Age



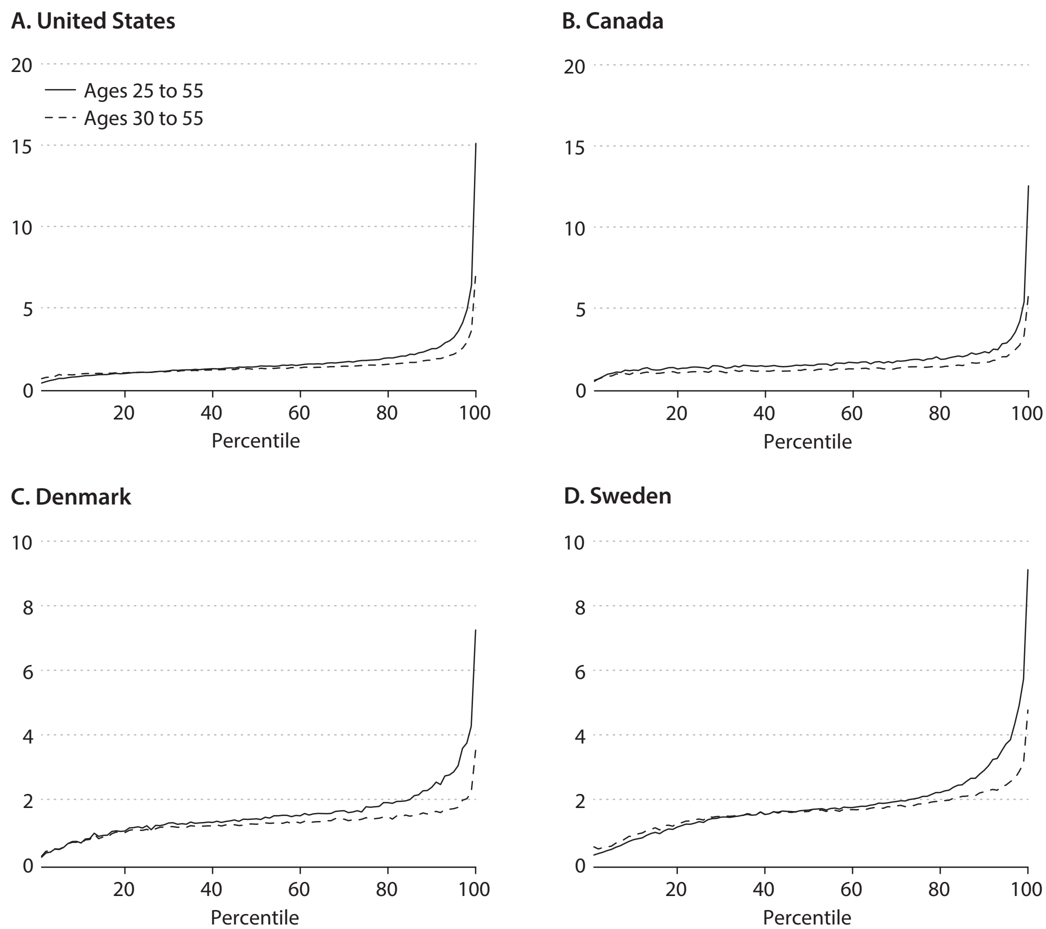
NOTE: The figure plots the estimated age coefficients after adding a vertical shift term, so each figure is normalized to equal the data value of the Pareto statistic at age 45 in the year 2010.

old is a growing multiple of the threshold with age. To the best of our knowledge, this fact has not been documented in the existing literature for a wide collection of countries.

It is interesting to compare the Pareto statistic in different age groups with the Pareto statistic in cross-sectional data previously documented in Figure 2. For the United States, the Pareto statistic at the 99th percentile in cross-sectional data is below 2 in the last two decades of the sample period. In Figure 5, it is below 2 in the United States for age groups above age 40, while it is above 2 for age groups below age 40. This suggests that the cross-sectional Pareto statistic for the United States is largely determined by the earnings distribution for males

Figure 6

Earnings Growth by Lifetime Earnings Group



NOTE: The figure plots the ratio of mean group earnings at age 55 to mean group earnings at age 25, as well as the ratio of mean group earnings at age 55 to mean group earnings at age 30 for groups sorted by percentile of lifetime earnings.

SOURCE: U.S. data are from Guvenen et al. (2015). The results for the other countries are based on our calculations from country longitudinal data.

age 40 and older. The same patterns hold in Canadian data. Thus, the cross-sectional Pareto statistic seems to be driven by the tail properties holding for older earners in both countries.

3.3 Fact 3: Top Lifetime Earners Have Dramatic Earnings Growth

We now use the longitudinal feature of each dataset. For each male in the longitudinal sample, we compute lifetime earnings *LE* as follows: $LE^i = \sum_{t \in T} \frac{\max\{e_t^i, \underline{e}_t\}}{p_t}$, where e_t^i is individual *i*'s nominal earnings in year *t*, $\underline{e}_t$ is the minimum earnings threshold used to construct the cross-section sample, p_t is a country price index in year t, and T is the set of years for which earnings observations are available.[11] We then sort males in the longitudinal sample into 100 bins based on the percentiles of the lifetime earnings distribution. Bin 100 corresponds to males with lifetime earnings above the 99th percentile, whereas bin 1 corresponds to males with lifetime earnings below the 1st percentile. Appendix A.1 describes the construction of the longitudinal data samples.

Figure 6 contains two plots for each country. It plots the ratio of mean real earnings at age 55 to mean real earnings at age 25 for individuals sorted by lifetime earnings bins, as well as the ratio of mean real earnings at age 55 to mean real earnings at age 30. In both plots the grouping of individuals into lifetime earnings bins is unchanged. Thus, for a given country, the two plots differ only insofar as there is growth in real mean earnings for the group from age 25 to age 30.

Figure 6 documents that earnings growth is greater for groups with larger lifetime earnings. It also documents the remarkable fact that the highest lifetime earnings groups (i.e., groups in lifetime earnings bins 96-100) have a much larger earnings growth rate than those with lifetime earnings close to the median (i.e., those in bin 50). The top lifetime earnings bin in the United States and Canada have a 13- to 15-fold increase in earnings from age 25 to age 55. The top lifetime earnings bin in Denmark and Sweden have a seven- to ninefold increase in earnings from age 25 to age 55. Thus, there are large, systematic differences in group earnings growth rates over the working lifetime, particularly at the very top. The large differences at the top imply that in each country, top lifetime earners tend to become top earners late in the working lifetime. We anticipate that Fact 3 will be particularly useful in empirically disciplining quantitative theories of top earners. We conjecture that theories built on temporary sources of earnings variation will struggle to produce Fact 3.

4 DISCUSSION

We close the article by discussing the potential relevance of the three earnings facts that we document for economic models of the distribution of earnings and wage rates over the working lifetime. We do so by briefly discussing two prominent articles that offer a quantitative-theoretical account of the changes in U.S. cross-sectional inequality measures.

4.1 Models of Changes in Cross-Sectional Inequality

Heathcote, Storesletten, and Violante (2010) provide a quantitative-theoretical account for the changes in U.S. cross-sectional earnings, consumption, and hours inequality. The key exogenous driving force in their model is changes in transitory and persistent idiosyncratic productivity shocks. They measure the time-varying variances of these shocks from panel data on U.S. wage rates. They find that both transitory and persistent innovation variances increase over time. They then show that their model accounts for the rise in measures of U.S. household earnings and consumption dispersion, among other facts, based on the measured process for productivity shocks.

Kaymak and Poschke (2016) provide a quantitative-theoretical account for changes in U.S. top-end wealth inequality over the past half century. They consider these three exogenous sources for the increase in U.S. wealth inequality: changes in taxes, transfers, and productivity shocks. They measure changes in U.S. corporate, estate, and income taxes over time, and they measure changes in the level and progressivity in Social Security benefits. Finally, they calibrate an idiosyncratic productivity shock process to match evidence for the rise in U.S. earnings/wage dispersion over time. Their shock process captures persistent and transitory sources of variation. They find that the rise in wage dispersion, the change in taxes (i.e., decrease in some top tax rates), and the increase in transfers all contributed to the increase in top-end U.S. wealth inequality. They find a particularly important contribution from the increase in top-end wage dispersion.

4.2 Idiosyncratic Productivity Shocks

The articles by Heathcote, Storesletten, and Violante (2010) and Kaymak and Poschke (2016) stress the role of idiosyncratic productivity shocks. Productivity in their models corresponds to wage rates in the data. We now compare properties of the process used in these articles with the facts for earnings that we document.

While earnings and wage rates are not strictly comparable, we think the comparison is still useful. Many age patterns in wage rate data also hold in earnings data. For example, Heathcote, Storesletten, and Violante (2005) show that the rise in both the variance of log earnings and the variance in log wage rates happens with age in U.S. data by similar amounts. In addition, cross-sectional inequality in log earnings and in log wage rates rises by a similar magnitude as documented by Heathcote, Storesletten, and Violante (2010). Finally, it is widely believed that productivity differences (i.e., earnings per work hour) are key in accounting for the earnings of top earners in U.S. data rather than work-hour differences.

The process used in each article is summarized below. Kaymak and Poschke (2016) model a worker's productivity as a finite Markov process, where the transition probabilities are given by the matrix Π. Productivity w takes on six values $(z_1,\ldots,z_6)$, where z_6 corresponds to an extraordinarily high level of productivity.[12] Heathcote, Storesletten, and Violante (2010) model log productivity as the sum of an age component μ_{j+1}, a persistent shock η_{j+1}, and a purely transitory shock v_{j+1}. The age component is common to all agents of age $j+1$, whereas the shock components are agent specific:

$$(KP)\ \ Prob\left(w_{j+1}=z' \mid w_j=z\right)=\Pi\left(z' \mid z\right)$$

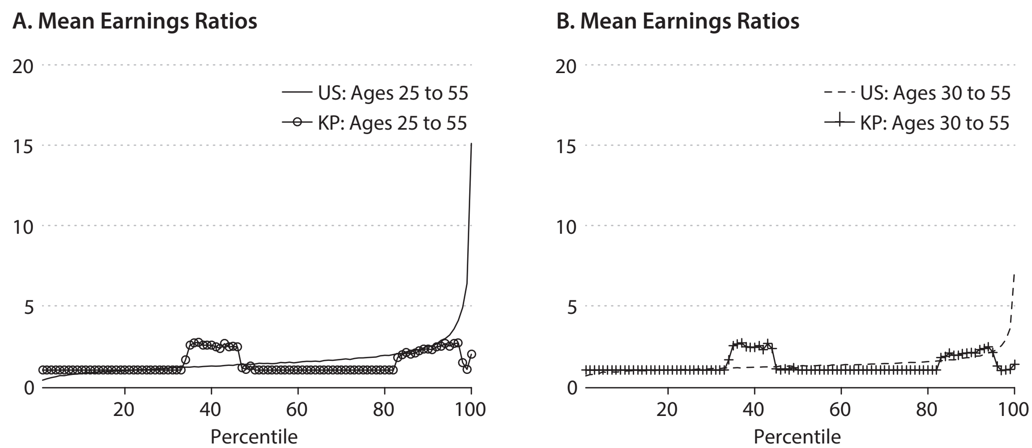
$$(HSV)\ \ \log w_{j+1}=\mu_{j+1}+\eta_{j+1}+v_{j+1},\ \text{and}\ \ \eta_{j+1}=\rho\eta_j+\omega_{j+1}.$$

We now simulate 2 million wage histories from age 20 to age 60 using the Kaymak-Poschke process above. The inputs are an initial distribution, the workers matrix Π above, and the six productivity values.[13] We highlight the implications of the Kaymak-Poschke process for Fact 3 from Section 3.3.

Figure 7 presents ratios of earnings across ages for different lifetime earnings groups in the Kaymak-Poschke model and in U.S. data from Figure 6. A measure of lifetime earnings is

Figure 7

Earnings Ratios: Kaymak-Poschke Model and U.S. Data



NOTE: Figure 7A plots the ratio of mean group earnings at age 55 to mean group earnings at age 25 in both U.S. data and the Kaymak-Poschke model. Figure 7B repeats this plot but uses data at age 55 and age 30. The horizontal axis sorts males and model agents by percentiles of lifetime earnings.

SOURCE: Kaymak and Poschke, 2016.

computed for each agent in the model based on earnings from age 25 to age 60. Agents are then placed into 100 bins according to their percentile of lifetime earnings. Thus, U.S. data and model data are treated symmetrically.

Figure 7 shows that the ratios in the model data are typically below the corresponding ratios found in U.S. earnings data. This holds most strikingly for the several highest lifetime earnings bins. For the highest bin in the Kaymak-Poschke model, the earnings ratio is below 2.5 for both the 55-25 age ratio and the 55-30 age ratio. In contrast, the corresponding ratios for the highest earnings bin in U.S. data are roughly 15 and 7.

One possible reason for the difference between the model and U.S. data is that there is mean reversion at the highest productivity state z_6 back to lower productivity states. Such mean reversion is one reason the model successfully concentrates a large fraction of wealth held by top wealth holders similar in magnitude to that found in U.S. data. Agents with shock z_6 save a large part of their labor income because this state is, to an important degree, transitory.[14]

We conjecture that models that rely only on purely temporary sources of earnings variation to account for the extreme right tail of the earnings distribution will also fail to produce the strong earnings growth for top lifetime earners documented in Figure 6. We suspect that simulations of productivity from the Heathcote-Storesletten-Violante model will also be below the patterns found in U.S. data for top lifetime earners. This is because their model relies on a persistent but mean-reverting component and a purely temporary component to account for the right tail of the productivity distribution.

We conjecture that models that allow for systematic differences in earnings growth over the working lifetime will be important to account for the earnings profiles of top lifetime earners. Human capital models are promising in this regard. Specifically, some human capital models allow agents to permanently differ in learning ability. Those with high learning ability optimally choose steeper mean earnings profiles via an investment in skill formation. Badel, Huggett, and Luo (2018) provide a model with this feature that can produce the properties that we document in Facts 1-3. Learning ability differences also help produce Fact 2—the fall in the Pareto statistic with age. The mechanism is the same. High productivity agents make skill investments, even late in the working lifetime, and these investments are a force that cause the earnings of agents above the 99th percentile within an age group to grow faster than those at the 99th percentile. ■

APPENDIX A

A.1 Longitudinal Samples

For Canada, our raw data consist of all individuals in the LAD dataset. The LAD is a 20 percent random subsample from the Canadian population that either filed a T1 form or received Canadian child benefits in any year since 1982 and had a social insurance number.[15] For Denmark we use tax registry data kept by Statistics Denmark. For Sweden we use tax registers kept in the Income and Taxation Register of Statistics Sweden. These data come from the Swedish Tax Agency, which collects information from virtually all persons who are Swedish citizens or who hold a residence permit.

We construct a longitudinal sample for Canada, Denmark, and Sweden. These three samples mimic the construction of the U.S. longitudinal sample described in Guvenen et al. (2015). The sample period is 1982-2013 for Canada, 1980-2013 for Denmark, and 1980, 1982, and 1985-2013 for Sweden. Thus, the sample period for each country spans a horizon of more than 30 years.

Our longitudinal sample for each of these three countries contains all individual histories that satisfy Conditions 1-4 below. The following notation is employed: e_t^i is individual i's nominal earnings, $\underline{e}_t$ is a minimum nominal earnings threshold, and se_t^i is individual i's self-employment income. (1) The individual is male and age 24, 25, or 26 in the first year of the sample period. (2) The individual has a valid non-missing earnings observation in every year of the sample period. (3) There are more than 15 years for which $e_t^i > \underline{e}_t$. (4) There are less than 9 years for which $se_t^i > \max\{\underline{e}_t, 0.1 e_t^i\}$.

We now provide a brief discussion of the specifics of imposing Conditions 1-4 in the longitudinal samples for each country. Condition 1 is straightforward to implement. All properties of mean earnings for groups by age are understood to be for the central age within the group. Condition 3 is straightforward to implement in each country. We simply employ the threshold used in the construction of each cross-sectional sample. We implement Condition 4 in Canada and Denmark by using the self-employment income measure described in Section 2 and employed in the construction of the cross-sectional sample. The longitudinal samples contain the following number of males after rounding to the nearest 100: (1) Canada 65,000, (2) Denmark 73,300, and (3) Sweden 143,400.

A.2 Pareto Statistic from SSA Data

Pareto statistics at the 99th percentile are not provided by Guvenen et al. (2014, 2015). Based on the statistics provided, we estimate the Pareto statistics for the United States in two different ways. First, for the Pareto statistics depicted in Figure 2D, we use the 99th and 99.999th percentiles of earnings, provided by Guvenen et al. (2014) for each sample year, to estimate the coefficient of a Type-I Pareto distribution for earnings above the 99th percentile. Such coefficient is the Pareto statistic. For the Pareto statistic at the 99th percentile by age group and year used to create the life cycle profiles in Figure 5, we employ the method described in Badel, Huggett, and Luo (2018), which uses the 95th and 99th percentiles that are provided by age group and year, to estimate a Pareto distribution for earnings above the 95th percentile.

A.3 Price Index Data Sources

Sources for price indexes are as follows:

Canada: Series number CPALCY01CAA661N, Consumer Price Index: Total, All Items for Canada©, Index 2010 = 1, Annual, Not Seasonally Adjusted; https://fred.stlouisfed.org.

Denmark: Available from Statistics Denmark's StatBank Denmark; http://www.statbank.dk/statbank5a/default.asp?w=1280. Year 2000 was the base year used.

Sweden: Statistics Sweden's official consumer price index series; http://www.scb.se/hitta-statistik/statistik-efter-amne/priser-och-konsumtion/konsumentprisindex/konsumentprisindex-kpi/pong/tabell-och-diagram/konsumentprisindex-kpi/kpi-faststallda-tal-1980100/.

A.4 Descriptive Statistics

Tables A1-3 present descriptive statistics for Canada, Denmark, and Sweden, respectively, for the cross-sectional sample.

Table A1

Summary Statistics: Cross-Sectional Samples, Canada

Year	*nobs*	*mean e*	*e50*	*e99*
1982	886,920	24,400	23,100	72,000
1983	886,310	25,400	24,200	74,800
1984	902,625	26,800	25,500	79,600
1985	914,700	28,100	26,700	84,300
1986	947,720	29,200	27,600	89,200
1987	956,945	30,700	28,800	95,300
1988	983,375	32,800	30,200	106,600
1989	1,012,465	34,700	31,600	114,500
1990	1,028,790	35,400	32,200	117,000
1991	1,022,650	36,000	32,900	119,300
1992	1,024,415	36,700	33,600	121,700
1993	1,028,755	37,300	33,800	124,900
1994	1,033,960	38,100	34,300	130,800
1995	1,044,510	39,100	34,900	139,000
1996	1,048,970	40,000	35,300	147,100
1997	1,058,555	41,900	36,100	160,500
1998	1,065,610	43,600	37,100	174,300
1999	1,084,320	45,100	38,100	182,600
2000	1,101,815	47,800	39,400	200,700
2001	1,140,225	49,000	40,200	210,500
2002	1,137,365	49,600	41,000	208,800
2003	1,149,010	50,800	42,000	214,300
2004	1,162,555	52,700	43,100	225,900
2005	1,177,270	55,200	44,400	243,500
2006	1,186,490	57,900	45,900	261,600
2007	1,199,525	60,000	47,400	273,200
2008	1,210,295	61,300	48,900	270,900
2009	1,201,615	59,500	48,000	257,800
2010	1,200,940	61,400	49,400	264,300
2011	1,221,400	63,700	51,100	275,600
2012	1,233,235	65,300	52,600	279,000
2013	1,241,750	67,000	53,900	284,200

NOTE: Earnings statistics are rounded to the nearest 100 for confidentiality. The notations *nobs*, *mean e*, *e*50, and *e*99 denote the number of observations, mean earnings, and the 50th and 99th earnings percentiles, respectively.

SOURCE: Statistics Canada.

Table A2

Summary Statistics: Cross-Sectional Samples, Denmark

Year	*nobs*	*mean e*	*e50*	*e99*
1980	871,620	118,228	113,579	294,873
1981	859,167	127,065	122,948	318,449
1982	866,315	141,548	137,413	352,795
1983	879,347	151,118	146,798	379,600
1984	890,302	160,415	154,354	408,569
1985	906,252	169,582	161,345	435,802
1986	917,972	181,792	172,378	464,136
1987	924,403	196,046	185,319	508,165
1988	926,431	206,465	195,574	535,390
1989	927,703	212,492	200,965	560,340
1990	936,043	217,956	206,008	579,697
1991	935,039	223,188	211,658	589,631
1992	943,109	228,499	217,469	605,339
1993	941,600	228,247	218,277	606,725
1994	951,024	239,598	227,380	647,593
1995	962,977	248,388	234,389	675,665
1996	972,286	255,076	240,615	695,749
1997	983,871	264,676	248,812	725,424
1998	998,120	273,832	255,626	765,677
1999	1,005,814	285,656	266,406	805,958
2000	1,011,325	296,760	275,335	850,898
2001	1,012,968	307,839	284,840	891,303
2002	1,009,869	315,493	292,833	917,934
2003	999,303	320,344	298,258	936,881
2004	993,586	328,709	306,149	960,951
2005	990,605	338,733	314,951	996,671
2006	989,524	351,789	325,779	1,040,338
2007	984,137	368,622	339,457	1,101,163
2008	969,799	386,238	353,782	1,160,732
2009	942,820	382,500	356,390	1,127,868
2010	923,739	402,285	365,001	1,304,489
2011	918,254	409,964	371,298	1,325,142
2012	913,586	417,780	376,711	1,355,580
2013	911,549	424,334	379,914	1,398,553

NOTE: The notations *nobs*, *mean e*, *e*50, and *e*99 denote the number of observations, mean earnings, and the 50th and 99th earnings percentiles, respectively. All percentiles calculated from Danish data are six observation local averages, a confidentiality requirement.

SOURCE: Statistics Denmark.

Table A3

Summary Statistics: Cross-Sectional Samples, Sweden

Year	*nobs*	*mean e*	*e50*	*e99*
1980	1,845,140	79,441	75,031	217,579
1981	—	—	—	—
1982	1,830,333	91,005	86,952	250,336
1983	—	—	—	—
1984	—	—	—	—
1985	1,615,820	118,882	111,795	311,643
1986	1,627,315	128,328	121,011	332,480
1987	1,644,682	138,457	130,327	363,534
1988	1,665,408	149,939	141,447	390,681
1989	1,691,587	164,961	156,272	423,839
1990	1,871,002	175,721	167,993	466,135
1991	1,898,011	187,592	178,511	524,915
1992	1,875,173	187,430	180,315	531,928
1993	1,840,234	189,948	183,420	554,667
1994	1,838,130	197,666	189,330	602,450
1995	1,856,135	204,850	197,545	594,136
1996	1,857,699	215,984	206,560	637,712
1997	1,860,797	225,708	215,075	672,172
1998	1,883,857	235,753	223,189	710,840
1999	1,914,785	244,286	229,501	745,901
2000	1,945,461	257,441	239,017	803,702
2001	1,962,558	270,076	248,633	853,088
2002	1,963,068	277,617	257,149	870,017
2003	1,945,148	282,512	263,059	880,435
2004	1,928,007	288,791	269,771	910,426
2005	1,914,243	299,104	278,085	946,117
2006	1,917,082	311,050	288,703	991,729
2007	1,913,805	325,196	300,023	1,043,752
2008	1,906,596	340,417	312,946	1,091,904
2009	1,875,741	342,493	316,552	1,097,314
2010	1,871,732	352,699	326,069	1,125,863
2011	1,885,636	365,852	336,988	1,166,034
2012	1,886,082	375,238	346,504	1,179,256
2013	1,886,746	381,534	353,416	1,193,581

NOTE: The notations *nobs*, *mean e*, *e*50, and *e*99 denote the number of observations, mean earnings, and the 50th and 99th earnings percentiles, respectively.

SOURCE: Statistics Sweden.

NOTES

1 Roine, Vlachos, and Waldenstrom (2009) and Alvaredo et al. (2013), among others, have documented inequality patterns over the past 100 years for many developed countries, including those in Figure 1.

2 Lifetime earnings are defined as a present value (or weighted sum) taken over the full history of annual earnings of a worker's life. Top lifetime earners are those in the top 1 percent of the lifetime earnings distribution.

3 See Deaton and Paxson (1994), Storesletten, Telmer, and Yaron (2004), Heathcote, Storesletten, and Violante (2005) or Huggett, Ventura, and Yaron (2011) for the United States; Creedy and Hart (1979) and Blundell and Etheridge (2010) for the United Kingdom; Brzozowski et al. (2010) for Canada; and Domeij and Floden (2010) for Sweden.

4 This variable, labeled LOENMV in the registers, has changed coverage over time. For example, the value of exercised stock options were not included prior to 2000.

5 The earnings measure comes from Statistics Sweden variable ARBINK up to 1985 and from variable LONEINK thereafter. These measures include some labor-related benefits such as parental leave benefits and short-term sick leave benefits. Variable LONEINK includes income from closely held businesses starting in 1994. Part of the value of realized stock options are included in the earnings measure.

6 For Canada, self-employment income is measured with the LAD variable SEI, which measures the sum of net income from self-employment. For Denmark, self-employment income is measured with the Statistics Denmark variable NETOVSKUDGL. For Sweden it is measured with variable FINK, which measures net entrepreneurial income.

7 Domeij and Floden (2010) provide evidence that the 99-50 earnings percentile ratio, based on family earnings, rises from about 1.6 to 1.8 in Sweden between 1990 and 2000.

8 For the United States, the available summary tables contain data for $j \in \{25,35,\dots,55\}$, so estimating one age coefficient β_j for each $j = 25,26,27,\dots,60$ is not possible. Therefore, we replace the age effects β_j in the regressions above with a third-order polynomial in age $P(j;\theta) = \theta_0 + \theta_1 j + \theta_2 j^2 + \theta_3 j^3$ and set the estimated age effects to $\hat{\beta}_j = P(j;\hat{\theta})$, where $\hat{\theta}$ are the estimated polynomial coefficients.

9 Appendix A.3 states sources for the price indexes that are used to deflate earnings.

10 For example, the *Review of Economic Dynamics* special issue on Cross-Sectional Facts for Macroeconomists in 2010 covers nine countries, and Lagakos et al. (2016) covers 18 countries.

11 The set T^{US} is based on 1978-2011, T^{CA} is based on 1982-2013, T^{DK} is based on 1980-2013, and T^{SW} is based on 1980, 1982, and 1985-2013. Price indexes are in Appendix A.3.

12 The process sets $(z_1,z_2,z_3,z_4,z_5,z_6) = (6.7,19.2,20.5,58.4,61.4,1222)$.

13 The initial distribution mentioned above is the initial distribution of descendants productivity constructed from the relevant matrices in Section 4 of Kaymak and Poschke (2016).

14 The retirement period is also an important force in the Kaymak-Poschke model for wealth accumulation for those with high productivity realizations.

15 A person who is sampled in a particular reference year is also selected in all other available years.

REFERENCES

Alvaredo, Facundo; Atkinson, Anthony B.; Piketty, Thomas and Saez, Emmanuel. "The Top 1 Percent in International and Historical Perspective." *Journal of Economic Perspectives*, Summer 2013, *27*(3), pp. 3-20.

Alvaredo, Facundo; Atkinson, Anthony B.; Piketty, Thomas and Saez, Emmanuel. The World Wealth and Income Database; http://topincomes.g-mond.parisschoolofeconomics.eu/.

Badel, Alejandro and Huggett, Mark. "The Sufficient Statistic Approach: Predicting the Top of the Laffer Curve." *Journal of Monetary Economics*, May 2017, *87*, pp. 1-12; https://doi.org/10.1016/j.jmoneco.2017.02.001.

Badel, Alejandro; Huggett, Mark and Luo, Wenlan. "Taxing Top Earners: A Human Capital Perspective." Manuscript, 2018.

Blundell, Richard and Etheridge, Ben. "Consumption, Income and Earnings Inequality in Britain." *Review of Economic Dynamics*, January 2010, *13*(1), pp. 76-102; https://doi.org/10.1016/j.red.2009.10.004.

Brzozowski, Matthew; Gervais, Martin; Klein, Paul and Suzuki, Michio. "Consumption, Income and Wealth Inequality in Canada." *Review of Economic Dynamics*, January 2010, *13*(1), pp. 52-75; https://doi.org/10.1016/j.red.2009.10.006.

Creedy, John and Hart, Peter E. "Age and the Distribution of Earnings." *Economic Journal*, June 1979, *89*(354), pp. 280-93; https://doi.org/10.2307/2231602.

Deaton, Angus and Paxson, Christina. "Intertemporal Choice and Inequality." *Journal of Political Economy*, June 1994, *102*(3), pp. 437-67; https://doi.org/10.1086/261941.

Domeij, David and Floden, Martin. "Inequality Trends in Sweden 1978-2004." *Review of Economic Dynamics*, January 2010, *13*(1), pp. 179-208; https://doi.org/10.1016/j.red.2009.10.005.

Guvenen, Fatih; Kaplan, Greg and Song, Jae. "How Risky Are Recessions for Top Earners?" *American Economic Review Papers and Proceedings*, May 2014, *104*(5), pp. 1-6; https://doi.org/10.1257/aer.104.5.148.

Guvenen, Fatih; Karahan, Fatih; Ozkan, Serdar and Song, Jae. "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" Federal Reserve Bank of New York *Staff Reports*, February 2015, Staff Report No. 710.

Guvenen, Fatih; Ozkan, Serdar and Song, Jae. "The Nature of Countercyclical Income Risk." *Journal of Political Economy*, June 2014, *122*(3), pp. 621-60; https://doi.org/10.1086/675535.

Heathcote, Jonathan; Storesletten, Kjetil and Violante, Giovanni. "Two Views of Inequality over the Life-Cycle." *Journal of the European Economic Association*, April 2005, *3*(2), pp. 765-75; https://doi.org/10.1162/1542476054473305.

Heathcote, Jonathan; Storesletten, Kjetil and Violante, Giovanni. "The Macroeconomic Implications of Rising Wage Inequality in the United States." *Journal of Political Economy*, August 2010, *118*(4), pp. 681-722; https://doi.org/10.1086/656632.

Huggett, Mark; Ventura, Gustavo and Yaron, Amir. "Sources of Lifetime Inequality." *American Economic Review*, December 2011, *101*(7), pp. 2923-54; https://doi.org/10.1257/aer.101.7.2923.

Kaymak, Baris and Poschke, Markus. "The Evolution of Wealth Inequality Over Half a Century: The Role of Taxes, Transfers and Technology." *Journal of Monetary Economics*, February 2016, *77*, pp. 1-25; https://doi.org/10.1016/j.jmoneco.2015.10.004.

Lagakos, David; Moll, Benjamin; Porzio, Tommaso; Qian, Nancy and Schoellman, Todd. "Life-Cycle Wage Growth across Countries." *Journal of Political Economy*, April 2018, *126*(2), pp. 797-849; https://doi.org/10.1086/696225.

Piketty, Thomas and Saez, Emmanuel. "Income Inequality in the United States, 1913-1998." *Quarterly Journal of Economics*, February 2003, *118*(1), pp. 1-39; https://doi.org/10.1162/00335530360535135.

Piketty, Thomas and Saez, Emmanuel. "Optimal Labor Income Taxation," in A. Auerbach, R. Chetty, M. Feldstein, eds., *Handbook of Public Economics*. Volume 5. Amsterdam: Elsevier, 2013, pp. 391-474; https://doi.org/10.1016/B978-0-444-53759-1.00007-8.

Roine, Jesper; Vlachos, Jonas and Waldenstrom, Daniel. "The Long-Run Determinants of Inequality: What Can We Learn From Top Income Data?" *Journal of Public Economics*, August 2009, *93*(7), pp. 974-88; https://doi.org/10.1016/j.jpubeco.2009.04.003.

Saez, Emmanuel and Veall, Michael. "The Evolution of High Incomes in Northern America: Lessons from Canadian Evidence." *American Economic Review*, June 2005, *95*(3), pp. 831-49; https://doi.org/10.1257/0002828054201404.

Storesletten, Kjetil; Telmer, Christopher and Yaron, Amir. "Consumption and Risk Sharing Over the Life Cycle." *Journal of Monetary Economics*, April 2004, *51*(3), pp. 609-33; https://doi.org/10.1016/j.jmoneco.2003.06.005.

Institutional Barriers and World Income Disparities

Ping Wang, Tsz-Nga Wong, and Chong K. Yip

Why have the income disparities between fast-growing economies and development laggards widened over the past five decades? How important is the role played by institutional barriers with relation to technology adoption? Using cross-country analysis, we find that more-severe institutional barriers in several representative lag-behind countries actually hinder the process of structural transformation and economic development, causing these countries to fall below a representative group of fast-growing economies despite having similar or even better initial states five decades ago. We also find that institutional barriers have played the most important role, accounting for more than half the economic growth in fast-growing and trapped economies and for more than 100 percent of the economic growth in the lag-behind countries. By conducting country studies, we identify that unnecessary protectionism, government misallocation, corruption, and financial instability have been key institutional barriers causing countries to either fall into the poverty trap or lag behind without a sustainable growth engine. (JEL O41, O43, O47)

Federal Reserve Bank of St. Louis *Review*, Third Quarter 2018, *100*(3), pp. 259-79.
https://doi.org/10.20955/r.2018.259-79

1 INTRODUCTION

Over the past half-century, world income disparities have widened. The gap in real gross domestic product (GDP) per capita relative to the United States between advanced and poor countries has increased. For example, the ratio of average real GDP per capita among the top 10 percent of countries to the bottom 10 percent has increased from less than 20 in 1960 to more than 40 in 1990, and to more than 50 since the turn of the new millennium (Table 1). The huge disparities remain even if we exclude outliners such as oil rich countries in OPEC and former members of the Soviet Union and Yugoslavia. This important issue has induced numerous studies seeking to understand the causes and consequences of such disparities.

Ping Wang is a research fellow at the Federal Reserve Bank of St. Louis, a professor at Washington University in St. Louis, and a research associate at the National Bureau of Economic Research; Tsz-Nga Wong is an economist at the Federal Reserve Bank of Richmond; and Chong K. Yip is a professor at the Chinese University of Hong Kong.

Table 1

Per Capita Income Ratio Between Top and Bottom 10% Countries

Year	Average per capita income, bottom 10%	Average per capita income, top 10%	Ratio
1960	606.2	12,015.0	19.8
1970	758.1	16,072.5	21.2
1980	716.9	24,545.9	34.2
1990	637.0	26,390.7	41.4
2000	657.9	35,522.1	54.0
2010	852.0	45,277.3	53.1

SOURCE: Penn World Table 8.1 (in millions of 2005 U.S. dollars).

The conventional framework developed by Lucas (1990, 2000) and Prescott (1998) uses a neoclassical aggregate production function to explain relative income gaps. In addition to differences in per capita capital, the residual gaps are calibrated as measures of relative total factor productivities (TFPs). This tends to result in unreasonably large gaps in TFPs, thereby leading to further analyses extending the basic framework. Among others, various forms of institutional barriers are believed to play an important role. Such barriers could relate to physical capital investment due to financial market frictions, as considered by Aghion, Howitt, and Mayer-Foulkes (2005) and Buera and Shin (2013). These barriers may also relate to frictions associated with human capital or health investment, as in Wang and Wang (2015). Moreover, they may relate to technology adoption, as in Basu and Weil (1998), or to technology assimilation, as in Wang, Wong, and Yip (2017).

The main point to be addressed in this article is why the income disparities between fast-growing economies and development laggards have widened. More specifically, we want to understand the following: How important is the role played by institutional barriers with relation to technology adoption? To illustrate, we select a set of 10 fast-growing economies. This set includes Asian countries and African economies that are perceived as better performing. In contrast, we select a set of 10 development laggards. Beyond the typical candidates of countries mired in the poverty trap, this set includes countries with similar or even better initial states than some of the fast-growing countries, but with divergent paths of development leading to worse macroeconomic outcomes. That is, among development laggards, we choose two subgroups, one consisting of trapped economies and another of lag-behind countries. Over the past five decades, the ratios of average real GDP per capita between the 10 fast-growing economies and the five trapped economies have widened by almost nine times, whereas the comparable figures between the fast-growing economies and the five lag-behind countries have been almost four times.

Using cross-country analysis, we find that a key factor for fast-growing countries to grow faster than the United States and for trapped economies to grow slower than the United States is the relative TFP, which may be technology driven and not related to institutional barriers.

Yet more-severe institutional barriers in the lag-behind countries actually hinder the process of structural transformation and economic development, causing these countries to fall behind the fast-growing economies despite having similar or even better initial states five decades ago. Overall, we find that institutional barriers have played the most important role, accounting for more than half the economic growth in fast-growing and trapped economies and for more than 100 percent of the economic growth in the lag-behind countries. By conducting country studies, we identify that unnecessary protectionism, government misallocation, corruption, and financial instability have been key institutional barriers causing countries to either fall into the poverty trap or lag behind without a sustainable growth engine. Such barriers have created frictions or distortions to capital markets, trade, and industrialization, subsequently preventing these countries from advancing.

2 THE ORGANIZING FRAMEWORK

To illustrate, we use Acemoglu's (2009) version of Basu and Weil's (1998) as the organizing framework. Specifically, consider a representative firm in country j that employs capital K and labor L to manufacture a final product Y using a well-behaved aggregate production function that satisfies the constant-returns-to-scale property. Thus, country j's per capita real income at time t is given by

$$y_{j,t} = \frac{Y_{j,t}}{L_{j,t}} = \frac{A_{j,t}F\left(K_{j,t}, L_{j,t}\right)}{L_{j,t}} = A_{j,t}f\left(k_{j,t}\right),$$

where A measures TFP, $k = K/L$ is the capital-labor ratio, and $f(k) = F(k,1)$. Assume that F takes the Cobb-Douglas form with a capital income share $\alpha \in (0,1)$. We consider the United States as on the technology frontier, whose per capita real income is expressed as

$$y_{US,t} = A_{US,t}k_{US,t}^{\alpha}.$$

We follow the Basu-Weil-Acemoglu framework, assuming that the ability for country j to adopt production technology from a source country on the technology frontier, namely, the United States, depends on the extent of adoption barriers. Let technology adoption of a firm in country j take a simple form:

$$A_{j,t} = \tau_{j,t}A_{US,t}\min\left[1,\left(k_{j,t}/k_{US,t}\right)^{\zeta_j}\right],$$

where τ_j measures the relative TFP of country j to the United States, and $\zeta_j \in [0,1-\alpha]$ captures the foreign technology's degree of inappropriateness, which may be referred to as country j's barriers to adoption. When $\zeta_j = 0$, $\min[1,(k_{j,t}/k_{US,t})^{\zeta_j}]$ is the highest (equal to 1)—we call this case the absence of adoption barriers. In contrast, when $\zeta_j = 1-\alpha$, the adoption barriers are most severe.

We can then write country j's per capita real income relative to the United States, or in short, relative income, as follows:

$$\frac{y_{j,t}}{y_{US,t}} = \tau_{j,t} \min\left[1, \left(\frac{k_{j,t}}{k_{US,t}}\right)^{\zeta_j}\right]\left(\frac{k_{j,t}}{K_{US,t}}\right)^{\alpha}.$$

It is clear that when $\zeta_j = 0$, the expression reduces to the Lucas (2000) benchmark. Hence, the consideration of adoption barriers generalizes the Lucas benchmark, permitting better development and accounting for cross-country income disparities.

Notably, the adoption barrier parameter ζ_j may also capture other types of barriers that are not directly linked to technology. Examples of such include corruption, waste of resources, mismatched deficiencies, and capital market frictions. This parameter is country-specific, depending on a country's market environment and economic institutions. Throughout the remainder of the article, we shall refer to this as the institutional barrier parameter.

3 CROSS-COUNTRY ANALYSIS

To elaborate on this parameter without using the entire sample of countries, we select two sets of representative countries. The first set contains 10 fast-growing economies:

(i) four Asian Tigers known for their growth miracles—Hong Kong, Singapore, South Korea, and Taiwan;
(ii) two potential future Tigers of the Association for Southeast Asian Nations (ASEAN)—Malaysia and Thailand;
(iii) two emerging giants—China and India; and
(iv) two African miracles—Botswana and Mauritius.

Over the past five decades, these countries all experienced rapid, sustained growth. While most of them grew 2 percent or more during the entire period, India started late and Mauritius slowed down after the turn of the new millennium. While some performed better initially (e.g., Hong Kong and Singapore), others caught up quickly (e.g., China and India). These countries are typically singled out as development miracles.

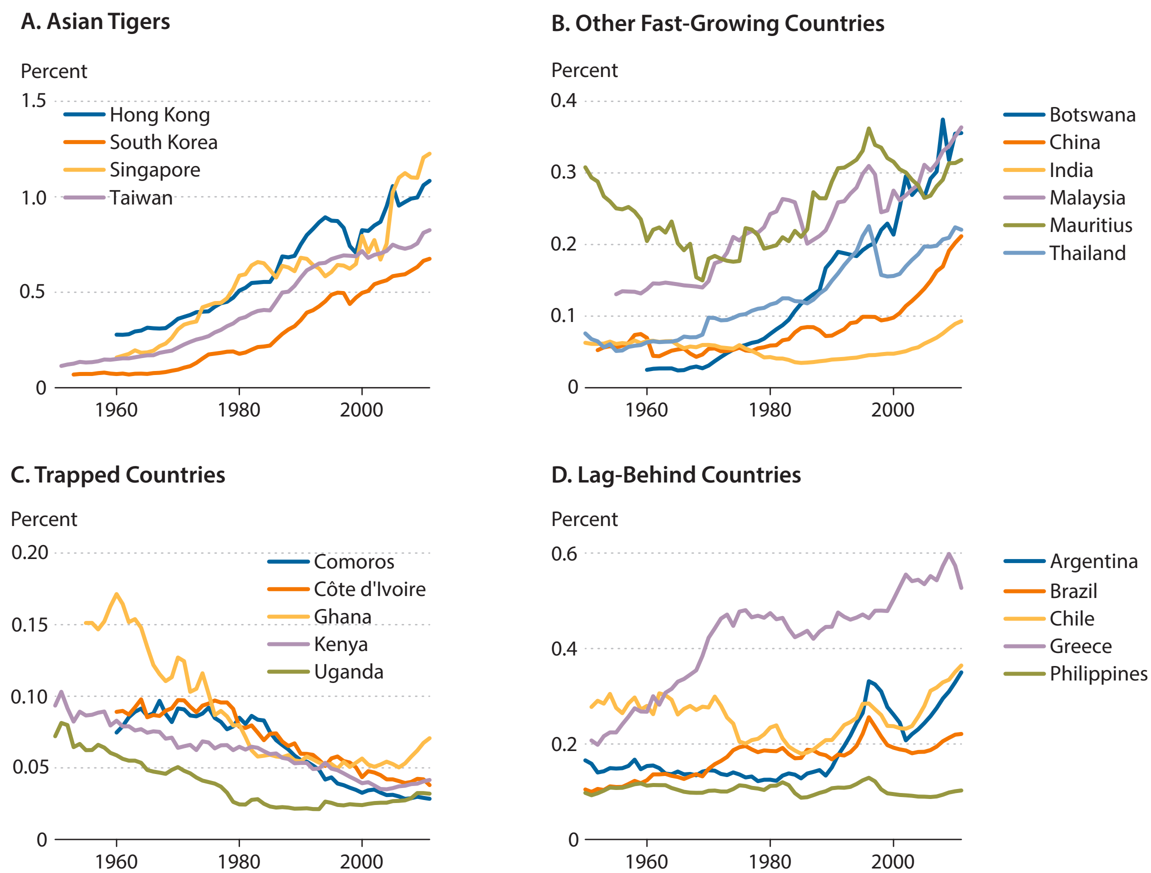
The second set includes 10 development laggards:

(i) five sub-Saharan African poor countries in poverty traps—Comoros, Côte d'Ivoire, Ghana, Kenya, and Uganda;
(ii) three Latin American countries falling behind newly industrialized Asian economies, despite better initial conditions—Argentina, Brazil, and Chile;
(iii) one lag-behind European country—Greece; and
(iv) one lag-behind Asian country—the Philippines.

All of the five trapped economies experienced negative growth and are frequently chosen in the literature when discussing poverty traps. The other five lag-behind countries' relative income growth is insignificant, between –1 and 1 percent. They are more or less trapped in the middle income level, around 20 to 30 percent of the U.S. income level since 1960. The selection of these countries is particularly interesting because they have similar or even better initial states than some of the fast-growing countries, but they take divergent paths of development resulting in less-desirable aggregate outcomes.

Figure 1

Relative Income of Asian Tigers, Other Fast-Growing, Trapped, and Lag-Behind Countries



SOURCE: Penn World Table 8.1 (in millions of 2005 U.S. dollars).

3.1 Data and Parameterization

For our quantitative analysis throughout (including Figure 1 and Tables 1-3), we use the Penn World Table 8.1 (PWT) for the period 1950-2011, for which real income measures are output-based and all relative income measures are with respect to the United States (see the Appendix for details). In Figure 1, we plot each country's relative income over the sample period. We divide the 20 countries into four panels: (A) the four Asian Tigers; (B) other fast-growing economies; (C) five sub-Saharan African poor countries in poverty traps; and (D) five lag-behind countries. One can see the miraculous development experienced by fast-growing economies during various episodes, as well as the economic miseries suffered by development laggards. One can also see that countries in Panels A and B experienced much faster growth over long episodes. In contrast, countries in Panel C showed a downward trend, whereas countries in Panel D had a moderate trend in relative income for a long time.

Our calibration exercise essentially summarizes the relative income of each country by the two parameters—the TFP ratio, τ_j, and the institutional barrier, ζ_j. Following Hall and Jones (1999), we set $\alpha = 1/3$. We calibrate τ_j and ζ_j to match the average level of relative income (in logs) and the average lag difference of relative income (in logs). The parameters are calibrated with the generalized method of moments whenever there does not exist any parameter within the parameter space to match the two targets.

3.2 Results

Our cross-country quantitative results are summarized in Table 2 (for fast-growing economies) and in Table 3 (for development laggards). One can see that the average growth rates of the fast-growing countries over the sample period of 1960-2011 is about 2.90 percent higher than that of the United States, much higher than that of the lag-behind countries (0.05 percent relative to the United States) and the trapped economies (–1.56 percent relative to the United States). Although the initial development status of these fast-growing economies (11.58 percent of the U.S. level) was ahead of the trapped ones (6.82 percent), it was below the other laggards (26.27 percent).

Simple development accounting analysis allows us to calibrate the institutional barrier parameters and the relative TFPs. For the fast-growing economies, the average institutional barrier parameter is about 0.4378, comparable with the figure for the trapped economies (0.4253) but noticeably lower than that for the lag-behind countries (0.4965). While the average relative TFP of the fast-growing economies (0.8769) is marginally higher than the figure for the lag-behind countries (0.8558), it is far above that for the trapped economies (0.6168).

One may therefore conclude that a key factor for fast-growing countries to grow faster than the United States and for trapped economies to grow slower than the United States is the relative TFP. More interestingly, more-severe institutional barriers in the lag-behind countries actually hinder the process of their structural transformation, causing these countries to fall below the fast-growing economies despite having similar or even better initial states five decades ago.

To complete the analysis, we perform standard growth accounting to compute the contribution of the institutional barriers to the long-term development of various countries compared with TFP advancement and capital deepening. We find that for all countries, institutional barriers have played the most important role: They account for more than half the economic growth in fast-growing and trapped economies—52.62 and 66.31 percent, respectively—and for more than 100 percent (101.70 percent) of the economic growth in the lag-behind countries.

While the overall message delivered above is clear-cut, the individual country's quantitative results are not entirely convincing. For example, the calibrated institutional barriers in Hong Kong and Taiwan may seem too high, whereas such barriers in China, Kenya, Uganda, and Brazil may seem too low. This is a typical problem of cross-country analysis. To further understand the role of institutional barriers, it is profitable to conduct more detailed country studies, to which we now turn.

Table 2

Cross-Country Quantitative Analysis—Fast-Growing Economies

	Average growth of relative income 1960-2011 (%)	Relative income in 1960 (%)	Development accounting: Relative TFP	Development accounting: Institutional barrier parameter	Growth accounting: Institutional barriers (%)	Growth accounting: Relative TFP advancement (%)	Growth accounting: Capital deepening (%)
Hong Kong	2.45	29.85	1.087	0.667	78.31	–17.47	39.16
Singapore	2.92	20.95	0.697	0.103	13.92	40.83	45.25
South Korea	4.06	10.64	0.901	0.391	57.61	–6.70	49.09
Taiwan	3.59	13.62	1.551	0.533	82.40	–33.96	51.56
Malaysia	2.21	10.73	1.067	0.667	102.46	–53.68	51.23
Thailand	2.70	5.10	0.405	0.296	36.33	22.71	40.95
China	4.95	2.67	0.406	0.389	38.81	27.91	33.28
India	1.11	4.85	1.544	0.667	92.94	–39.41	46.47
Botswana	4.03	3.09	0.244	0.000	0.00	68.16	31.84
Mauritius	0.97	14.31	0.867	0.667	23.44	64.84	11.72
Average	2.90	11.58	0.877	0.438	52.62	7.32	40.05

SOURCE: Penn World Table 8.1.

Table 3

Cross-Country Quantitative Analysis—Development Laggards

	Average growth of relative income 1960-2011 (%)	Relative income in 1960 (%)	Development accounting: Relative TFP	Development accounting: Institutional barrier parameter	Growth accounting: Institutional barriers (%)	Growth accounting: Relative TFP advancement (%)	Growth accounting: Capital deepening (%)
A. Trapped Economies							
Comoros	–2.01	4.34	0.596	0.667	61.65	7.53	30.82
Côte d'Ivoire	–2.76	8.26	1.420	0.667	74.95	–12.42	37.47
Ghana	–0.83	10.28	0.508	0.509	187.76	–210.81	123.05
Kenya	–1.32	6.46	0.132	0.000	0.00	52.03	47.97
Uganda	–0.85	4.73	0.428	0.285	7.18	84.42	8.40
Average	–1.56	6.82	0.617	0.425	66.31	–15.85	49.54
B. Other Laggards							
Argentina	–0.58	43.51	0.734	0.667	73.66	–10.48	36.83
Brazil	0.42	15.56	0.328	0.000	0.00	36.79	63.21
Chile	0.45	25.87	1.382	0.667	284.46	–326.69	142.23
Greece	0.61	34.71	1.113	0.667	185.22	–177.84	92.61
Philippines	–0.65	11.68	0.722	0.483	–34.85	158.93	–24.07
Average	0.05	26.27	0.856	0.497	101.70	–63.86	62.16

SOURCE: Penn World Table 8.1.

4 COUNTRY STUDIES

In what follows, we document for each country the key country-specific economic conditions and institutions that are relevant to our study. For general institutional background for each of the 20 countries, we refer the reader to the references listed in the Appendix.

4.1 Fast-Growing Economies

We begin by documenting some developmental and institutional details of the 10 fast-growing economies. For brevity, we focus on factors more likely to account for our quantitative findings.

Hong Kong. As shown in Figure 1A, Hong Kong's relative income (to the United States) rose sharply, except during the Asian financial crisis and the internet bubble periods (late 1990s and early 2000s). In the 1960s and 70s, Hong Kong's wider manufacturing industry successfully developed a reputation as a low-cost, labor-intensive original equipment manufacturing center by producing goods for export to Western countries. Morawetz (1981) argues that the mass production of Hong Kong's garment and general textile industry started with the American production lines built to produce supplies for the Korean War. Wan (2004) also provides cases for the electronics industry, but they are not that significant. Ever-increasing wages and land prices in the late 1970s seriously threatened the original equipment manufacturing strategy adopted by most manufacturing firms in Hong Kong. Since China's opening-up policy in 1978, many firms in Hong Kong have transformed themselves into service providers, from manufacturing to trading. A rapid transition to a service-based economy then took place in the 1980s, and Hong Kong further grew to become a financial center in the 1990s.

Singapore. The growth of Singapore's relative income is particularly impressive before the mid-1980s and after the internet bubble (see Figure 1A). By the turn of the century, the manufacturing and financial sectors contributed to almost half the country's GDP. The government set as high priority the high-tech industries, such as electronics, chemicals, and biotechnology. To achieve this goal, Singapore switched from labor-intensive to capital-intensive production by adopting a high-wage policy (compare with Otani and Sassanpour, 1988). But the exchange rate fluctuations in the early 1980s caused a serious recession that forced the government to set ceilings for nominal wages. Following the exchange rate depreciation in the mid-1980s, the economy started its recovery. Toward the end of the 1990s, Singapore became a world technology leader, particularly in its main electronics and biotech industries (compare with Lim and Lloyd, 1986).

South Korea. Figure 1A suggests South Korea's relative income rose steadily over the entire sample period. South Korea pursued a government-led, export-oriented growth strategy and shifted from import-substitution to export-oriented industrialization in the early 1960s. The Five Year Economic Development Plan of 1967-71 focused on shifting from primary exports to labor-intensive manufacturing sectors. Starting in 1970, the country shifted to promote heavy industries through the supply of cheap credit. In the 1980s, although the country had to deal with overinvestment or excess capacity negative growth, it stuck to its strategy of long-term investment in the high-tech exports, which contributed to the great takeoff in the following decade.

During its industry transformation, South Korea benefitted greatly from Japan, as documented by Kim (1997) and Wan (2004). Note that South Korean high school students were required to take Japanese language courses. Moreover, the country's business structures, inclusive of both large conglomerates (chaebols) and organized international trading companies (general trading corporations), resembled those of Japan, not to mention its significant technology transfers in the electronics industry (Sony to Samsung and Hitachi to LG) and the automotive industry (Honda to Hyundai).

With the onset of the Asian financial crisis in 1997, the financial sector liberalized and the proactive industrial policy to promote exports almost came to an end. However, with international lending and economic restructuring, South Korea recovered after only 18 months from the start of the crisis. Globalization of South Korea's industries since then reached a milestone in the nation's history, and the country joined the high-income group shortly after, becoming only the second in Asia to do so following the lead of Japan.

Taiwan. Similar to South Korea, Taiwan's relative income rose steadily over the entire sample period (see Figure 1A). During the 1960s and 70s, the Taiwan economy became industrialized and technology oriented. It was the second-fastest growing state in Asia after Japan. Starting in 1974, Chiang Ching-kuo implemented the Ten Major Construction Projects, the beginning foundations that helped Taiwan transform into an export-led growing economy. Exports became the main momentum to growth and industrialization and accumulated huge foreign reserves from a trade surplus.

Note that the Japanese rule before and during World War II brought changes in the public and private sectors, including rapid communications, good transportation systems throughout much of the island, and compulsory primary education. As documented by Kuo (1983) and Li (1988), Taiwan has benefitted from Japan, particularly in agricultural and textile technology. Moreover, Tatung was established in 1918, and it later became Taiwan's consumer electronics giant. The Sanyo to Sampo technology transfer also played an important role in broadening the range of electronics products.

However, the rapid development of Taiwan's modern IT industry in the past two decades was mostly influenced by the United States. As documented by Wan (2004) and Lee and Wang (2011), a series of key stepping stones helped Taiwan grow into an IT giant. In 1966, General Instrument built the largest foreign headquarters in Taipei County, and it was crucial to the development of machineries and tools. In 1970, RCA started its technology transfer to Taiwan. Robert Tsao was trained at RCA and later returned to Taiwan to establish the country's first major IT firm, the United Microelectronics Corporation (UMC), in 1980. Similarly, Texas Instruments started its technology transfer to Taiwan in 1970. Its senior vice president Morris Chang returned to Taiwan to establish the Taiwan Semiconductor Manufacturing Company (TSMC) in 1987. The semiconductor industry has formed a major part of Taiwan's IT industry. It overtook that of the United States, which was second only to Japan, in 2007. TSMC and UMC are the two largest contract chipmakers in the world, and MediaTek is the fourth-largest fabless supplier globally. Despite their success, small and medium-sized businesses still make up a significant portion of the businesses in Taiwan (i.e., they account for 85 percent of the industrial output), unlike in Japan and South Korea. The small and medium-sized enterprises rely on the importation of key components and advanced technology from the United States and Japan.

Malaysia. Since it became independent in 1957, Malaysia has grown remarkably in the region. It is the third-largest economy among countries in the ASEAN. During the second half of the past century, Malaysia grew an average of about 6.5 percent per annum, reaching a growth rate of 8 percent during its golden era, between 1980 and 1995 (Malaysia Yearbook, 2015). Its growth trend in relative income is plotted in Figure 1C. The major economic activities of the country are international trade and manufacturing. Malaysia is an exporter of natural and agricultural resources and is the largest Islamic banking center in the world.

Since the implementation of the controversial New Economic Policy in 1971, the predominantly mining and agricultural-based economy began a transition toward a more multi-sector industrial economy. The Mahathir Mohamad administration carried out many mega-projects in the 1980s that generated rapid economic growth and urbanization. Nevertheless, the financial markets were hit hard by the Asian financial crisis by the end of the past century. Although Malaysia's GDP declined dramatically by 7.5 percent in 1998, it rebounded by 5.6 percent in the following year (Malaysia Directory, 2008). The post Y2K slump of 2001 did not affect Malaysia as much as other countries.

Thailand. From Figure 1B, one can see that Thailand's relative income rose, except during the Asian financial crisis and internet bubble periods. Similar to the Asian Tigers, Thailand is a heavily export-dependent country, with exports accounting for more than two-thirds of its GDP. It completed industrial transformation two decades ago despite its relatively slow urbanization process.

The Srisdi regime, in power from 1957-73, introduced the market-oriented import-substitution industrialization that led to a period of rapid growth in 1958, with an average growth rate of 7 percent a year since then. From the 1970s to 1984, Thailand suffered from many economic problems, including reduced American investments, sizable current account deficits, high inflation due to two oil crises, unstable domestic politics, and unfriendly international politics. To deal with these economic problems, the Thai government devalued the baht three times in the early 1980s. As a result, beginning in 1987 the economy improved because net exports rose and foreign direct investment increased (especially from Japan); however, many economic problems persisted. For example, along with a huge current account deficit (averaged at 5.4 percent of GDP per year), a domestic capital shortage and rising external debt also occurred. All this led to the currency attack in 1997 that started the regional Asian financial crisis—the baht was forced to float starting July 2, 1997. Nevertheless, the economy began to recover in 1999 because of both internal and external factors. Internally, the government expanded its investment under the Thaksin Shinawatra administration. Externally, exports increased sharply because of both a weak baht and strong growth of the region.

China. China pursued an export-oriented growth strategy in the context of late industrialization toward the end of the 1970s, when East Asian early-starters felt the need to relocate and diversify their costly production bases. Among the late-industrializing nations, China has adopted a comprehensive export-led growth strategy, particularly since the 1992 Southern Tour announcement by Deng Xiao-ping. It provided the correct incentives from all government policies, including industry and trade policies, regulations (new laws and rules in all aspects of Chinese economic and trade reforms), administrative guidance and support (establishment of economic and technological development zones and directive measures to lead finance and investment to the key sectors), and a foreign direct investment (FDI) policy. As indicated in Figure 1B, China's relative income has risen sharply since then.

China's policies concurrently affect all sectors of exports, including primary goods, intermediate goods, and finished products. Its processing trade policy exempts imports for production to exports from tariffs and value-added tax, and it is a major support to exporters. Its exports through export and special economic zones have been encouraged, and its imports for research and development (R&D)-center establishment and operation are also exempt from tariffs and value-added tax. China's broad support for exports represents a departure from the promotion of selective sectors or the products approach of its neighbors in the 1970s. China took only two decades to make itself the number-one exporter of the world. The FDI share in China's exports increased from 1.94 percent in 1986 to 54.8 percent in 2003.

On average, during 1990-2004, China's annual real GDP growth rate was 10 percent, which remains one of the highest in the world. International trade makes up a major portion of China's GDP, and it exceeded $2.4 trillion at the end of 2008. However, aside from the external factors of trade and FDI, the Chinese economy experienced a series of internal structural transformations since the opening-up economic reform in 1979. These transformations include rural industrialization, reform of state-owned enterprises, modernization of the banking sector, and the recent proposal of fighting corruption.

India. As indicated by Figure 1B, India's growth came much later than China's. During 1950-90, India's per capita income grew at an average annual rate of only about 2 percent, a result due to the Indian government's implementation of restrictive trade, financial, and industrial policies. The Indian state took control of major heavy industries, by including additional licensing requirements, capacity restrictions, and limits on the regulatory framework. Following the foreign exchange crisis of 1957-58 (Sinha, 2004), trade policies shifted toward self-sufficiency, and the government gradually tightened control by increasing statutory liquidity and cash reserve requirements.

Despite all these regulations, the government invested heavily in R&D in the post-independence era. Many organizations were established to commercialize research outcomes. In terms of human capital accumulation, the government started different programs for developing engineers and scientists. The average R&D expenditure was 0.4 percent of GDP during 1950-90, but it surged to an average of 0.8 percent of GDP during 1991-2005, twice as much as the pre-reform period. In the late 1970s, the Indian government opened the economy by liberalizing both international trade and the capital market, leading to rapid growth in the early 1990s. As argued by Rodrik and Subramanian (2005), the trigger for India's economic growth was an attitudinal shift on the part of the national government in 1980 in favor of private businesses. Specifically, this shift is the pro-business attitude of Indira Gandhi, who returned to power in 1980 to garner political support from existing business groups. Such a change in the attitude of the national government toward the private sector created a positive atmosphere for investors' animal spirit.

The final trigger of the major economic reform of Manmohan Singh in the 1990s was due to the well-known 1991 balance-of-payment crisis. The economic reform that helped India's recovery was the condition of the emergency loan imposed by the International Monetary Fund (IMF) on the Narasimha Rao government. This reform ended the protectionist policies followed by previous Indian governments and started the liberalization of the economy toward a free-market system. This event led to an average annual growth rate that exceeded 6 percent in per capita terms during 1990-2005.

Botswana. During 1966-88, Botswana's GDP grew at an annual rate of 14.5 percent (The World Bank Profile, 2004). Its strong performance branded Botswana as the African Miracle. Rapid export growth followed the discovery of diamonds, with mining contributing 13 percent to GDP in 1975 to more than 50 percent by the end of the 1980s. Botswana then sought to diversify the economy by encouraging investment from abroad.

According to Sachs and Warner (1997), reform took place in 1979 that led to the country being open to international markets. During 1997-2002, its average GDP growth rate decreased to 5.9 percent. Moreover, AIDS has been a serious problem for Botswana: In 2002, HIV prevalence in Botswana was 39 percent of the people 15-49 years of age (The World Bank Profile, 2004), one of the highest in the world. This has not only raised mortality but also reduced labor productivity.

Mauritius. Another miraculous performer in Africa is Mauritius, whose success is due to its early reform that took place in 1968. The reform instituted a spectacular economic transformation into tourism and out of sugar production. Consequently, agriculture in GDP composition dropped from 24 percent in 1970 to 5 percent in 2007, and services increased from 56 to 74 percent in the same period. During 1970-99, the average annual growth of export volumes was 5 percent. It began at only 1 percent in the 1970s, increased to about 10 percent in the 1980s, and then was 5 percent (mode 12) in the 1990s. Given its economic focus on tourism, the technological assimilation of Mauritius is not expected to be important, as is confirmed by its low value of the assimilation parameter at 0.0037. As shown in Figure 1B, the growth trend in relative income is most noticeable during 1968-96.

4.2 Development Laggards

We next turn to documentation of developmental and institutional details of the five trapped and five lag-behind countries.

Comoros. Income disparity between Comoros and the United States has widened, as shown in Figure 1C. Comoros was colonized by France in 1841 and was transformed into a plantation-based economy. It declared independence in 1975. However, Comoros has experienced political interruptions since then, with its on-and-off independence from France.

Agriculture dominates the production activity of Comoros, and it features three main cash crops: ylang-ylang (for perfume), vanilla, and coconut. The country's high total fertility rates created serious population pressure because of the people's fear of hunger. Not surprisingly, its education level was low as well. Moreover, the government deficits were large because the country lacked stable tax revenues. In the past few decades, the economy of Comoros mainly relied on foreign financial and technical assistance.

Côte d'Ivoire. Real GDP per capita of Côte d'Ivoire rose in the 1960s-70s but fell afterward (see Figure 1C). Its real GDP per capita in 2003 was about the same as that in the early 1960s. Underpinning the per capita output decline since the late 1970s are the following: rapid population growth (3.95 percent for the first period of 1960-79 and 3.26 percent for the second period) and a decrease in capital investment. (The growth rate of gross capital formation per capita was 5.61 percent for the first period and –2.90 percent for the second period.)

Côte d'Ivoire has become dependent on raw commodities, particularly cash crops such as cocoa. Over the past 40 years, cocoa output has grown to dominate Côte d'Ivoire's economy and world production. Its cocoa output was 1.4 million tons in 2001, equivalent to about 40 percent of the world's output. The nominal and real cocoa prices in the world market increased until they hit an all-time high in 1977. Then the prices collapsed and never fully recovered. And cocoa exports dominate Côte d'Ivoire's trade and economy. By 2000, raw cocoa represented 80 percent of the country's commodity exports, more than 50 percent of all exported goods and services and 21 percent of GDP. The structural problem of the Côte d'Ivoire economy was caused by its dependence on cocoa production. Cocoa production is not a very profitable business because there is a long period of investment and no alternative use of the land. The life cycle of cocoa is nearly four decades while it takes seven years to finish just the planting process. In addition, the country's terms of trade have deteriorated seriously since the late 1970s because of the fall in international cocoa prices and the fluctuations in foreign exchange rates. In short, the gamble of Côte d'Ivoire on cocoa to finance development efforts failed.

The support of foreign aid (from IMF and World Bank) is crucial to the economy. Note that Côte d'Ivoire is one of the most corrupt countries in the world (compare with Easterly, 2001).

Ghana. As shown in Figure 1C, income disparity between Ghana and the United States had widened until very recently because the 2008 financial tsunami did not harm Ghana's economy as much as it did the rest of the world. Attaining its political independence in 1957, the Ghana government adopted a fast-track strategy by launching state-owned import substitution industries in the 1960s. In 1970, Ghana had one of the most diverse and dynamic manufacturing sectors in sub-Saharan Africa. To understand the surge that occurred until the early 1970s, we find that, by ignoring the outlier of 1968, a clear negative correlation exists between the government spending-GDP ratio and per capita GDP. This finding can be understood by Prescott's (1998) suggestion that, for developing countries, government size can be used as a measure of distortions. However, as Ghana pursued non-selective industrialization policies behind high barriers of protection, its industry failed to develop adequate industrial capabilities and infrastructure. The Ghanaian industrial structure is typical of a low-level industrialization, showing a natural evolution from traditional to simple processing and assembly activities. Consequently, growth rates slowed down and, with declining revenues from the primary exports that had financed industry, even turned negative. Thus, structural reform policies were undertaken in the 1980s. Growth resumed by the end of the 1990s. Unlike the formal technological effort in Ghanaian manufacturing, practically all industrial R&D in Ghana was conducted by public institutions rather than by enterprises. The R&D effort declined sharply in the 1980s, and it was well below the critical mass needed to make a significant contribution to the absorption, adaptation, or creation of technology (Lall, 1990).

Kenya. Income disparity between Kenya and the United States widened over the entire sample period (see Figure 1C). Following Kenya's emergence as an independent nation in 1963, the years were marked by rapid growth of domestic products, strong fiscal performance, low rates of inflation, and manageable external accounts. Since the early 1970s, growth has slowed because of two oil crises, as Kenya is heavily dependent on imported petroleum. Other international economic events occurred as well, such as the collapse of the East African Community, the major market for Kenyan manufactured goods. International recession and high international interest rates resulted in a steady tightening of the foreign exchange constraint on growth. Dramatic swings in the prices of Kenya's key exports—coffee and tea—compounded the already difficult problems of economic management. Real prices for raw cocoa, cotton, and coffee increased in world commodity markets in 1960 until they hit all-time highs in 1977. After 1977, real prices crashed, and on average they have declined steadily ever since. During 1977-2003, coffee production did not increase, and real prices declined. Other factors contributing to the poor growth performance of Kenya include rapid population growth, an AIDS outbreak, poor infrastructure, and the extended and recurrent banking crisis since the mid-1980s.

Uganda. Figure 1C indicates that income disparity between Uganda and the United States widened before 1980; afterward, the gap remained largely unchanged. Uganda became independent from the United Kingdom in 1962. At that time, more than 80 percent of households lived in rural areas, with the majority earning a living from quasi-subsistence agriculture. Unfortunately, this independent nation suffered a sharp economic downturn in the 1970s and 80s. The prolonged recession was due to the expulsion of South Asians in 1972, the administration of Idi Amin (the so-called State of Blood), the outbreak of the Uganda-Tanzania War in 1978-79, and the Bush War by the National Resistance Army in 1981-86.

In 1986, the government (with the support of foreign countries and international agencies) attempted to rehabilitate the economy devastated during the regime of Idi Amin and the subsequent civil war. However, the economy was still in very poor condition. Its agriculture sector accounted for 56 percent of real GDP, with coffee as its main export. (Coffee's price fell sharply in the late 1970s.) The reform in 1988 helped the agriculture-based export expansion, causing an annual growth rate of 14 percent. Inflation was at 240 percent in 1987 and 42 percent in June 1992.

Argentina. As shown in Figure 1D, Argentina's relative income fell slightly until 1990; since then there have been some upward, though volatile, movements. Argentina is the third-largest economy in Latin America and was one of the richest countries in the world in the early twentieth century. However, after the Great Depression, import substitution generated a cost-push effect of high wages on inflation. During 1975-90, growing government spending, large wage increases, and inefficient production created chronic inflation that increased until the 1980s, and real per capita income fell by more than 20 percent. The annual rate of inflation never fell below 100 percent during this period. Starting with the *Rodrigazo* (the group of economic policies announced in Argentina on June 4, 1975), inflation rose sharply, reaching an average of more than 300 percent per year during 1975-91, and prices increased by a factor of 20 billion pesos. Foreign debt also reached a level that was equal to three quarters of gross national product. Record foreign debt interest payments, tax evasion, and capital flight resulted in a balance-of-payments crisis that plagued Argentina with severe stagflation for this period.

In 1991, the government attempted to control inflation by pegging the peso to the U.S. dollar. In addition, it began to privatize state-run enterprises on a broader basis and stop the run of government debt. Unfortunately, lacking a full commitment, the economy continued to crumble slowly and eventually collapsed in 2001 when the Argentine government defaulted on its debt. Its GDP declined by nearly 20 percent in four years, unemployment reached 25 percent, and the peso depreciated by 70 percent after being devalued and floated. Since then, the economy has started to recover.

Brazil. Figure 1D suggests that Brazil gained ground by catching up with the United States before the mid-1970s, but since then it has been unable to move up further. Brazil is the largest national economy in Latin America and is famous for its corruption (ranking 69th among 178 countries in 2012 in the Transparency International's Corruption Perceptions Index). Corruption alone costs Brazil an estimated $41 billion a year.

Since the end of the Second World War, the Brazilian economy started intense import substitution, transformed from a largely agricultural to an industrial society, and grew more than 7 percent during 1951-60. Yet, without setting up the institutions for export growth, the economy ran a large current account deficit. With the industrialization process continuing over the following decade, export expansion led to double-digit growth.

After the first oil shock in the early 1970s, Brazil continued its high-growth policy by running up foreign debt because of increasing import requirements of industrialization. Another feature of the post-1973 period is the acceleration of inflation. During 1973-91, Brazil experienced high inflation; its annual rate was 30 to 40 percent in the 1970s and accelerated in the 1980s. Its government deficit (corrected for inflation) fell from an average of 13 percent of GDP in the early 1980s (peaking at 22.1 percent in 1983) to 3.3 percent in 1987. Capital continued to flow out of the country for more than a decade, during 1982-91, and was estimated at more than $10 trillion per year for the second half of the 1980s (Edwards, 1998). Its inflation peaked as follows: 1986:02 (20 percent), 1987:06 (24 percent), 1989:01 (33 percent), and 1990:03 (59 percent). The 1986 Cruzado Plan and the 1990 Collor I Plan helped only temporarily. With the 1991 Collor II Plan still lacking success, Brazil continued to suffer crises, including those in 1999 and 2002; although, these crises were considerably less severe than the Argentine crisis in 2001-02.

Chile. Chile's relative income did not rise but fell during 1970-2001 (see Figure 1D). From the Second World War to 1970, real GDP per capita of Chile increased at an average annual rate of 1.6 percent, and its economic performance was behind those of Latin America's large and medium-sized countries. Chile pursued an import-substitution strategy, which resulted in an acute overvaluation of its currency that intensified inflation. The government carried out several anti-inflation programs but failed, and inflation continued to be a serious problem of Chile's in the 1970s. However, since its adoption of its inflation target in 1991, the Chilean economy has stabilized.

Although most Latin American countries have practiced strong government intervention in the markets since the mid-1970s, Chile pursued free market reform. A group of Chicago economists—the so-called Chicago boys—were brought in for the formulation and implementation of Chile's reforms. The outcomes are as follows: Exports grew rapidly, per capita income took off, inflation declined to single digits, wages increased substantially, and the incidence of poverty plummeted (compare with Edwards and Edwards, 1991). Since the democratic administration of Patricio Aylwin in 1990, the economic reform has been accelerated and Chile has become one of the healthiest economies in Latin America.

Greece. As indicated in Figure 1D, Greece's relative income rose before 1970; since then it has been largely flat. After World War II and the civil war between communist and anti-communist forces in the 1940s, the Greek economy enjoyed rapid growth that was propelled in part by the Marshall Plan. During the 1960s-70s, the country went through a political transition, known as the Metapolitefsi.

Greece became a member of both NATO (1980) and the European Communities (1981). This facilitated foreign investments in industrial enterprises and heavy infrastructure. The rapid expansion of the service sector (mainly due to tourism) further contributed to the fast-growing performance of the economy. In the first decade of this century, the service sector made up 85 percent of the national economic output. In addition, the country adopted the euro in 2001 and successfully hosted the 2004 Summer Olympic Games in Athens. However, by the end of 2009, Greece suffered greatly from the financial crisis that began in the United States and that has been central to the related European sovereign debt crisis recently.

Philippines. The Philippines was one of the wealthiest countries of the region in the post-war period. However, as shown in Figure 1D, its relative income has been flat over the entire sample period since then. Since the Marcos administration in the 1960s, the Philippines was overtaken by other Asian economies. The People Power Revolution of 1986 resulted in Marcos's exile to Hawaii, and Corazón Aquino became the president. However, during her administration, the economy suffered from national debt, government corruption, coup attempts, and natural disasters. When Fidel V. Ramos replaced Aquino in 1992 and liberalized the economy, there was a short recovery, which was interrupted by the 1997 Asian financial crisis.

Similar to its neighbors in the region, the Philippines suffered a fall in both the stock market and its currency value during the Asian financial crisis. Fortunately, the government was fiscally conservative because of the supervision of the IMF, and the economy has recovered since the millennium. During Gloria Macapagal-Arroyo's 9-year administration beginning in

2001, the economy experienced GDP growth from 4 percent in 2002 to 7 percent in 2007 with the completion of infrastructure projects, such as the Manila Light Rail Transit System Line 2 in 2004, and it managed to avoid the Great Recession caused by the financial crisis of 2007-08.

4.3 Summary and Comparison

By reviewing the previous country-specific details, one can see that the 10 fast-growing countries have all adopted an open policy with an export-led development strategy. For instance, the 1967-71 Five-Year Economic Development Plan of South Korea focused on shifting from primary exports to manufacturing sectors. In 1974, Chiang Ching-kuo implemented the Ten Major Construction Projects that helped Taiwan transform into a growing, export-led economy. These countries have set up the environment to promote FDI. Their governments have undertaken serious reforms, particularly in both labor and financial markets. Most of them have also implemented various effective industry polices suitable to their competitive edges. Examples include the pro-market reform of China since the 1992 Southern Tour announcement by Deng Xiao-ping. In particular, the Chinese government has instituted a number of ownership, trade, and hukou (household registration) reforms and established economic and technological development zones to attract FDI. In the 1980s, Singapore set as high priority the high-tech industries (such as electronics, chemicals, and biotechnology) and pursued a high-wage policy during 1979-84 to shift production away from labor-intensive to both capital-intensive and high-tech activities.

In contrast, those African economies mired in the poverty trap have adopted import substitution policies that attracted foreign investment through protected markets. Many have financed public investment in heavy industry, which allocated lots of resources to only a thin market. To private industries, strong protectionism has stimulated domestic investment but retarded productivity and competitiveness. There has been a lack of both export diversification and production sophistication. Moreover, corruption has been severe while government allocation has been perceived as inefficient. As a result, these economies remain primarily agricultural societies. For instance, agriculture dominates the production activity of Comoros, focusing on only three main cash crops: ylang-ylang (for perfume), vanilla, and coconut. Over the past 40 years, cocoa output has grown to dominate Côte d'Ivoire's economy and world production. And GDP of both Kenya and Uganda have relied heavily on coffee production.

While the Latin American countries were at much better stages of development even relative to most of the fast-growing ones, they also have employed an import substitution strategy based on protectionism and government-heavy industrialization as in the trapped African economies. Since the oil crises in 1973 and 1979, they have all accumulated external debt at an unsustainable pace. Undisciplined fiscal expansion together with loose monetary policy have led to various high-inflation episodes, causing instability and uncertainty. Corruption and inefficient government allocation have caused resource waste. For instance, after the Great Depression, Argentina pursued a strategy of import substitution to achieve industrialization, but it eventually led to the nightmare of cost-push inflation. Similarly, the import substitution industrialization of Brazil also resulted in a substantial increase in imports and,

Table 4

Summary of Development Drivers from Country Studies

Development-enhancing factors	Development-retarding factors
(1) Export-led open policy	(1) Unnecessary protectionism
(2) FDI incentives	(2) Government misallocation
(3) Business infrastructure	(3) Corruption
(4) High-tech promotion	(4) Financial instability
(5) Pro-market labor institutions	
(6) Pro-market financial institutions	

SOURCE: Penn World Table 8.1.

hence, a large foreign debt, so the country's balance-of-payments problems increased dramatically. The remaining two development laggards, Greece and the Philippines, have experienced paths similar to the three Latin American countries, except that they have had more moderate inflation.

We summarize our findings in Table 4. On the one hand, we identify that export-led open policy, FDI incentives, solid infrastructure for business, and promotion of high-tech industries, together with other pro-market reforms in labor and financial markets, have helped fast-growing countries outgrow the United States. On the other hand, we see that unnecessary protectionism, government misallocation, corruption, and financial instability have caused the trapped and the lag-behind countries to be unable to develop along a sustainable growth path. These institutional factors can be barriers to capital markets, trade, and industrialization, which prevent the development laggards from catching up with the Joneses.

5 CONCLUSION

In this article, we established quantitatively that while relative TFP is a key factor for fast-growing countries to rise and for trapped economies to fall in the development process, more-severe institutional barriers are important drivers in the laggards falling behind and the poor continuing to be mired in the development trap. We also illustrated that institutional barriers have played the most important role, accounting for more than half the economic growth in fast-growing and trapped economies and for more than 100 percent of the economic growth in the lag-behind countries. We further identified that unnecessary protectionism, government misallocation, corruption, and financial instability have caused this second set of countries to either fall into the poverty trap or lag behind without a sustainable growth engine. Thus, the establishment of correct institutions and individual incentives for better access to capital markets, international trade, and industrialization can be viewed as crucial for a country to advance with sustained economic growth. ■

APPENDIX

In this appendix, we provide detailed documentation for data and institutional background used herein.

Data

Our data are based on the PWT for the period 1960-2011. Real income measures are output-based real GDP ($rgdp^o$), using prices for final goods, exports, and imports in millions of 2005 U.S. dollars that are constant across countries and over time. The capital stocks are measured in the PWT by rk^{NA}, at constant national prices in millions of 2005 U.S. dollars based on investment and prices of structures and equipment. All relative income and relative factor endowment measures are relative to the comparable incomes and endowment measures in the United States.

Documentation for Country-Specific Institutional Background

General institutional background for each of the 20 countries is based on the references listed below. In cases where our citations are more substantive, we also include the references in the main text.

- Hong Kong: Morawetz (1981); Wan (2004)
- Singapore: Lim and Lloyd (1986); Otani and Sassanpour (1988)
- South Korea: Kim (1997); Wan (2004)
- Taiwan: Kuo (1983); Li (1988); Wan (2004); Lee and Wang (2011)
- Malaysia: Malaysia Directory, (2008); Malaysia Yearbook, (2015)
- Thailand: Xu and Yu (2001)
- China: Naughton (2007)
- India: Sinha (2004); Rodrik and Subramanian (2005)
- Botswana: The World Bank Profile #34145 (2004); Sachs and Warner (1997)
- Mauritius: Svirydzenka and Petri (2014)
- Comoros: The World Factbook (2012)
- Côte d'Ivoire: Easterly (2001)
- Ghana: Lall (1990); Prescott (1998)
- Kenya: The World Factbook (2012); The World Bank (2012)
- Uganda: Byrnes (1992)
- Argentina: Rogers and Wang (1993); Donghi (2011)
- Brazil: Rogers and Wang (1993); Edwards (1998); Ferraz and Finan (2008)
- Chile: Edwards and Edwards (1991)
- Greece: Allison and Nicholaidis (1997)
- Philippines: Canlas, Khan, and Zhuang (2011); Hutchcroft (1999)

REFERENCES

Acemoglu, D. *Introduction to Modern Economic Growth*. Princeton University Press, 2009.

Aghion, P.; Howitt, P. and Mayer-Foulkes, D. "The Effect of Financial Development on Convergence: Theory and Evidence." *Quarterly Journal of Economics*, 2005, *120*, pp. 173-222.

Allison, G. T. and Nicholaidis, K. *The Greek Paradox: Promise Vs. Performance*. Cambridge, MA: MIT Press, 1997.

Basu, S. and Weil, D. "Appropriate Technology and Growth." *Quarterly Journal of Economics*, 1998, *113*, pp. 1025-1054; https://doi.org/10.1162/003355398555829.

Buera, F. J. and Shin, Y. "Financial Frictions and the Persistence of History: A Quantitative Exploration." *Journal of Political Economy*, 2013, *121*, pp. 221-272; https://doi.org/10.1086/670271.

Byrnes, R. M. *Uganda: A Country Study*. Washington, D.C.: Federal Research Division and Thomas Leiper Kane Collection, Library of Congress, 1992.

Canlas, D. B.; Khan, M. E. and Zhuang, J. *Diagnosing the Philippine Economy: Toward Inclusive Growth*. Anthem Press, 2011.

Donghi, T. H. *Argentina-Economy*. December 14, 2011; http://www.webcitation.org/63w7tsNNw.

Easterly, W. *The Elusive Quest for Growth*. Cambridge, MA: MIT Press, 2001.

Edwards, S. "Capital Flows, Real Exchange Rates, and Capital Controls: Some Latin American Experiences." NBER WP# 6800, 1998.

Edwards, S. and Edwards, A. C. *Monetarism and Liberalization: the Chilean Experiment*. Chicago, IL: University of Chicago Press, 1991.

Ferraz, C. and Finan, F. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics*, 2008, *123*, pp. 703-745; https://doi.org/10.1162/qjec.2008.123.2.703.

Hall, R. and Jones, C. "Why Do Some Countries Produce So Much More Output per Worker than Others?" *Quarterly Journal of Economics*, 1999, *114*, pp. 83-116; https://doi.org/10.1162/003355399555954.

Hutchcroft, P. "Neither Dynamo nor Domino: Reforms and Crises in the Philippine Political Economy," in T. J. Pempel, ed., *The Politics of the Asian Economic Crisis*. Cornel University Press, 1999.

Kim, L. *Imitation to Innovation: The Dynamics of Korea's Technological Learning*. Harvard Business School Press, 1997.

Kuo, S. W.-Y. *The Taiwan Economy in Transition*. Westview Press, 1983.

Lall, S. *Building Industrial Competitiveness in Developing Countries*. Paris OECD Development Centre, 1990.

Lee, W.-C. and Wang, P. *Taiwan Economic Development: From the Past to the Future* (in Chinese). Taipei, Taiwan: Wunan Publishing, 2011.

Li, K. T. *The Evolution of Policy Behind Taiwan's Development Success*. Yale University Press, 1988.

Lim, C. Y. and Lloyd, P. *Singapore: Resources and Growth*. Oxford, England: Oxford University Press, 1986.

Lucas, R. E. Jr. "Why Doesn't Capital Flow from Rich to Poor Countries?" *American Economic Review*, 1990, *80*, pp. 92-96.

Lucas, R. E. Jr. "Some Macroeconomics for the 21st Century." *Journal of Economic Perspectives*, 2000, *14*, pp. 159-168; https://doi.org/10.1257/jep.14.1.159.

Malaysia Industrial and Business Directory. *International Business Publications*, 2008.

Malaysia Recent Economic and Political Developments Yearbook. "Strategic Information and Developments." *International Business Publications*, 2015.

Morawetz. *Why the Emperor's New Clothes Are Not Made in Colombia: A Case Study in Latin American and East Asian Manufactured Exports*. Oxford, England: Oxford University Press, 1981.

Naughton, B. *The Chinese Economy: Transitions and Growth*. Cambridge, MA: MIT Press, 2007.

Otani, I. and Sassanpour, C. "Financial, Exchange Rate, and Wage Policies in Singapore, 1979-86." *IMF Staff Papers*, 1988, *35*(3), pp. 474-95; https://doi.org/10.2307/3867183.

Prescott, E. "Needed: A Theory of Total Factor Productivity." *International Economic Review*, 1998, *39*, pp. 525-553; https://doi.org/10.2307/2527389.

Rodrik, D. and Subramanian, A. "From Hindu Growth to Productivity Surge: The Mystery of the Indian Growth Transition." *IMF Staff Papers*, 2005, *52*, pp. 193-228.

Rogers, J. and Wang, P. "High Inflation: Causes and Consequences." Federal Reserve Bank of Dallas *Economic Review*, 1993, pp. 19-33.

Sachs, J. D. and Warner, A. M. "Sources of Slow Growth in African Economies." *Journal of African Economies*, 1997, *6*, pp. 335-76; https://doi.org/10.1093/oxfordjournals.jae.a020932.

Sinha, J. B. P. *Multinationals in India: Managing the Interface of Cultures*. New Delhi, India: Sage Publications, 2004.

Svirydzenka, K. and Petri, M. "Mauritius: The Drivers of Growth—Can the Past be Extended?" IMF Working Paper WP/14/134, 2014.

Wan, H. *Economic Development in a Globalized Environment: East Asian Evidence*. Chaps. 4, 7, 8. Kluwer Academic Publishers, 2004; https://doi.org/10.1007/978-1-4419-8941-3.

Wang, Y. and Wang, P. *Barriers to Health and Poverty Trap*. Unpublished manuscript. The Chinese University of Hong Kong, 2015.

Wang, P.; Wong, T. N. and Yip, C. K. T*echnology Assimilation and Development Accounting*. Unpublished manuscript. Washington University in St. Louis, 2017.

The World Bank Profile. "Botswana—Towards Prosperity for All: A Comprehensive Development Framework Profile." #34145, 2004.

The World Bank. "Kenya Economic Update." June 2012; http://www.worldbank.org/content/dam/Worldbank/document/Africa/Kenya/kenya-economic-update-june-2012.pdf.

The World Factbook. Central Intelligence Agency. 2012; https://theodora.com/wfbcurrent/comoros/index.html.

Xu, D. and Yu, Z. "From Crisis to Recovery: East Asia Rising Again?" *World Scientific*, 2001.

Intergenerational Mobility and the Effects of Parental Education, Time Investment, and Income on Children's Educational Attainment

George-Levi Gayle, Limor Golan, and Mehmet A. Soytas

This article analyzes the mechanisms through which parents' and children's education are linked. It estimates the causal effect of parental education, parental time with children, and parental income during early childhood on the educational outcomes of children. Estimating the causal effects of time with children, income, and parental education is challenging because parental time with children is usually unavailable in many datasets and because of the problem of endogeneity of parental income, time with children, and education. The authors, therefore, use an instrumental variables approach to estimate the causal effects. They find that once they account for the parental time input with children, parental income during the first five years is no longer statistically significant. The parental time investments of both parents in early childhood are each statistically and quantitatively significant determinants of the educational outcomes of children. (JEL C13, J13, J22, J62)

Federal Reserve Bank of St. Louis *Review*, Third Quarter 2018, *100*(3), pp. 281-95.
https://doi.org/10.20955/r.2018.281-95

1 INTRODUCTION

Income inequality has increased substantially in the United States since the 1970s. One concern about higher income inequality is that it is linked to intergenerational mobility. Intergenerational mobility refers to the change in socioeconomic outcomes from one generation to the next—from parents to their children as adults—and usually is measured by the intergenerational correlation in income, education, or social class. A strong link between the incomes and educations of parents and children across generations is a sign of low intergenerational income mobility and likely means that children born to less-advantaged households will be less advantaged as adults. Economies with higher inequality tend to have lower intergenerational mobility (a phenomenon known as the "Great Gatsby" curve). Since education is an important determinant of earnings and many other life outcomes, we focus on inter-

George-Levi Gayle and Limor Golan are associate professors at Washington University in St. Louis and research fellows at the Federal Reserve Bank of St. Louis. Mehmet A. Soytas is an assistant professor at Ozyegin University.

generational transmission of education, specifically, the mechanisms that link children's education to their parents' education. It is well documented that highly educated parents are more likely to have highly educated children. Since many skills are formed before individuals enter college, we seek to understand early determinants that affect economic mobility across generations. In particular, we focus on the impact of parental monetary investment and the time parents spend with their children when their children are young on the children's completed education.

There are many possible underlying mechanisms for the intergenerational correlation in education, and it is difficult to determine causal effects of the different determinants (see Black and Devereux, 2011, for a survey of the literature). First, more-educated parents often have higher incomes, which may affect the educational attainment of their children. However, more-educated parents may have more skills and abilities that are directly transmitted to their children. Therefore, the intergenerational correlation in education observed in the data may reflect the fact that more-able parents get more education and have more-able children who get more education. This is known as the "ability bias." Lastly, more-educated parents tend to spend more time with their children. By time, we mean parental time allocated to the care of their children, which is an essential input in the development of the children's skills, or human capital. Parental time with children, then, captures the investment in human-capital-enhancing activities for the children. However, it is possible that parents with higher ability tend to spend more time with their children, which introduces an endogeneity problem. This problem arises since it is hard to distinguish whether parental time investment itself enhances the human-capital development of children or whether it simply reflects the fact that more-educated parents tend to spend more time with their children and that their parental skills and education affect the development of the children's skills.

Moreover, both income and parental time with children depend on how households choose to allocate their time. That is, college-educated mothers married to college-educated men may spend different amounts of time with their children and in the labor market than college-educated mothers married to men with only a high school education. Since there is typically assortative mating—that is, people tend to marry people with similar education—it is important to account directly for the characteristics and inputs of both parents. These issues, in addition to the fact that there is a lack of data on parental time investment, pose difficulties for estimation of the causal effect of each of the above factors on children's educational outcomes.

In this article, we estimate the causal effects of parental education and income and time spent with children in the first five years of life on a child's educational outcome. It is based on Gayle, Golan, and Soytas (2015). There is an extensive literature on estimation of the production function of skills (see Heckman and Mosso, 2014, for a comprehensive survey); however, this literature does not identify the causal impact of parental time on completed education (typically due to lack of data on parental time investment). An exception is Del Boca, Flinn, and Wiswall (2014), who estimate a detailed process of skill formation of young children using data on parental time with children to identify the causal effect of parental time. However, they measure the effect on test scores and not on completed education. We first show that fathers' income in the first five years of their children's lives has a positive effect on the educational outcomes of their children. This outcome still holds once we account for endogeneity by using instrumental variables techniques. However, once we include parental time measures, the income variables are no longer statistically significant. Measures of parental time are not typically accounted for in the literature since data on them are not typically available. We find, however, that parental education is still important. Parental time of both mothers and fathers each has a significant impact on children's educational attainment. Our results, therefore, do not support the idea that household credit constraints, at least when children are young, have a direct effect on children's educational outcomes. However, credit constraints could have an indirect effect if they reduce the amount of time parents spend with their children.

This article is organized as follows: Section 2 describes the data and documents the relationship between parental education and income and children's educational outcomes. Section 3 estimates the causal effect of parental time and income on children's educational outcomes. Section 4 provides discussion and concludes.

2 EDUCATIONAL OUTCOMES, PARENTAL EDUCATION, AND INCOME

We begin by documenting the correlation between children's educational outcomes and the socioeconomic status of their parents, using Panel Study of Income Dynamics (PSID) data of two generations of parents and children. The PSID is the longest-running longitudinal household survey in the world. The survey started in 1968 with a nationally representative sample of 5,000 U.S. families and includes continuously collected information on employment, income, wealth, expenditures, health, marriage, childbearing, child development, education, and various other aspects of the family members and their descendants. The PSID is directed by the University of Michigan, and the data are freely available to researchers.

We select individuals from 1968 to 1997 by setting the individual level variables "Relationship to Head" to head, or wife, or son, or daughter. Our main sample contains 423,631 individual-year observations. Only white and black individuals between 17 and 55 years of age are kept in our sample. In the data, we observe the number of children, annual labor income, labor market hours, housework hours, and parental time with children. Table 1 presents the summary statistics of the main variables used in the estimation.

The PSID measures annual hours of housework for each individual; however, it does not provide data on the time parents spend on childcare.[1] The variable for time spent with children is estimated using a version of the approach used in the literature. Hours with children are computed as the deviation of housework hours in a particular year from the average housework hours of individuals with no children, by gender, education, and year (Hill and Stafford, 1974, 1980; Leibowitz, 1977; and Datcher-Loury, 1988). Negative values are set to zero, and childcare hours for individuals with no children are also set to zero. In addition, in the estimation and in the analysis, we do not use the hours with children measure directly; instead, we compute a discrete version of this measure with three levels of time spent with children (low, medium, and high) for fathers and mothers separately, which reduces the concerns about the representativeness of the measure.[2]

Table 1

Summary Statistics

	N	Mean
Female	2,693	0.476 (0.499)
Black	2,693	0.266 (0.441)
Child's education	2,693	13.33 (2.007)
Number of siblings under age 3	2,307	1.17 (1.026)
Number of siblings between ages 3 and 6	2,307	0.50 (0.648)
Mother's education	2,693	13.10 (2.072)
Father's education	2,693	13.25 (2.408)
Mother's age when the child is age 1	2,227	25.8 (4.753)
Father's age when the child is age 1	2,227	28.3 (5.594)
Mother's time with child	1,461	5.49 (2.590)
Father's time with child	1,479	2.75 (2.748)
Mother's labor supply	1,544	4.77 (3.905)
Father's labor supply	1,586	9.56 (1.422)
Mother's labor income	1,576	5.22 (6.113)
Father's labor income	1,588	20.55 (11.991)
Year when the child is age 1	2,227	1977 (5.372)

NOTE: Standard deviations are in parentheses. *N* is the number of observations. All variables are measured annually. Education measures the years of completed education. There are fewer observations for annual housework hours than time spent with children because single individuals with no child are coded as missing housework hours and by definition their hours are set to zero for time spent with children. Labor income is measured yearly in 2005 dollars.

SOURCE: Data are from the "Family-Individual" file of the PSID and include individuals surveyed between 1968 and 1997.

Table 2

OLS Estimation of the Production Function

	(1)			(2)		
Variable	**High school**	**Some college**	**College**	**High school**	**Some college**	**College**
Female	0.0047 (0.0130)	0.1257*** (0.0194)	0.0671*** (0.0163)	0.0027 (0.0150)	0.1250*** (0.0239)	0.0850*** (0.0201)
Black	–0.0188 (0.0152)	–0.0233 (0.0227)	–0.0473** (0.0191)	0.0283 (0.0176)	0.0256 (0.0281)	–0.0190 (0.0236)
High school father	0.0478** (0.0216)	0.0723** (0.0322)	0.0251 (0.0271)	0.0521** (0.0256)	0.0664 (0.0408)	0.0058 (0.0343)
Some college father	0.0401** (0.0175)	0.1083*** (0.0260)	0.0623*** (0.0219)	0.0274 (0.0203)	0.0900*** (0.0324)	0.0423 (0.0273)
College father	0.0016 (0.0199)	0.1172*** (0.0296)	0.1538*** (0.0249)	–0.0196 (0.0229)	0.0699* (0.0366)	0.1321*** (0.0308)
High school mother	0.1346*** (0.0240)	0.1781*** (0.0357)	0.0481 (0.0299)	0.0831*** (0.0293)	0.1631*** (0.0468)	0.0472 (0.0393)
Some college mother	–0.0031 (0.0169)	0.0718*** (0.0252)	0.0691*** (0.0211)	–0.0022 (0.0195)	0.0877*** (0.0311)	0.0921*** (0.0262)
College mother	0.0200 (0.0206)	0.0687** (0.0307)	0.0917*** (0.0258)	0.0004 (0.0241)	0.047 (0.0384)	0.0388 (0.0323)
Mother's labor income				–0.0014 (0.0013)	0.0001 (0.0021)	0.0013 (0.0017)
Father's labor income				0.0026*** (0.0007)	0.0033*** (0.0012)	0.0043*** (0.0010)
Constant	0.7028*** (0.0252)	0.1234*** (0.0375)	0.0222 (0.0315)	0.7181*** (0.0321)	0.0917* (0.0512)	–0.0506 (0.0430)
Observations	2,306	2,306	2,306	1,541	1,541	1,541

NOTE: High school, high school graduate. College, college graduate. Standard errors are in parentheses. All results are relative to "high school dropout."

We then describe the relationship between a child's educational outcome and parental education and income, controlling for race. In the relationship, we control for the race and gender of the child. Table 2 presents the results from the ordinary least-squares (OLS) estimation of a linear probability model. The outcome (dependent variable) is the educational attainment of the child. It is divided into four ordered categories according to the highest educational achievement of the child: high school dropout, high school graduate, some college education, and a four-year college degree or above. The category "high school dropout" is omitted from the table. Two specifications, labeled (1) and (2), estimate the linear probability system using children's educational outcomes as the only endogenous variables in the system.

Specification (1) presents the correlation between the educational outcomes of parents and children, controlling for the race and gender of the child. Consistent with findings in the literature, having a college-educated father or mother increases the probability of graduating from college and of having some college education. Having a father or a mother with some college education increases the probability of graduating from college (to a lesser degree than having college-educated parents) or having some college education. Having a mother or a father who is a high school graduate increases the probability of graduating from high school or having some college education but has a small and statistically insignificant effect on the probability of graduating from college (recall that the estimates are relative to having parents with less than a high school education).

Specification (2) adds the mother's and father's income in the first five years of their child's life to the variables already controlled for in specification (1). The income variables are an additional proxy for the parents' socioeconomic status. However, income can also directly affect educational attainment. We look at income only in the first five years because, if there are credit constraints, it may capture the inability to borrow and invest in children during those years. We find that a father's income has a positive, significant effect on a child's educational outcome. A mother's income, however, has no significant effect, which may not be surprising since many women take time off from the labor market to care for young children. Such time off might have a positive effect on a child's educational outcome that offsets the decline in income during the child's early years. The addition of the income variable changes the coefficients on the education variables. For example, there is no longer a significant effect of a mother's college education on her child's educational outcome, while there is still a significant effect of a father's college education on the outcome, although the effect is smaller than before. Clearly there is no causal interpretation because of endogeneity and selection issues.

3 CAUSAL EFFECT OF PARENTAL EDUCATION, INCOME, AND TIME

The above results on the impact of fathers' education on the education of their children are consistent with findings in the literature. However, there is disagreement in the literature about the importance of income and credit constraints on children's educational outcomes. We present below a simple framework for estimation of the production function of education that extends the regression models above. The specification of the production function facilitates a more transparent discussion of the empirical challenges of estimating the causal effect of parental education and skills, as well as investments and the assumptions made about them, on the educational outcomes of children.

3.1 Framework

To capture the impact of parents' characteristics and inputs on children's educational attainment, we specify an "education production function" that accounts for parental time and monetary investment in children as well as the parents' characteristics and skills. We denote a child's education by e' and the child's innate ability by η'. The characteristics of children in the next generation $x' \equiv (e', \eta')$ are affected by their parents' characteristics $x \equiv (e, \eta)$, early childhood monetary investment, early childhood time investments, and the presence and timing of siblings in early childhood. We further index the variables by gender; for instance, e'_f represents the educational outcome of a daughter in the next generation. This intergenerational production function written generically for a child of any gender is determined by the following set of equations:

(1a) $$e'_{f(m)} = \Gamma_{f(m)}\left(x, d^{(0)}, \ldots, d^{(5)}, w^{(0)}, \ldots, w^{(5)}, S_{-5}\right) + \omega'_{f(m)}$$

(1b) $$\eta'_{f(m)} = \Gamma_{f(m)\eta}\left(e'_{f(m)}\right) + \tilde{\eta}'_{f(m)}$$

(1c) $$\Pr\left(\tilde{\eta}' = \tilde{\eta}_i\right) = F_{f(m)}\left(e_f, e_m, \eta_f, \eta_m\right).$$

In the empirical implementation, $\Gamma_{f(m)}$ and $\Gamma_{f(m)\eta}$ are both linear functions. The vector $d^{(j)} = (d_f^{(j)}, d_m^{(j)})$ is the parental time investment at age j of the child, $w^{(j)}$ is the household earnings at age j of the child, S_{-5} is the gender-adjusted number of young siblings present in the household during early childhood, and $\omega'_{f(m)}$ is the gender-specific luck component that determines the educational outcome of the child.[3] A child's innate ability, $\eta'_{f(m)}$, is determined once the education level is determined as the sum of systematic, $\Gamma_{f(m)\eta}(e')$, and random, $\tilde{\eta}'_{f(m)}$, components. The random component, $\tilde{\eta}'_{f(m)}$, is assumed to have finite support and to be independent of $\omega'_{f(m)}$, with probability distribution function $F_{f(m)}(e_f, e_m, \eta_f, \eta_m)$. An important feature of this specification is that it divides the child's ability into a (i) a component determined by parental inputs through the effect of the educational outcome, (ii) innate ability, and (iii) a separable component that is directly transmitted through the parents' innate ability.

3.2 Estimation

Next, we estimate equations (1a) to (1c), which specify the intergenerational production function. We use an instrumental variable identification strategy with a linear probability model (IV-LPM).[4] Although there are three other methods of estimating discrete choice models with endogenous regressors,[5] given the other issues (discussed below) in estimating the intergenerational production functions, the IV-LPM is the most straightforward method for simultaneously dealing with all these issues.

There is a large literature on the estimation of the direct effect of parental traits and investment on children's income in adulthood (see Behrman, 1996; Behrman and Rosenzweig, 2002; and Lee, Roys, and Seshadri, 2014, among others). There are two well-known fundamental problems with estimating the causal intergenerational schooling effect of parents' education. The first is the standard-ability "bias" from the literature on the estimation of the returns to education. That is, more "able" parents may obtain more schooling: If schooling or earnings ability is genetically transmitted to their children, the intergenerational education correlation between children and parents may merely reflect that more-able parents who have more schooling have more-able children who obtain more schooling. The second problem is that the relationship among parental traits, investment, and children's educational outcomes is normally estimated for mothers and children only. Thus, even among mothers with the same abilities, those with higher education may have children with greater educational outcomes and labor market performance because of assortative mating (that is, more-educated women are more likely to marry more-educated men; thus, some of the estimated effects of mothers reflects the unobserved effects of their spouse).

The specification of the education production function in our model, equations (1a) to (1c), internalizes all these concerns, which are accounted for in the estimation as follows. First, we assume that observed ability in the labor market is a monotonic transformation of academic ability; therefore, by using the panel structure of our data, we are able to estimate fixed effects for both parents and children using data on earnings.[6] These estimated fixed effects are then used in the estimation of the education production function to mitigate the ability bias. Second, we include a father's education and time with his child in the education production function while explicitly accounting for household interactions.

However, this approach leads to a third problem: the simultaneity of the inputs of both fathers and mothers and the endogeneity of which parent and type of parent spends time with their child (by "type of parent" we refer to the education and skills of the parent and their spouse). The output of the intergenerational education production function (i.e., the completed education level) is determined across generations, while the inputs, such as parental time investment, are determined over the life cycle of each generation. Therefore, we treat the inputs as predetermined and use instruments from within the system to estimate the production function. This leads to a system of equations that needs to be estimated simultaneously: equations (1a) to (1c), the education production function, as well as equations for the parental labor supply, income, and time spent with children.

To estimate our system, we need a number of exclusion restrictions. The first is the sex composition of siblings; it enters the education production function but not the labor-supply equations. This is similar to the siblings-sex ratio, first used by Angrist and Evans (1998) and is justified on the basis that the sex composition of the children does not have a direct effect on labor supply or the outcome of the child (again, the outcome of the child depends on the child's gender and number of siblings, but not the siblings' sex composition). However, sex composition has an indirect effect on parental time investment in children because (i) parents potentially spend different amounts of time with boys and girls and (ii) it may affect fertility decisions (parents might have preferences, for example, for a balanced sex composition of their children). The second set of instruments—the difference in the age-earnings profile by education—is used to provide quasi-experimental variation in income, labor supply, and subsequent fertility decisions.[7] See Gayle, Golan, and Soytas (2014) for more details and a theoretical intergenerational model that justifies these exclusion restrictions.

3.3 Results

Table 3 presents results of a three-stage least-squares estimation of the system of individual educational outcomes for the educational outcomes equations only.[8] The entire system includes equations for educational outcomes of children, parental time investments, parental incomes, and the parental labor supply. Therefore, in total, a system with nine endogenous variables is estimated. Parental time investment is the sum of the parental time investment over the first five years of the child's life for each parent. The total time investment is a variable that ranges from 0 to 10, since low parental investment is coded as 0 and high parental invest-

Table 3

Three-Stage Least-Squares Estimation of the Production Function

Variable	High school	Some college	College
Female	–0.0036	0.1363***	0.0863***
	(0.0166)	(0.0275)	(0.0222)
Black	0.0070	0.0658	0.0268
	(0.0388)	(0.0626)	(0.0506)
High school father	0.0731**	0.0045	–0.0055
	(0.0324)	(0.0525)	(0.0426)
Some college father	0.0548**	0.1301***	0.0547*
	(0.0235)	(0.0381)	(0.0309)
College father	–0.0416	0.0083	0.1169***
	(0.0319)	(0.0513)	(0.0419)
High school mother	0.0911**	0.0949	–0.0116
	(0.0403)	(0.0646)	(0.0523)
Some college mother	0.0251	–0.0257	0.0213
	(0.0306)	(0.0491)	(0.0398)
College mother	0.0872**	0.1271**	0.0432
	(0.0364)	(0.0575)	(0.0472)
Mother's labor income	–0.0277***	–0.0170	0.0035
	(0.0087)	(0.0137)	(0.0114)
Father's labor income	0.0011	0.0010	0.0024
	(0.0025)	(0.0039)	(0.0033)
Number of siblings under age 3	–0.0063	–0.1018***	–0.0376*
	(0.0166)	(0.0270)	(0.0218)
Number of siblings between ages 3 and 6	–0.0292	–0.0456	–0.0117
	(0.0186)	(0.0302)	(0.0246)
Mother's time with child	–0.0331	0.0668**	0.0595**
	(0.0207)	(0.0334)	(0.0270)
Father's time with child	0.0283	0.1019***	0.0328
	(0.0186)	(0.0293)	(0.0246)
Constant	0.9533***	–0.1619	–0.3259**
	(0.1087)	(0.1738)	(0.1414)
Observations	1,332	1,332	1,332

NOTE: High school, high school graduate. College, college graduate. Standard errors are in parentheses. All results are relative to "high school dropout." Instruments: the sibling sex composition (i.e., the fractions of female siblings under age 3 and between ages 3 and 6) and age-earnings profile (i.e., the linear and quadratic terms of the mother's and father's ages when the child was age 5).

ment is coded as 2. Income variables are constructed as the sum of the annual income of each parent over the first five years of the child's life and measured in 2005 dollars. The labor supply is coded for each parent annually as 0, 1, or 2 corresponding to no work, part-time work, and full-time work. Therefore, the total labor supply is a variable that sums the labor supply over the first five years of the child's life and ranges from 0 to 10.

The education and the race of the parents remain as the exogenous variables in the system estimation, as does the sex of the child. However, sex composition of the siblings and the ages of the parents serve as instruments. Age is measured as the age of the father (mother) when the child was five years old. Two variables are constructed for measuring the effect of sex composition. The variable "Number of siblings under age 3" is the number of siblings less than age 3 when the child was less than 6 years old. Similarly, the variable "Number of siblings between ages 3 and 6" is the number of siblings between the ages of 3 and 6 when the child was less than 6 years old. In the income and labor supply equations only, we use the age of each parent, the age of each parent squared, the age of each parent cubed, and the age of each parent interacted with his or her education. Therefore, we use the exclusion restriction that this set of instruments does not affect the educational outcome equations. Secondly, we use the variables "Number of siblings under age 3" and "Number of siblings between ages 3 and 6" only in the parental time and educational outcome equations, imposing the exclusion restriction that sex composition does not affect the labor supply and income outcomes.

The estimation results show that controlling for all inputs, a child whose mother has a college education has a higher probability of obtaining at least some college education and a significantly lower probability of not graduating from high school relative to a child with a less-educated mother; while the probability of the child graduating from college is also larger, it is not statistically significant. If a child's father, however, has some college or a college education, the child has a higher probability of graduating from college.

Table 3 also shows that while a mother's time investment significantly increases the probability of a child graduating from college or having some college education, a father's time investment significantly increases the probability of the child graduating from high school or having some college education. These estimates suggest that a mother's time investment increases the probability of a high educational outcome, while a father's time investment truncates a low educational outcome. However, the time investment of both parents is productive in terms of a child's educational outcome.

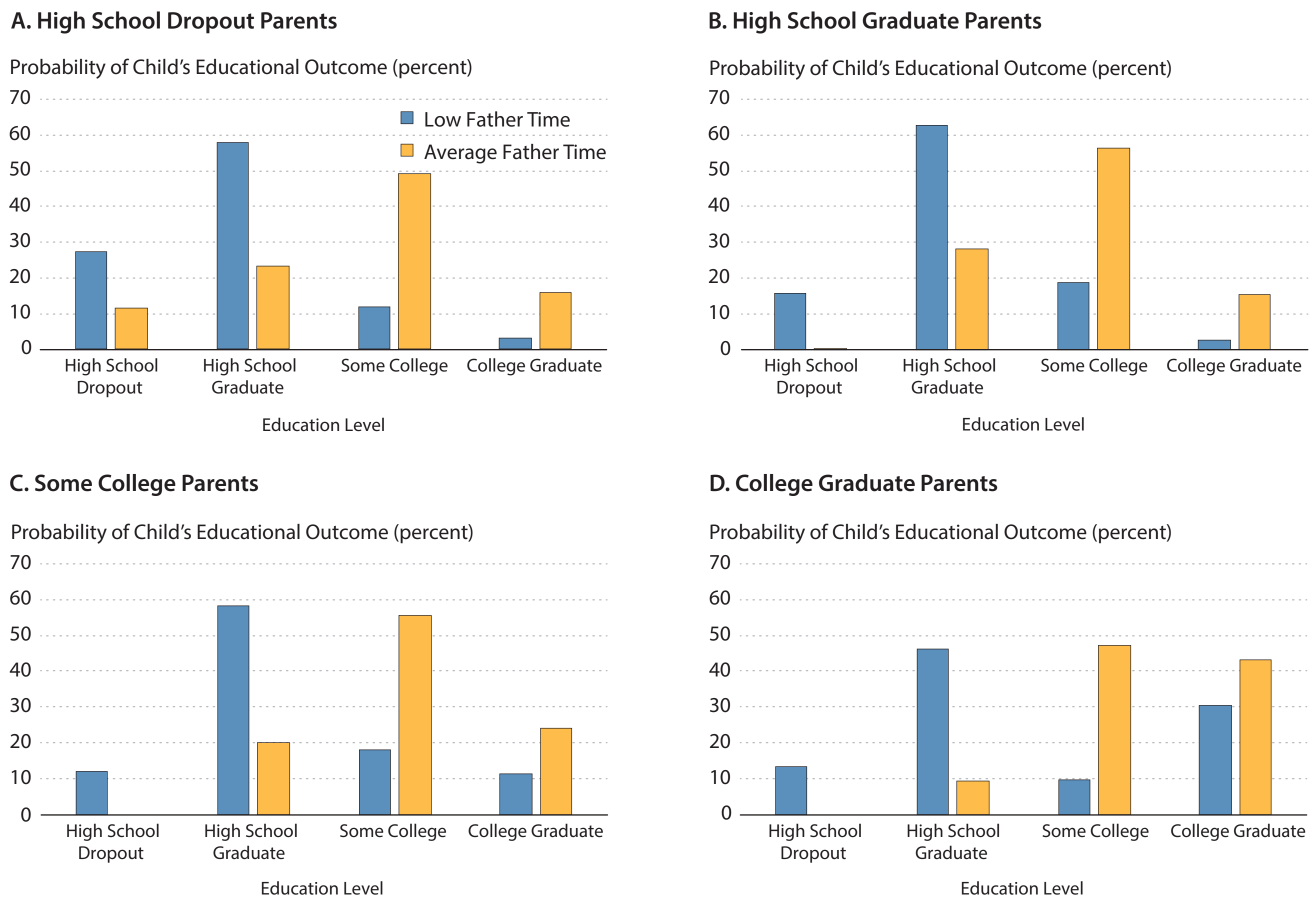
It is important to note that the hours mothers and fathers spend with their children are at different margins, with mothers spending significantly more time with the children than fathers. Thus, the magnitudes of the discrete levels of time investment of mothers and fathers are not directly comparable since what constitutes low and high investment differs across genders.

Figure 1 highlights the relative magnitudes. It shows that fathers' time investment does have a significant impact on the educational outcomes of children. For example, in a household where both parents are high school dropouts, a daughter would have a 3 percent chance of graduating from college if the mother has the sample average time investment and the father has a low time investment for the first five years of the child's life. However, the chance of graduating from college increases to 16 percent if the father increases his time investment to the sample average while the mother's time investment remains at the sample average. A similar pattern holds for all other household types.

Figure 2 highlights the relative importance of parental time investments versus the automatic transmission of education level from parents to children. It highlights the role of both

Figure 1

Three-Stage Least-Squares Estimates of the Relative Importance of Fathers' and Mothers' Time Investment on Children's Educational Outcomes



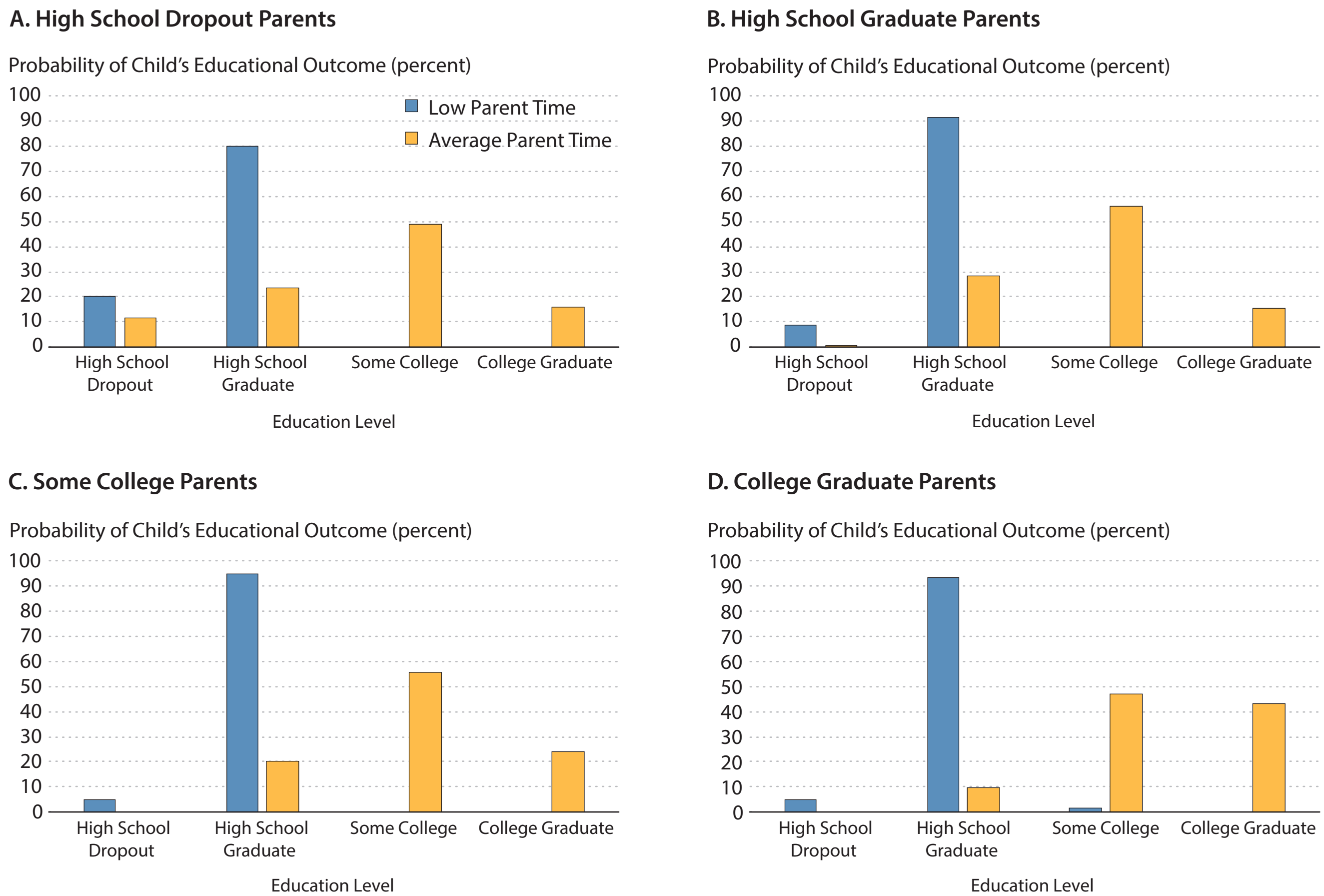
"nature" (education status is automatically transferred from parents to children) and "nurture" (more parental time with children increases the probability of the children having higher educational outcomes). The relative importance of nature versus nurture in accounting for the persistence of earnings across generations is a quantification question that needs to be answered with an optimizing behavioral framework, and parents may take actions that either enhance or diminish the relative effects of nature versus nurture.

4 DISCUSSION AND CONCLUSIONS

In this article, we identify the causal effects of parental traits, time investment, and income in early childhood on children's educational outcomes. We find that after accounting for parental education, skills, and income, both a father's and mother's time investment in the

Figure 2

Three-Stage Least-Squares Estimates of Parental Time Investment Versus the Causal Impact of Parental Education on Children's Educational Outcomes



first five years of a child's life have a large effect on the child's completed education. We find that when a father increases his time investment from low to the average while the mother is at the average, the probability that the child graduates from college increases by 13 percent. When a mother increases her time investment from low to average while the father is at the average, the probability of a child graduating from college increases by 16 percent.

Controlling for all other inputs and parental characteristics, we find that girls have a higher probability of graduating from college. In addition, we do not find evidence for the importance of income in early childhood on children's educational attainment. There is debate in the literature on the effect of credit constraints on children's educational attainment (see Heckman and Mosso, 2014, for a comprehensive review). However, we look at income only in the first five years of a child's life, and it is possible that credit constraints and monetary investment play a more important role later on in a child's development and quality of schooling. Furthermore, credit constraints can have an effect on children's educational outcomes to the extent that credit constraints affect the time parents spend with young children. These issues are important and will be analyzed in future work.

While estimating the causal effects on education in early childhood is important, it is not sufficient for policy assessment. Time investment and monetary inputs, as well as fertility decisions, all potentially affect children's educational outcomes. However, these are all parental decisions that may be affected by different policies. Currently, there is little known on the underlying mechanisms affecting parental choices and input.

A handful of recent papers have analyzed the underlying mechanisms affecting parental choices (see Gayle, Golan, and Soytas, 2014, 2015, and Lee and Seshadri, 2015). Specifically, Lee and Seshadri (2015) and Gayle, Golan, and Soytas (2015) account for the role of investment decisions in the underlying intergenerational correlation in earnings. As we concluded, there is a causal effect of fathers' education and of fathers' and mothers' time during early childhood on children's educational outcomes. However, due to assortative mating, which has increased over the past decades, and because parental time with children depends on how households allocate time among labor market activities, leisure, and household work, understanding the role of the marriage market and households is central to understanding intergenerational mobility patterns.

Gayle, Golan, and Soytas (2014, 2015) explicitly analyze the impact of markets on fertility decisions, household time allocation, and the educational outcomes of children. They find that parental time with children is greatly impacted by marriage markets, and the labor market structure (meaning the nonlinear nature of the return to the labor market and full-time versus part-time work as well as the racial and gender pay gaps). Therefore, the marriage market and the labor market structure have a significant impact on intergenerational mobility, long-term labor market outcomes, and the welfare of children. These findings thus support the large research emphasizing the importance of policies aimed at investment in early childhood. We suggest that it is important to consider the impact of policies on parental time investment in children. ■

NOTES

[1] Starting in 1997, the Child Development Supplement of the PSID provides more extensive family data on children, including parental time on childcare.

[2] For robustness, we benchmarked our constructed parental time variable from the PSID with the data from the American Time Use Survey. See Gayle, Golan, and Soytas (2014).

[3] Indexes (*f*,*m*) are used similarly to represent the mother's and father's time investments.

[4] One reason to use a linear specification is that the nonlinearity in the intergenerational production function itself can generate persistence in earnings across generations. However, we wanted to focus on the economic mechanism that generates persistence of earnings across generations.

[5] The three are the maximum likelihood, control variable, and special regressor approaches. See Lewbel, Dong, and Yang (2012) for a comparison of the different approaches.

[6] For complete estimation details and results of the fixed effects in the labor markets, see Gayle, Golan, and Soytas (2014).

[7] More specifically, the ages of the parents when the child is five years old has an effect on the labor supply because age and education have a deterministic effect on earnings. However, age does not affect directly time investment in children.

[8] For the estimates of the entire system, see Gayle, Golan, and Soytas (2014).

REFERENCES

Angrist, Joshua D. and Evans, William N. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review*, June 1998, *88*(3), pp. 450-77; http://www.jstor.org/stable/116844.

Behrman, Jere R. "From Parent to Child," in Jon R. Neill, ed., *Poverty and Inequality: The Political Economy of Redistribution*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1996, pp. 105-26; https://doi.org/10.17848/9780585314402.ch6.

Behrman, Jere R. and Rosenzweig, Mark R. "Does Increasing Women's Schooling Raise the Schooling of the Next Generation?" *American Economic Review*, March 2002, *92*(1), pp. 323-34; https://doi.org/10.1257/000282802760015757.

Black, Sandra E. and Devereux, Paul J. "Recent Developments in Intergenerational Mobility," in David Card and Orley Ashenfelter, eds., *Handbook of Labor Economics*. Volume 4, Part B. Amsterdam, Netherlands: Elsevier, 2011, pp. 1487-541; https://doi.org/10.1016/S0169-7218(11)02414-2.

Datcher-Loury, Linda. "Effects of Mother's Home Time on Children's Schooling." *Review of Economics and Statistics*, August 1988, *70*(3), pp. 367-73; https://doi.org/10.2307/1926773.

Del Boca, Daniela; Flinn, Christopher and Wiswall, Matthew. "Household Choices and Child Development." *Review of Economic Studies*, January 2014, *81*(1), pp. 137-85; https://doi.org/10.1093/restud/rdt026.

Gayle, George-Levi; Golan, Limor and Soytas, Mehmet A. "What Accounts for the Racial Gap in Time Allocation and Intergenerational Transmission of Human Capital?" Unpublished manuscript, Washington University in St. Louis, Department of Economics, October 2014.

Gayle, George-Levi; Golan, Limor and Soytas, Mehmet A. "What Is the Source of the Intergenerational Correlation in Earnings?" Working Paper 2015-019A, Federal Reserve Bank of St. Louis, August 2015; https://research.stlouisfed.org/wp/2015/2015-019.pdf.

Heckman, James J. and Mosso, Stefano. "The Economics of Human Development and Social Mobility." *Annual Review of Economics*, August 2014, *6*, pp. 689-33; https://doi.org/10.1146/annurev-economics-080213-040753.

Hill, C. Russell and Stafford, Frank P. "Allocation of Time to Preschool Children and Educational Opportunity." *Journal of Human Resources*, 1974, *9*(3), pp. 323-41; http://doi.org/10.2307/144690.

Hill, C. Russell and Stafford, Frank P. “Parental Care of Children: Time Diary Estimates of Quantity, Predictability, and Variety.” *Journal of Human Resources*, 1980, *15*(2), pp. 219-39; http://doi.org/10.2307/145332.

Lee, Sang Yoon; Roys, Nicholas and Seshadri, Ananth. “The Causal Effect of Parental Human Capital on Children’s Human Capital.” Unpublished manuscript, University of Wisconsin–Madison, Department of Economics, December 2014.

Lee, Sang Yoon and Seshadri, Ananth. “On the Intergenerational Transmission of Economic Status.” Unpublished manuscript, University of Wisconsin–Madison, Department of Economics, January 2015.

Leibowitz, Arleen. “Parental Inputs and Children’s Achievement.” *Journal of Human Resources*, 1977, *12*(2), pp. 242-51; http://doi.org/10.2307/145387.

Lewbel, Arthur; Dong, Yingying and Yang, Thomas T. “Comparing Features of Convenient Estimators for Binary Choice Models with Endogenous Regressors.” *Canadian Journal of Economics*, August 2012, *45*(3), pp. 809-29; http://doi.org/10.1111/j.1540-5982.2012.01733.x.