tion (5). This abnormal earnings growth rate corresponds to a real growth in rents of 10 percent (assumed long-term inflation rate is 5.04 percent), which is clearly an unreasonably optimistic assumption.

In sum, our estimates of the equity risk premium using the abnormal earnings approach are considerably lower than the Ibbotson rate, even though we believe the analyst forecasts we use, as well as the terminal growth assumptions we make, are optimistic. Adjusting for such optimism would lower our estimates further. While our estimates from the dividend growth approach are much closer to the Ibbotson rate, we believe they are biased upward because the assumed growth rate ($g = g_5$) is too high an estimate for dividend growth in perpetuity. The estimates from the abnormal earnings approach are more reliable because we use more available information to reduce the importance of assumed growth rates, and we are better able to reject growth rates as being infeasible by projecting rents rather than dividends. Additional benefits of using the abnormal earnings approach are illustrated in Section V.

IV. Equity Premium Estimates from Other Markets

Other equity markets offer a convenient opportunity to validate our domestic results. As long as the different markets are integrated with the United States and are of similar risk, those markets' estimates should proxy for the equity premium in the United States. We replicated the U.S. analysis on five other important equity markets with sufficient data to generate reasonably representative samples of those markets. Only a summary of our results is provided here; details of those analyses are in Claus and Thomas (1999b). The six markets exhibit considerable diversity in performance and underlying fundamentals over our sample period. This across-market variation increases the likelihood that the estimates we obtain from each market offer independent evidence.

As with the U.S. data, earnings forecasts, actual earnings per share, dividends per share, share prices, and the number of outstanding shares are obtained from I/B/E/S. Book values of equity as of the end of year 0 are collected from COMPUSTAT and Global Vantage for Canada and from Datastream for the remaining four countries. Unlike I/B/E/S and COMPUSTAT, Datastream drops firms that are no longer active. While such deletions are less frequent outside the United States, only surviving firms are included in our sample. Fortunately, no bias is created in this study since we equate market valuations with contemporaneous forecasts, and do not track performance.[15] Therefore, even if the surviving firms (included in our sample) performed systematically better or worse than firms that were dropped, our equity premium estimates are unbiased as long as market prices and earnings forecasts in each year are efficient and incorporate the same information.

[15] Note that there is no "backfilling" in our sample, where prior years' data for successful firms are entered subsequently.

All data are denominated in local currency. Currency risk is not an issue here, since it is present in the required rates of returns for both equities and government bonds. Thus the difference between the two rates should be comparable across countries.

We find that analysts' forecasts in these five markets exhibit an optimism bias, similar to that observed in the United States. We considered other potential sources of measurement error in the forecasts, but are confident that any biases created by these errors are unlikely to alter our equity premium estimates much. For example, in Germany, earnings could be computed in as many as four different ways: GAAP per International Accounting Standards, German GAAP, DVFA, and U.S. GAAP.[16] I/B/E/S employees indicated that they have been more successful at achieving consistency in recent years (all forecasts are on a DVFA basis), but they are not as certain about earlier years in their database. While differences in basis between forecast and actual items would affect analyst bias, they do not affect our estimates of market discount rates. Differences in basis across analysts contaminate the consensus numbers used, but the estimated market discount rates are relatively insensitive to changes in the near-term forecasts used.

To select the month of analysis for each country, we followed the same logic as that for the U.S. analysis. December was the most popular fiscal year-end for all countries except for Japan, where it was March. We then identified the period after the fiscal year-end by which annual earnings are required to be disclosed. This period differs across countries (see Table 1 in Alford et al. (1993)): it is three months for Japan and the United States, four months for France, six months for Canada and the United Kingdom, and eight months for Germany. We selected the month following the reporting deadline as the "sure to be disclosed" month to collect forecasts for any given year.

To include a country-year in our sample, we required that the total market value of all firms in our sample exceed 35 percent of the market value of "primary stock holdings" for that country, as defined by Datastream. Although we used a low hurdle to ensure that our sample contained contiguous years for all countries, a substantially greater proportion of the Datastream Market Index than our minimum hurdle is represented for most country-years.

The equity-premium estimates using the abnormal earnings and dividend growth approaches as well as the prevailing risk-free rates for different country-year combinations with sufficient data are reported in Table III. The number of years with sufficient firms to represent the overall market was highest for Canada (all 14 years between 1985 and 1998), and lowest for Japan (8 years). As with the U.S. sample, we use a 50 percent aggregate

[16] The German financial analyst society, Deutsche Vereinigung für Finanzanalyse (DVFA), has developed a system used by analysts (and often by firms) to adjust reported earnings data to provide a measure that is closer to permanent or core earnings. The adjustment process uses both reported financial information as well as firms' internal records. GAAP refers to Generally Accepted Accounting Principles or the accounting rules under which financial statements are prepared in different domiciles.

Table III

Implied Equity Premium Using Abnormal Earnings and Dividend Growth Approaches ($k - r_f$ and $k^* - r_f$) for International Stocks (1985 to 1998)

The market is an aggregate of firms on the I/B/E/S Summary files with forecasts for years +1, +2, and a five-year earnings growth estimate (g_5) as of April each year, and actual earnings, dividends, number of shares outstanding, and prices as of the end of the prior full fiscal year (year 0). Book values of equity for year 0 (bv_0) are obtained from COMPUSTAT, Global Vantage, and Datastream. Forecasted earnings for years +3, +4, and +5 are determined by applying g_5, the forecasted 5-year growth rate, to year +2 forecasted earnings. All amounts are measured in local currencies. r_f is the 10-year government bond yield. The implied discount rate that satisfies the valuation relation in equation (5) below is k. Abnormal earnings (ae_t) equal reported earnings less a charge for the cost of equity (= beginning book value of equity $*$ k). Assuming that 50% of earnings are retained allows the estimation of future book values from current book values and forecast earnings. The terminal value represents all abnormal earnings beyond year +5. Those abnormal earnings are assumed to grow at a constant rate, g_{ae}, which is assumed to equal the expected inflation rate, and is set equal to r_f less 3 percent. The expected rate of return on the market is also estimated using equation (1), and is labeled k^*. Equation (1) is derived from the dividend growth model, and dividend growth in perpetuity, g, is assumed to equal the five-year earnings growth rate, g_5.

$$p_0 = bv_0 + \frac{ae_1}{(1+k)} + \frac{ae_2}{(1+k)^2} + \frac{ae_3}{(1+k)^3} + \frac{ae_4}{(1+k)^4} + \frac{ae_5}{(1+k)^5} + \left[\frac{ae_5(1+g_{ae})}{(k-g_{ae})(1+k)^5}\right] \quad (5)$$

$$k^* = \frac{d_1}{p_0} + g \quad (1)$$

	Canada			France			Germany			Japan			U.K.		
Year	r_f	$k - r_f$	$k^* - r_f$	r_f	$k - r_f$	$k^* - r_f$	r_f	$k - r_f$	$k^* - r_f$	r_f	$k - r_f$	$k^* - r_f$	r_f	$k - r_f$	$k^* - r_f$
1985	10.50%	4.41%	7.45%												
1986	8.82%	2.93%	6.64%												
1987	9.16%	1.56%	4.53%	8.72%	2.06%	6.06%									
1988	9.66%	2.83%	4.67%	9.35%	4.00%	3.90%	6.78%	3.43%	4.59%						
1989	9.29%	3.08%	3.66%	8.76%	3.64%	6.11%	6.83%	3.87%	5.48%				10.16%	3.17%	7.24%
1990	10.69%	1.51%	2.97%	9.66%	3.04%	4.23%	8.99%	1.10%	3.23%				11.39%	2.57%	5.06%
1991	10.08%	0.75%	3.71%	8.81%	2.94%	4.41%	8.42%	1.03%	4.72%	6.72%	−0.95%	0.38%	10.49%	2.47%	7.27%
1992	8.18%	0.42%	6.36%	8.74%	2.26%	5.81%	7.89%	2.16%	5.03%	5.38%	−0.86%	−0.34%	9.12%	2.77%	8.69%
1993	7.32%	1.69%	6.59%	7.18%	2.31%	10.57%	6.14%	0.70%	4.19%	4.45%	−1.05%	4.36%	7.64%	3.29%	10.75%
1994	9.29%	1.65%	7.67%	6.82%	1.70%	8.24%	7.26%	1.30%	8.77%	4.24%	−1.04%	4.56%	8.63%	2.87%	8.50%
1995	7.93%	2.71%	6.77%	7.80%	2.06%	10.04%	6.70%	2.22%	9.84%	2.80%	1.12%	9.50%	8.44%	3.02%	8.59%
1996	7.69%	2.69%	6.89%	6.39%	2.38%	12.26%	6.41%	2.14%	8.40%	3.17%	0.79%	7.82%	7.92%	3.34%	8.43%
1997	6.35%	2.28%	7.10%	5.66%	2.28%	9.69%	5.68%	2.28%	11.56%	2.47%	1.65%	9.46%	7.02%	2.53%	7.81%
1998	5.36%	2.68%	7.44%	5.02%	2.53%	13.44%				1.65%	1.99%	10.89%	5.84%	2.09%	6.77%
Mean	8.59%	2.23%	5.89%	7.74%	2.60%	7.90%	7.11%	2.02%	6.58%	3.86%	0.21%	5.83%	8.66%	2.81%	7.91%
S.D.	1.55%	1.04%	1.62%	1.51%	0.68%	3.27%	1.04%	1.03%	2.82%	1.67%	1.31%	4.27%	1.68%	0.40%	1.49%

dividend payout ratio to generate future dividends and book values, and assume that abnormal earnings grow at the expected inflation rate, which is assumed to be three percent less than the prevailing risk-free rate. For the few years when r_f in Japan is below three percent, we set $g_{ae} = 0$.

The equity premium values based on the abnormal earnings approach $(k - r_f)$ generally lie between two and three percent, except for Japan, where the estimates are considerably lower (and even negative in the early 1990s). Finding that none of the almost 70 estimates of $k - r_f$ reported in Tables II and III are close to the Ibbotson estimate suggests strongly that that historical estimate is too high. In contrast, the equity premium estimates based on the dividend growth approach with dividends growing in perpetuity at the five-year earnings growth forecast (g_5) are considerably higher, similar to the pattern observed in the United States. The dividend growth estimates are very close to those reported in Khorana et al. (1997), which uses a similar approach and a similar sample.

Repeating the sensitivity analyses conducted on the United States (described in Section V) on these five markets produced similar conclusions. The abnormal earnings estimates generate projections that are consistent with experience, but the dividend growth estimates are biased upward and generate projections that are too optimistic because the five-year earnings growth forecast (g_5) is too high an estimate for dividend growth in perpetuity. The values of g_5 suggest mean real dividend growth rates in perpetuity that range between 6.09 percent for Canada and 8.25 percent for Japan. These real rates exceed historic real earnings growth rates, and are at least twice as high as the real GDP growth rates forecast for these countries.

The results observed for Japan are unusual and invite speculation. While our results suggest that the equity premium in Japan increased during the sample period, from about −1 percent in the early 1990s to 2 percent in the late 1990s, these results are also consistent with a stock market bubble that has gradually burst. That is, early in our sample period, prices were systematically higher than the fundamentals (represented by analysts' forecasts) would suggest, and have gradually declined to a level that is supported by analysts' forecasts. Note that our sample excludes the peak valuations in the late 1980s before the crash. Perhaps the implied equity premium in that period would be even more negative than the numbers we estimate for the early 1990s. Regardless of whether the poor performance of Japanese equities in the 1990s is due to correction of an earlier mispricing, it is useful to contrast the inferences from a historic approach with those from a forward-looking approach such as ours: the former would conclude that equity premia have fallen in Japan during the 1990s, whereas our approach suggests the opposite.

V. Sensitivity Analyses

This section summarizes our analysis of U.S. equity data designed to gauge the robustness of our conclusion that the equity premium is much lower than historic estimates. We begin by considering two relations for P/B and P/E ratios that allow us to check whether our projections under the dividend growth and abnormal earnings models are reasonable. Next, we document the extent of analyst optimism in our data. Finally, we consider the sensitivity of our risk premium estimates to the assumed abnormal earnings growth rate (g_{ae}).[17]

A. P/B Ratios and the Level of Future Profitability

The first relation we examine is that between the P/B ratio and future levels of profitability (e.g., Penman (1999)), where future profitability is the excess of the forecast market accounting rate of return (roe_t) over the required rate of return, k.

$$\frac{p_0}{bv_0} = 1 + \frac{roe_1 - k}{(1+k)} + \frac{roe_2 - k}{(1+k)^2}\left(\frac{bv_1}{bv_0}\right) + \frac{roe_3 - k}{(1+k)^3}\left(\frac{bv_2}{bv_0}\right) + \dots, \tag{6}$$

where $roe_t = e_t/bv_{t-1}$ is the accounting return on equity in year t.

This relation indicates that the P/B ratio is explained by expected future profitability ($roe_t - k$).[18] Firms expected to earn an accounting rate of return on equity equal to the cost of capital should trade currently at book values ($p_0/bv_0 = 1$). Similarly, the P/B ratio expected in year +5 (p_5/bv_5), which is determined by the assumed growth in abnormal earnings after year +5 (g_{ae}), should be related to profitability beyond year +5. To investigate the validity of our assumed growth rates, we examine the profiles of future P/B ratios and profitability levels to check if they are reasonable and related to each other as predicted by equation (6). Future book values are generated by adding projected earnings and subtracting projected dividends (assuming a 50 percent payout) to the prior year's book value. Similarly, projected market values are obtained by growing the prior year's market value at the discount rate (k) less projected dividends.

Table IV provides data on current and projected values of P/B ratios and profitability. Current market and book values are reported in columns 1 and 2, and projected market and book values in year +5 are reported in columns

[17] We also examined Value Line data for the DOW 30 firms for two years: 1985 and 1995 (details in Claus and Thomas (1999a)). Value Line provides both dividend forecasts (over a four- or five-year horizon) and a projected price. This price is, in effect, a terminal value estimate, which obviates the need to assume dividend growth in perpetuity. Unfortunately, those risk premium estimates appear to be unreliable: The estimated discount rate is 20 percent (8.5 percent) for 1985 (1995). These results are consistent with Value Line believing that the DOW 30 firms are undervalued (overvalued) in 1985 (1995); that is, current price does not equal the present value of forecast dividends and projected prices. This view is supported by their recommendations for the proportion to be invested in equity: it was 100 percent through the 1980s, and declined through the 1990s (it is currently at 40 percent).

[18] The growth in book value terms in equation (6), bv_t/bv_0, which add a multiplicative effect, have been ignored in the discussion because of the built-in correlation with $roe_t - k$. Higher roe_t results in higher e_t, which in turn causes higher growth in bv_t because dividend payouts are held constant at 50 percent for all years.

Table IV

Price-to-Book Ratios (p_t/bv_t), Forecast Accounting Return on Equity (roe_t) and Expected Rates of Return (k) for U.S. Stocks (1985 to 1998)

To examine the validity of assumptions underlying k, which is the implied discount rate that satisfies the valuation relation in equation (5), current price-to-book ratios are compared with estimated future returns on equity (roe_t) to examine fit with equation (6) below. The market is an aggregate of firms on the I/B/E/S Summary files with forecasts for years +1, +2, and a five-year earnings growth estimate (g_5) as of April each year, and actual earnings, dividends, number of shares outstanding, and prices as of the end of the prior full fiscal year (year 0). Book values of equity for year 0 (bv_0) are obtained from COMPUSTAT. When missing, forecasted earnings for years +3, +4, and +5 are determined by applying g_5 to year +2 forecasted earnings. Assuming that 50 percent of earnings are retained allows the estimation of future book values from current book values and forecast earnings. Return on equity (roe_t) equals forecast earnings scaled by beginning book value of equity (bv_{t-1}). Market and book value amounts are in millions of dollars.

$$p_0 = bv_0 + \frac{ae_1}{(1+k)} + \frac{ae_2}{(1+k)^2} + \frac{ae_3}{(1+k)^3} + \frac{ae_4}{(1+k)^4} + \frac{ae_5}{(1+k)^5} + \left[\frac{ae_5(1+g_{ae})}{(k-g_{ae})(1+k)^5}\right] \quad (5)$$

$$\frac{p_0}{bv_0} = 1 + \frac{roe_1 - k}{(1+k)} + \frac{roe_2 - k}{(1+k)^2}\left(\frac{bv_1}{bv_0}\right) + \ldots \quad (6)$$

Forecasts as of April	Year 0 Equity Values: Market Value (p_0)	Year 0 Equity Values: Book Value (bv_0)	Year +5 Equity Values: Market Value (p_5)	Year +5 Equity Values: Book Value (bv_5)	Price/Book Ratio: In Year 0 (p_0/bv_0)	Price/Book Ratio: In Year 5 (p_5/bv_5)	Forecast Accounting Return on Equity: In Year 1 (roe_1)	In Year 2 (roe_2)	In Year 3 (roe_3)	In Year 4 (roe_4)	In Year 5 (roe_5)	In Year 6 (roe_6)	k from Eq. (5)
	1	2	3	4	5	6	7	8	9	10	11	12	13
1985	1,747,133	1,191,869	2,676,683	1,768,036	1.5	1.5	15%	16%	16%	17%	17%	17%	14.38%
1986	2,284,245	1,214,454	3,197,490	1,783,987	1.9	1.8	15%	16%	16%	17%	17%	17%	11.28%
1987	2,640,743	1,323,899	3,727,459	1,936,215	2.0	1.9	14%	16%	16%	16%	17%	17%	11.12%
1988	2,615,857	1,430,672	3,779,033	2,122,648	1.8	1.8	16%	16%	16%	17%	17%	17%	12.15%
1989	2,858,585	1,541,231	4,200,867	2,341,029	1.9	1.8	17%	17%	17%	18%	18%	18%	12.75%
1990	3,143,879	1,636,069	4,589,685	2,465,373	1.9	1.9	16%	17%	17%	18%	18%	18%	12.33%
1991	3,660,296	1,775,199	5,181,184	2,597,264	2.1	2.0	14%	16%	16%	17%	17%	17%	11.05%
1992	4,001,756	1,911,383	5,574,848	2,773,918	2.1	2.0	13%	15%	16%	16%	17%	17%	10.57%
1993	4,918,359	2,140,668	6,595,210	3,139,088	2.3	2.1	14%	16%	16%	17%	17%	17%	9.62%
1994	5,282,046	2,168,446	7,336,322	3,301,664	2.4	2.2	16%	17%	18%	18%	19%	18%	10.47%
1995	6,289,760	2,670,725	8,837,148	4,132,682	2.4	2.1	17%	18%	18%	19%	19%	19%	11.03%
1996	8,207,274	3,182,952	11,206,787	4,853,189	2.6	2.3	16%	17%	18%	18%	19%	18%	9.96%
1997	10,198,036	3,679,110	14,103,523	5,708,609	2.8	2.5	17%	18%	18%	19%	20%	19%	10.12%
1998	12,908,495	3,412,303	16,838,377	5,378,478	3.8	3.1	17%	18%	19%	20%	21%	20%	8.15%
Mean					2.2	2.1	15%	17%	17%	18%	18%	18%	11.04%

3 and 4. These values are used to generate current and year +5 P/B ratios, reported in columns 5 and 6. Columns 7 through 12 contain the forecasted accounting rate of return on equity for years 1 to 6, which can be compared with the estimated market discount rate, k, reported in column 13, to obtain forecasted profitability.

The current P/B ratio has been greater than 1 in every year in the sample period, and has increased steadily over time, from 1.5 in 1985 to 3.8 in 1998. Consistent with equation (6), all forecasted *roe* values for years 1 through 6 in Table IV exceed the corresponding values of k. Increases in the P/B ratio over the sample period are mirrored by corresponding increases in forecast profitability ($roe_t - k$) in years +1 through +5 as well as forecast profitability in the posthorizon period (after year +5), as measured by the implied price-to-book ratio in year +5. Finally, the tendency for P/B ratios to revert gradually over the horizon toward one (indicated by the year +5 values in column 6 being smaller than the year 0 values in column 5) is consistent with intuition (e.g., Nissim and Penman (1999)).

We also extended our investigation to years beyond year +5 for the assumptions underlying the abnormal earnings estimates, and find that the pattern of projections for P/B and *roe* remain reasonable. In contrast, those projections for the assumptions underlying the dividend growth model estimates suggest that the underlying growth rates are unreasonably high. To provide an illustrative example of those results, we contrast in Figure 2 the patterns for future *roe* and P/B that are projected for the dividend growth and abnormal earnings approaches for 1991. The *roe* levels are marked off on the left scale, and P/B ratios are shown on the right scale. Recall that the market discount rates estimated for the abnormal earnings and dividend growth approaches are 11.05 percent (k) and 15.16 percent (k^*) and the corresponding terminal growth rates for abnormal earnings and dividends are 5.04 percent and 12.12 percent.

The projections for the abnormal earnings method (indicated by bold lines) continue to remain reasonable. The P/B ratio always exceeds one, but it trends down over time. Consistent with P/B exceeding one, the *roe* is always above the 11.05 percent cost of capital, and trends toward it after year +5. Note that the optimistic analyst forecasts cause *roe* projections to climb for years +1 through +5, but the subsequent decline in *roe* is because the profitability growth implied by g_{ae} (our assumed growth in abnormal earnings past year +5) is lower than that implied by g_5.

The results for the dividend growth approach illustrate the benefits of using projected accounting ratios to validate assumed growth rates. The profitability (*roe*) is actually below the cost of equity of 15.16 percent (k^*), for the first three years, even though the P/B ratio is greater than one. Thereafter, the profitability keeps increasing, to a level above 20 percent by year +15. Both the high level of profitability and its increasing trend are not easily justified, especially when they are observed repeatedly for every year in our sample. Similarly, the increasing pattern for P/B, which is projected to increase from about two to about three by year +15, is hard to justify.

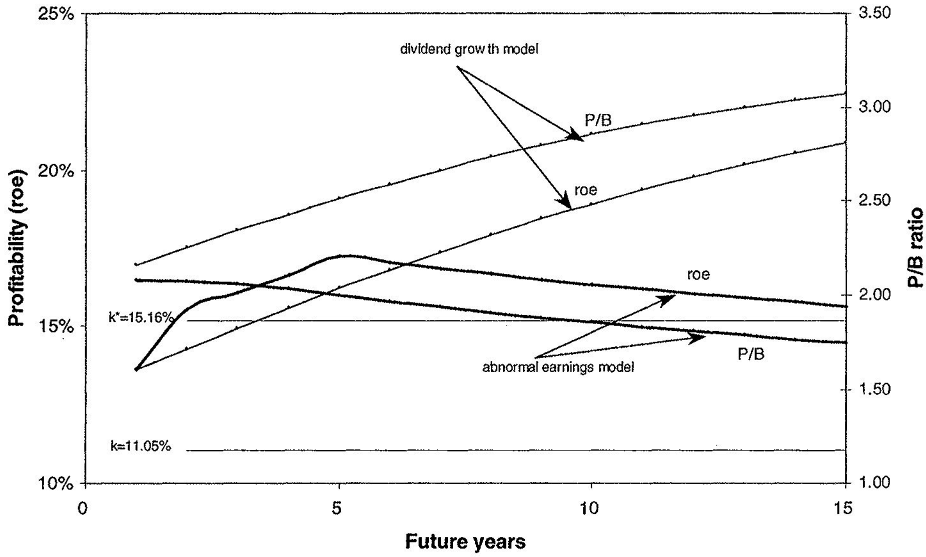


Figure 2. Pattern of future price-to-book (P/B) ratios and profitability, measured as excess of accounting return on equity (*roe*) over estimated discount rates (k^* and k), for dividend growth and abnormal earnings approaches for U.S. stocks as of April, 1991. For the dividend growth model described by equation (1) in Table II, dividends are assumed to grow at the consensus five-year earnings growth rate of 12.12 percent, and future *roe* is compared with the estimated market discount rate of 15.16 percent (k^*). For the abnormal earnings model described by equation (5) in Table II, abnormal earnings are assumed to grow at an anticipated inflation rate of 5.04 percent, and *roe* is compared with the estimated market discount rate of 11.05 percent (k). Projected P/B ratios are shown for both models.

These projections are, however, consistent with an estimated discount rate that is too high. Since near-term analysts' forecasts of profitability are below this discount rate, future levels of profitability have to be unreasonably high to compensate.

B. P/E Ratios and Forecast Growth in Profitability

The second relation we use to check the validity of our assumptions regarding g_{ae} is the price–earnings ratio, described by equation (7) (see derivation in Claus and Thomas, 1999a). Price–earnings ratios are a function of the present value of future changes in abnormal earnings, multiplied by a capitalization factor (= $1/k$).

$$\frac{p_0}{e_1} = \frac{1}{k}\left[1 + \frac{\Delta ae_2}{e_1(1+k)} + \frac{\Delta ae_3}{e_1(1+k)} + \dots\right], \tag{7}$$

where $\Delta ae_t = ae_t - ae_{t-1}$ is the change in expected abnormal earnings over the prior year.

The price–earnings ratio on the left-hand side deviates slightly from the traditional representation in the sense that it is a "forward" price–earnings ratio, based on expected earnings for the upcoming year, rather than a "trailing" price–earnings ratio (p_0/e_0), which is based on earnings over the year just concluded. The relation between future earnings growth and forward price–earnings ratios is simpler than that for trailing price–earnings ratios.[19] Therefore, we use only the forward price–earnings ratio here and refer to it simply as the P/E ratio.

The results reported in Table V describe P/E ratios and growth in abnormal earnings derived from analysts' forecasts for the market. The first four columns provide market values and the corresponding upcoming expected earnings for year 0 and year +5. These numbers are used to generate the current and year +5 P/E ratios reported in columns 5 and 6, which can be compared to the values of $1/k$ reported in column 18.[20] According to equation (7), absent growth in abnormal earnings, the P/E ratio should be equal to $1/k$, and the P/E ratio should be greater (less) than $1/k$ for positive (negative) expected growth in abnormal earnings. Forecast growth rates in abnormal earnings for years +2 through +6 are reported in columns 7 through 11. To maintain equivalence with the terms in equation (7), growth in abnormal earnings is scaled by earnings expected for year +1 (e_1) and then discounted.

To understand the relations among the numbers in the different columns, consider the row corresponding to 1991. The market P/E ratio of 15.1 is higher than the inverse of the discount rate ($1/k$ = 9.0). That difference of 6.1 is represented by the sum of the present value of the abnormal earnings growth terms in future years, scaled by e_1 (this sum needs to be multiplied by $1/k$ as shown in equation (7)). These growth terms decline from 13 percent in year 2 to 2 percent in year 6, and continue to decline thereafter. By year +5, the market P/E is expected to fall (to 11.7), since some of the growth in abnormal earnings (represented by the amounts in columns 7 through 11) is expected to have already occurred by then. Turning to the other sample years, the P/E ratios in year 0 (column 5) have generally increased through the sample period, and so have the values of $1/k$. Consistent with P/E ratios exceeding $1/k$ in every year, abnormal earnings are forecast to exhibit positive growth for all cells in columns 7 to 11. Also, the P/E ratios in year +5 are forecast to decline, relative to the corresponding year 0 P/E values, because of the value represented by the amounts in columns 7 to 11.

[19] Since the numerator of the P/E ratio is an ex-dividend price (p_0), the payment of a large dividend (d_0) would reduce p_0 without affecting trailing earnings (e_0), thereby destroying the relation between p_0 and e_0. This complication does not arise when expected earnings for the upcoming period (e_1) is used instead of e_0.

[20] If the numbers in Table V appear to be not as high as the trailing P/E ratios commonly reported in the popular press, note that forward P/E ratios are generally smaller than trailing P/E ratios for the following reasons. First, next year's earnings are greater than current earnings because of earnings growth. Second, current earnings contain one-time or transitory components that are on average negative, whereas forecast earnings focus on core or continuing earnings.

For purposes of comparison with other work, we also report in columns 12 through 17 of Table V the growth in forecast earnings (as opposed to growth in abnormal earnings) for years +1 through +6. Forecasted growth in earnings declines over the horizon, similar to the pattern exhibited by growth in abnormal earnings. Note the similarity in the pattern of earnings growth for all years in the sample period: the magnitudes of earnings growth estimates appear to settle at around 12 percent by year +5, before dropping sharply to values around 7 percent in the posthorizon period (year +6). Again, this decline occurs because the earnings growth implied by g_{ae} (our assumed growth in abnormal earnings past year +5) is lower than g_5.

The results in Table V confirm the predictions derived from equation (7) as well as the intuitive links drawn in the literature. As with the results for P/B ratios, the trends for P/E ratios and growth in abnormal earnings exhibit no apparent discrepancies that might suggest that the assumptions underlying our abnormal earnings model are unreasonable.

C. Bias in Analyst Forecasts

We considered a variety of biases that may exist in the I/B/E/S forecasts, but found only the well-known optimism bias to be noteworthy (details provided in Claus and Thomas (1999a)).[21] We compute the forecast error for each firm in our sample, representing the median consensus forecast as of April less actual earnings, for different forecast horizons (year +1, +2, . . . +5) for each year between 1985 and 1997. Table VI contains the median forecast errors (across all firms in the sample for each year), scaled by share price. In general, forecasted earnings exceed actual earnings, and the extent of optimism increases with the horizon.[22] There is, however, a gradual reduction in optimism toward the end of the sample period.

Since the forecast errors in Table VI are scaled by price, comparing the magnitudes of the median forecast errors with the inverse of the trailing P/E ratios (or E/P ratios) is similar to a comparison of forecast errors with earnings levels. While the trailing E/P ratios for our sample vary between 5 and 9 percent, the forecast errors in Table VI vary between values that are in the neighborhood of 0.5 percent for year +1 to around 3 percent in year +5. Comparing the magnitudes of year +5 forecast errors with the implied E/P ratios indicates that forecasted earnings exceed actual earnings by as

[21] I/B/E/S removes one-time items (typically negative) from reported earnings. That is, the level of optimism would have been even higher if we had used reported numbers instead of actual earnings according to I/B/E/S.

[22] In addition to increasing with forecast horizon, the optimism bias is greater for certain years where earnings were depressed temporarily. The higher than average dividend payouts observed in Table I for 1987 and 1992 indicate temporarily depressed earnings in those years, and the forecast errors are also higher than average for those years. For example, the two largest median year +2 forecast errors are 1.86 and 1.81 percent, and they correspond to two-year out forecasts made in 1985 and 1990.

Table V

Forward Price-to-Earnings Ratios (p_t/e_{t+1}) and Growth in Forecast Abnormal Earnings and Earnings for U.S. Stocks (1985 to 1998)

To examine the validity of assumptions underlying k, which is the implied discount rate that satisfies the valuation relation in equation (5), current and forecast forward price-to-earnings ratios are compared with growth in forecast abnormal earnings to examine fit with equation (7) below. The market is an aggregate of firms on the I/B/E/S Summary files with forecasts for years +1, +2, and a five-year earnings growth estimate (g_5) as of April each year, and actual earnings, dividends, number of shares outstanding, and prices as of the end of the prior full fiscal year (year 0). Book values of equity for year 0 (bv_0) are obtained from COMPUSTAT. Abnormal earnings (ae_t) equal reported earnings less a charge for the cost of equity (= beginning book value of equity $*$ k). Future market values are projected for each year by multiplying beginning market values by $(1 + k)$ and subtracting dividends. When missing, forecasted earnings for years +3, +4, and +5 are determined by applying g_5 to year +2 forecasted earnings. Assuming that 50 percent of earnings are retained allows the estimation of future book values from current book values and forecast earnings. Market equity values and earnings amounts are in millions of dollars.

$$p_0 = bv_0 + \frac{ae_1}{(1+k)} + \frac{ae_2}{(1+k)^2} + \frac{ae_3}{(1+k)^3} + \frac{ae_4}{(1+k)^4} + \frac{ae_5}{(1+k)^5} + \left[\frac{ae_5(1+g_{ae})}{(k-g_{ae})(1+k)^5}\right] \quad (5)$$

$$\frac{p_0}{e_1} = \frac{1}{k}\left[1 + \frac{\Delta ae_2}{e_1(1+k)} + \frac{\Delta ae_3}{e_1(1+k)^2} + \dots\right] \quad (7)$$

Forecasts as of April	Year 0 Values: Market Value (p_0)	Year 0 Values: Earnings (e_1)	Year +5 Values: Market Value (p_5)	Year +5 Values: Earnings (e_6)	Forward P/E Ratio: In Year 0 (p_0/e_1)	Forward P/E Ratio: In Year 5 (p_5/e_6)	PV of ae Growth (Δae_t), Scaled by e_1: +2	+3	+4	+5	+6	Growth in Forecast Earnings: +1	+2	+3	+4	+5	+6	$1/k$ from Eq. (5)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1985	1,747,133	180,945	2,676,683	308,308	9.7	8.7	5%	3%	3%	3%	1%	17%	13%	11%	11%	12%	9%	7.0
1986	2,284,245	178,024	3,197,490	299,896	12.8	10.7	7%	4%	5%	5%	1%	15%	14%	11%	11%	11%	7%	8.9
1987	2,640,743	186,319	3,727,459	324,573	14.2	11.5	10%	5%	5%	5%	1%	27%	18%	11%	11%	11%	7%	9.0
1988	2,615,857	222,497	3,781,766	364,583	11.8	10.4	4%	4%	4%	4%	1%	33%	11%	11%	11%	11%	8%	8.2
1989	2,858,585	261,278	4,200,867	420,673	10.9	10.0	2%	3%	3%	4%	1%	14%	9%	11%	11%	11%	8%	7.8
1990	3,143,879	257,657	4,589,685	442,911	12.2	10.4	7%	4%	4%	4%	1%	13%	15%	11%	12%	12%	8%	8.1
1991	3,660,296	241,760	5,181,184	442,291	15.1	11.7	13%	5%	6%	6%	2%	11%	22%	12%	12%	12%	7%	9.0
1992	4,001,756	252,109	5,574,848	463,780	15.9	12.0	14%	6%	6%	6%	2%	25%	22%	12%	12%	12%	7%	9.5
1993	4,918,359	295,862	6,595,210	531,812	16.6	12.4	13%	6%	7%	7%	1%	19%	20%	12%	12%	12%	6%	10.4
1994	5,282,046	339,694	7,174,214	604,559	15.5	11.9	11%	6%	6%	7%	1%	17%	19%	12%	12%	12%	6%	10.0
1995	6,289,760	444,593	8,837,148	783,736	14.1	11.3	9%	5%	6%	6%	2%	22%	17%	12%	12%	12%	7%	9.1
1996	8,207,274	512,921	11,206,787	893,185	16.0	12.5	8%	6%	7%	7%	2%	15%	15%	12%	13%	13%	7%	10.0
1997	10,198,036	614,932	14,103,523	1,100,714	16.6	12.8	8%	7%	7%	8%	2%	19%	16%	11%	12%	12%	7%	9.9
1998	12,908,495	577,297	16,838,377	1,069,786	22.4	15.7	12%	9%	10%	11%	2%	19%	16%	11%	12%	12%	7%	12.3
Mean					14.6	11.6	9%	5%	6%	6%	1%	19%	16%	11%	12%	12%	7%	9.2

Table VI

Optimism Bias in I/B/E/S Forecasts for U.S. Stocks: Median Forecast Errors for Forecasts Made Between 1985 and 1997

The following table represents the median of all forecast errors scaled by share price for each year examined. The forecast error is calculated for each firm as of April each year, and equals the median consensus forecasted earnings per share minus the actual earnings per share, scaled by price. The year when the forecasts were made is listed in the first row, while the first column lists the horizon of that forecast. For each year and horizon combination, we report the median forecast error and the number of firms in the sample. To interpret the Table, consider the values of 0.78 percent and 1,680 reported for the +1/1985 combination., in the top left-hand corner of the table. This means that the median value of the difference between the forecasted and actual earnings for 1986 was 0.78 percent of price, and that sample consisted of 1,680 firms with available forecast errors. The results confirm that analyst forecasts are systematically positively biased and that this bias increases with the forecast horizon; however, the extent of any such bias has been declining steadily over time.

		Year Forecast Was Made													
		1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	Mean
Forecast	Median	0.78%	0.65%	0.37%	0.07%	0.44%	0.58%	0.39%	0.17%	0.15%	0.03%	0.04%	0.00%	0.00%	0.28%
Year +1	Obs.	1,680	1,707	1,878	1,815	1,868	1,932	1,959	2,176	2,492	2,710	2,895	3,261	3,462	
Forecast	Median	2.05%	1.40%	0.79%	0.99%	1.74%	1.88%	1.21%	0.87%	0.58%	0.34%	0.32%	0.27%	—	1.04%
Year +2	Obs.	1,545	1,572	1,732	1,701	1,757	1,815	1,896	2,084	2,287	2,594	2,694	2,852	—	
Forecast	Median	2.84%	0.99%	1.44%	2.22%	2.78%	2.39%	1.50%	0.95%	0.63%	0.54%	0.45%	—	—	1.52%
Year +3	Obs.	1,406	1,449	1,596	1,576	1,634	1,744	1,826	1,936	2,159	2,396	2,346	—	—	
Forecast	Median	2.63%	2.04%	2.80%	3.19%	3.17%	2.83%	1.54%	0.91%	0.77%	0.60%	—	—	—	2.05%
Year +4	Obs.	1,285	1,344	1,492	1,474	1,586	1,696	1,724	1,825	2,024	2,132	—	—	—	
Forecast	Median	3.54%	3.44%	3.86%	3.59%	3.43%	2.91%	1.36%	0.94%	0.74%	—	—	—	—	2.65%
Year +5	Obs.	1,201	1,260	1,411	1,432	1,528	1,621	1,618	1,704	1,815	—	—	—	—	

much as 50 percent at that horizon. These results suggest that our equity premium estimates are biased upward because we do not adjust for the considerable optimism in earnings forecasts for years +1 to +5. They also suggest that we are justified in dropping assumed growth rates for earnings past year +5 (column 17 versus column 16 in Table V).

D. Impact of Variation in the Assumed Growth Rate in Abnormal Earnings Beyond Year +5 (g_{ae})

We begin by considering two alternative cases for g_{ae}: three percent less and three percent more than our base case, where g_{ae} is assumed to equal the expected inflation rate. As mentioned in the Appendix, our base growth rate of $g_{ae} = r_f - 3\%$ is higher than any rate assumed in the prior abnormal earnings literature. Adding another three percent to the growth rate, which would require rents to grow at a three percent real rate in perpetuity, raises the level of optimism further. Dropping three percent from the base case, in the lower growth scenario, would be equivalent to assuming a very low nominal growth rate in abnormal earnings, and would be only slightly more optimistic than the assumptions in much of the prior abnormal earnings literature.

For the higher (lower) growth rate scenario, corresponding to $g_{ae} = r_f$ ($g_{ae} = r_f - 6\%$), the average risk premium over the 14-year sample period increases (decreases) to a mean of 4.66 (2.18), from a mean of 3.40 percent for the base case. Even for the high growth rate in abnormal earnings, the increase in the estimated risk premium is modest, and leaves it substantially below the traditional estimates of the risk premium. While increasing (decreasing) the growth rate increases (decreases) the terminal value, it also reduces (increases) the present value of that terminal value because of the higher (lower) discount rate it engenders.

We also considered a synthetic market portfolio each year constructed to have no expected future abnormal earnings, to avoid the need for an assumed abnormal earnings growth rate beyond year +5. As described in equation (6), portfolios with P/B = 1 should exhibit no abnormal earnings; that is, the roe_t should on average equal k for this synthetic market. The last term in equation (5), representing the terminal value of abnormal earnings beyond year +5, is set to zero and the estimates for k obtained iteratively each year. The mean estimate for $k - r_f$ from this synthetic market is 2.20 percent, which is slightly lower than the mean risk premium of 3.40 percent in Table II. Note that a lower discount rate is not expected for the synthetic market, since it has a beta close to one each year and has a lower P/B than the market. (Low P/B firms are expected to generate higher returns (e.g., Gebhardt, Lee, and Swaminathan (forthcoming).) The higher discount rates observed for the assumptions underlying our abnormal earnings model support our view that the analyst forecasts we use and our assumption that the terminal growth in abnormal earnings equals expected inflation ($g_{ae} = r_f - 3\%$) are both optimistic.

VI. Conclusion

Barring some notable exceptions (e.g, Siegel (1992 and 1998), Blanchard (1993), Malkiel (1996), and Cornell (1999)), academic financial economists generally accept that the equity premium is around eight percent, based on the performance of the U.S. market since 1926. We claim that these estimates are too high for the post-1985 period that we examine, and the equity premium is probably no more than three percent. Our claim is based on estimates of the equity premium obtained for the six largest equity markets, derived by subtracting the 10-year risk-free rate from the discount rate that equates current prices to forecasted future flows (derived from I/B/E/S earnings forecasts). Growth rates in perpetuity for dividends and abnormal earnings need to be much higher than is plausible to justify equity premium estimates of about eight percent. Not only are such growth rates substantially in excess of any reasonable forecasts of aggregate growth (e.g., GDP), the projected streams for various indicators, such as price-to-book and price-to-earnings ratios, are also internally contradictory and inconsistent with intuition and past experience.

We agree that the weight of the evidence provided by the historical performance of U.S. stock markets since 1926 is considerable. Yet there are reasons to believe that this performance exceeded expectations, because of potential declines in the equity premium, good luck, and survivor bias. While projecting dividends to grow at earnings growth rates forecast by analysts provides equity premium estimates as high as eight percent, we show that those growth forecasts exhibit substantial optimism bias and need to be adjusted downward. In addition to our results, theory-based work, historical evidence from other periods and other markets, and surveys of institutional investors all suggest that the equity premium is much lower than eight percent. Overall, we believe that an eight percent equity premium is not supported by an analysis that compares current market prices with reasonable expectations of future flows for the markets and years that we examine.

Appendix: Assumed Growth Rates in Perpetuity for Dividends (g) and Abnormal Earnings (g_{ae})

While the conceptual definition of g is clear—it is the dividend growth rate that can be sustained in perpetuity, given current capital and future earnings[23]—determining this rate from fundamentals is not easy. To illustrate, take two firms that are similar in every way, except that they have announced different dividend policies in the current period, which results in a higher expected forward dividend yield (d_1/p_0) for one firm than the other, say 7 percent and 1 percent. What can be said about g for the two firms?

[23] Assuming too high a rate would cause the capital to be depleted in some future period, and assuming too low a rate would cause the capital to grow "too fast."

Examination of equation (1) indicates that g for the low dividend yield firm must be 6 percent higher than g for the higher dividend yield firm, assuming they both have the same discount rate (k^*). If k^* equals 10 percent, for example, the value of g for the two firms must be 3 percent and 9 percent. These two values of g are substantially different from each other, even though the two firms are not.

In addition to being a hypothetical rate, g need not be related to historic or forecasted near-term growth rates for earnings or dividends. Dividend payout ratios can change over time because of changes in the investment opportunity set available and the relative attractiveness of cash dividends versus stock buybacks. Since changes in dividend payout affect the dividend yield, which in turn affects g, historic growth rates may not be relevant for g. Also, if dividend policies are likely to change over time, g need not be related to g_5 (the growth rate forecast for earnings over the next five years), a rate that is frequently used to proxy for g. Various scenarios can be constructed for the two firms in the example above to obtain similar historic and/or near-term forecast growth rates and yet have substantially different values for g.

Despite the difficulties noted above, both historic and forecast rates for aggregate dividends, earnings, and other macroeconomic measures (such as GDP) have been used as proxies for g. We note that these proxies create additional error. First, it is important to hold the unit of investment constant through the period where growth is measured. In particular, any growth created at the aggregate level by the issuance/retirement of equity since the beginning of the period should be ignored. Second, profits from all activities conducted outside the publicly traded corporate sector that are included in the macroeconomic measures should be deleted, and all overseas profits relating to this sector that are excluded from some macroeconomic measures should be included.

To control for the unit of investment problem, we use forecasted growth in per-share earnings rather than aggregate earnings, and to mitigate the problems associated with identifying g, we focus on growth in rents (abnormal earnings), g_{ae}, rather than dividends. To understand the benefits of switching to g_{ae}, it is important to describe some features of abnormal earnings. Expected abnormal earnings would equal zero if book values of equity reflected market values.[24] If book values measure input costs fairly, but do not include the portion of market values that represent economic rents (not yet earned), abnormal earnings would reflect those rents. However, the magnitude of such rents at the aggregate market level is likely to be small, and any rents that emerge are likely to be dissipated over time for the usual reasons (antitrust actions, global competition, etc.). As a result, much of the

[24] That is, if market prices are efficient and book values are marked to market values each period, market (book) values are expected to adjust each period so that no future abnormal returns (abnormal earnings) are expected.

earlier literature using the abnormal earnings approach has assumed zero growth in abnormal earnings past the "horizon" date.[25]

Returning to the two-firm example, shifting the focus from growth in dividends to growth in rents removes much of the confusion caused by transitory changes in dividend payouts and dividend yields: these factors should have no impact on growth in rents, since the level of and growth in rents are determined by economic factors such as monopoly power. That is, even though the two firms have different forecasted earnings and dividends, the forecasted abnormal earnings and growth in abnormal earnings should be identical.

We believe, however, that the popular assumption of zero growth in abnormal earnings may be too pessimistic because accounting statements are conservative and understate input costs: assets (liabilities) tend to be understated (overstated) on average. For example, many investments (such as research and development, advertising, and purchased intangibles) are written off too rapidly in many domiciles. As a result, abnormal earnings tend to be positive, even in the absence of economic rents. Growth in abnormal earnings under conservative accounting is best understood by examining the behavior of the excess of *roe* (the accounting rate of return on the book value of equity) over k (the discount rate). Simulations and theoretical analyses (e.g., Zhang (2000)) of the steady-state behavior of the accounting rate of return under conservative accounting suggest two important determinants: the long-term growth in investment and the degree of accounting conservatism. These analyses also suggest that *roe* approaches k, but remains above it in the long-term.

Even though a decline in the excess of *roe* over k should cause the magnitude of abnormal earnings to fall over time, a countervailing factor is the growth in investment, which increases the base on which abnormal earnings are generated. We assume as a first approximation that the latter effect is greater than the former, and that abnormal earnings increase in perpetuity at the expected inflation rate. Since we recognize that this assumption is an approximation, we elected to err on the side of choosing too high a growth rate to ensure that our equity premium estimates are not biased downward. Also, we conduct sensitivity analyses to identify the impact on our equity premium estimates of varying the assumed growth rate within a reasonable range.

REFERENCES

Abarbanell, Jeffrey and Victor Bernard, 1999, Is the U.S. stock market myopic? *Journal of Accounting Research* 38, 221–242.

→ Abel, Andrew B., 1999, Risk premia and term premia in general equilibrium, *Journal of Monetary Economics* 43, 3–33.

→ Alford, Andrew, Jennifer Jones, Richard Leftwich, and Mark Zmijewski, 1993, The relative informativeness of accounting disclosures in different countries, *Journal of Accounting Research* 31, 183–223.

[25] That is, abnormal earnings persist, but show no growth. Some papers are even more conservative, and have assumed that abnormal earnings drop to zero past the horizon date.

Ang, Andrew and Jun Liu, 2000. A generalized affine earnings model, Working paper, Columbia University.

Bagnoli, Mark, Messod D. Beneish, and Susan G. Watts, 1999, Whisper forecasts of quarterly earnings per share, *Journal of Accounting and Economics* 28, 27–50.

Benore, Charles, 1983, A survey of investor attitudes toward the electrical power industry, Paine Webber Mitchell Hutchins, Inc., New York.

Blanchard, Oliver J., 1993, Movements in the equity premium, *Brookings Papers on Economic Activity* 2, 75–138.

Bogle, John C., 1999, *Common Sense on Mutual Funds* (John Wiley & Sons, New York).

Botosan, Christine A., 1997, Disclosure level and the cost of equity capital, *Accounting Review* 72, 323–350.

Brigham, Eugene F., Dilip K. Shome, and Steve R. Vinson, 1985, Risk premium approach to measuring a utility's cost of equity, *Financial Management* 14, 33–45.

Brown, Stephen J., William N. Goetzmann, and Stephen A. Ross, 1995, Survival, *Journal of Finance* 50, 853–873.

Burr, Barry B., 1998, 20%+ again? Keep dreaming, *Pensions and Investments* 26, 33–34.

Campbell, John Y., 1991, A variance decomposition for stock returns, *The Economic Journal* 101 (405), 157–179.

Campbell, John Y. and Robert J. Shiller, 1988, The dividend-price ratio and expectations of future dividends and discount factors, *Review of Financial Studies* 1, 195–228.

Claus, James J. and Jacob K. Thomas, 1999a, The equity risk premium is lower than you think it is: Empirical estimates from a new approach, Working paper, Columbia University.

Claus, James J. and Jacob K. Thomas, 1999b, Measuring the equity premium using earnings forecasts: An international analysis, Working paper, Columbia University.

Cochrane, John H., 1997, Where is the market going? Uncertain facts and novel theories, *Economic Perspectives* 21, 3–37.

Cornell, Bradford, 1999, *The Equity Risk Premium: The Long-Run Future of the Stock Market* (John Wiley & Sons, New York).

Edwards, Edgar O. and Philip W. Bell, 1961, *The Theory and Measurement of Business Income* (University of California Press, Berkeley, CA).

Fama, Eugene F. and Kenneth R. French, 2000, The equity premium, Working paper, University of Chicago.

Frankel, Richard and Charles M. C. Lee, 1998a, Accounting valuation, market expectation, and cross-sectional stock returns, *Journal of Accounting and Economics* 25, 283–319.

Frankel, Richard and Charles M. C. Lee, 1998b, Accounting diversity and international valuation, Working paper, University of Michigan.

Gebhardt, William, Charles M. C. Lee and Bhaskaran Swaminathan, Toward an ex ante cost of capital, *Journal of Accounting Research* forthcoming.

Glassman, James K. and Kevin A. Hassett, 1998, Are stocks overvalued? Not a chance, *Wall Street Journal*, March 30.

Gordon, Myron, 1962, *The investment, financing and valuation of the corporation* (Irwin, Homewood, IL).

Ibbotson Associates, Inc., 1999, Stocks, Bonds, Bills and Inflation, *Yearbook*, Chicago, IL.

Jorion, Philipe and William N. Goetzmann, 1999, Global stock markets in the twentieth century, *Journal of Finance* 54, 953–980.

Khorana, Ajay, R. Charles Moyer, and Ajay Patel, 1997, An examination of equity risk premium forecasts in G-6 countries, Working paper, Babcock Graduate School of Management, Wake Forest University.

Kocherlakota, Naryana, 1996, The equity premium: It's still a puzzle, *Journal of Economic Literature* 34, 42–71.

Liu, Jing and Jacob K. Thomas, 2000, Stock returns and accounting earnings, *Journal of Accounting Research* 38, 71–101.

Malkiel, Burton G., 1996, The equity premium puzzle, Keynote speech at 1996 Ibbotson Associates equity risk premium conference, Chicago, IL.

Matsumoto, Dawn A., 1999, Management's Incentives to Guide Analysts' Forecasts, Working paper, Harvard Business School.

Mayfield, E. Scott, 1999, Estimating the market risk premium, Working paper, Harvard Business School.

Mehra, Rajnish and Edward C. Prescott, 1985, The Equity Premium: A Puzzle, *Journal of Monetary Economics* 15, 145–161.

Moyer, R. Charles, and Ajay Patel, 1997, The equity market risk premium: a critical look at alternative ex ante estimates, Working paper, Babcock Graduate School of Management, Wake Forest University.

Myers, James N., 1999, Conservative accounting and finite firm life: Why residual income valuation estimates understate stock price, Working paper, University of Washington.

Nissim, Doron and Stephen H. Penman, 1999, Ratio analysis and equity valuation, Working paper, Columbia Business School.

Penman, Stephen H., 1999, *Financial Statement Analysis and Security Valuation* (Irwin/McGraw Hill, New York).

Preinreich, Gabriel A. D., 1938, Annual survey of economic theory: the theory of depreciation, *Econometrica* 6, 219–241.

Samuelson, Paul, 1965, Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review* 6, 41–49.

Siegel, Jeremy J., 1992, The equity premium: Stock and bond returns since 1802, *Financial Analysts Journal* January/February, 28–38.

Siegel, Jeremy J., 1998, *Stocks for the long run,* 2nd edition (McGraw-Hill, New York).

Siegel, Jeremy J., and Richard H. Thaler, 1997, The equity risk premium puzzle, *Journal of Economic Perspectives* 11, 191–200.

Stulz, René M., 1999, Globalization, corporate finance, and the cost of capital, *Journal of Applied Corporate Finance* 12(3), 8–25.

Vuolteenaho, Tuomo, 1999, Understanding the aggregate book-to-market ratio, Working paper, University of Chicago.

Welch, Ivo, 1999, Views of financial economists on the equity premium and professional controversies, Working paper, UCLA.

Wien, Byron R., 1998, U.S. strategy: Risk premium—R.I.P. Morgan Stanley Dean Witter.

Williams, John. B., 1938, *The Theory of Investment Value* (Harvard University Press, Cambridge, MA).

Zhang, Xiao-jun, 2000, Conservative accounting and equity valuation, *Journal of Accounting and Economics* 29, 125–149.

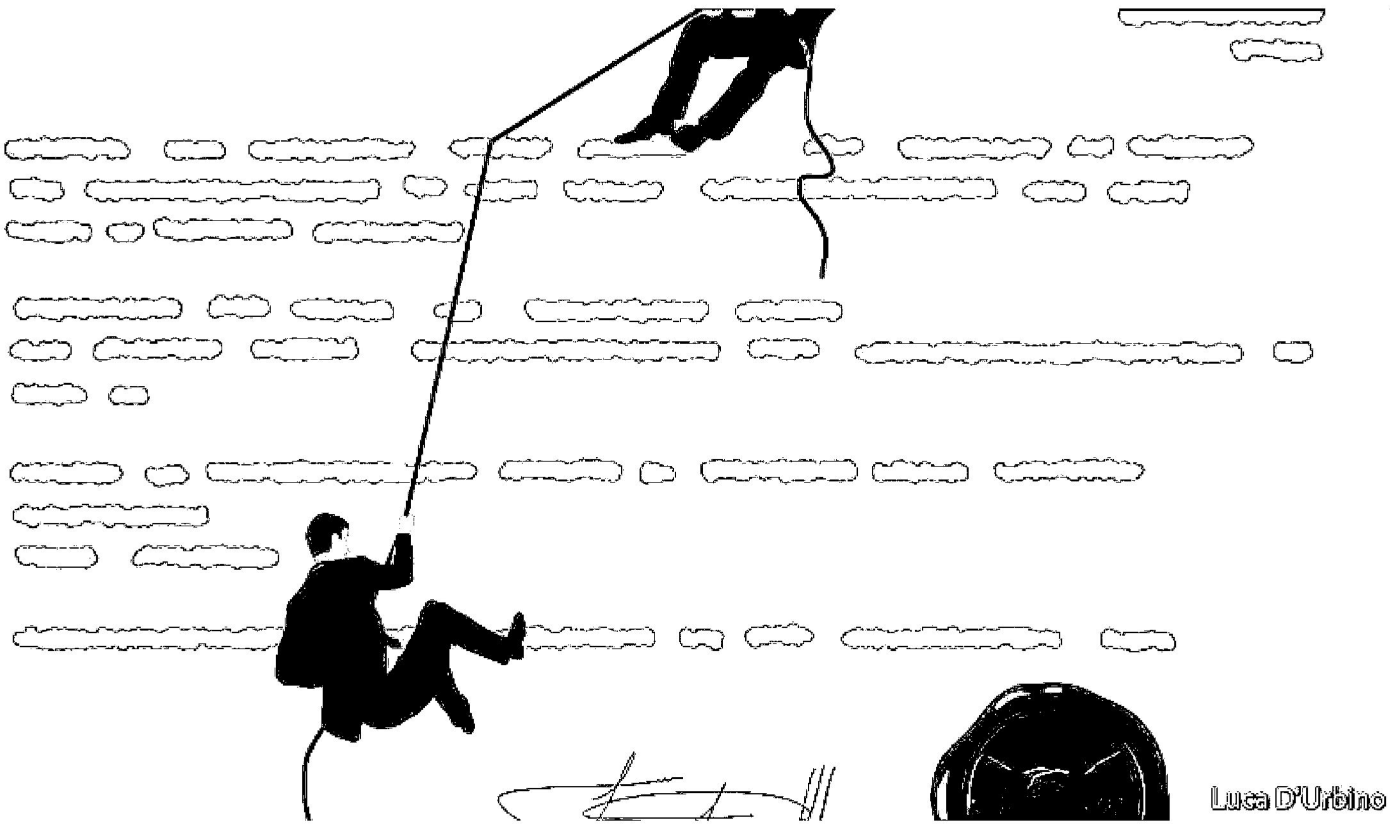
≡ Menu

RW

Schools brief | Six big ideas

Coase's theory of the firm

If markets are so good at directing resources, why do companies exist? The first in our series on big economic ideas



Luca D'Urbino

Jul 29th 2017

Saved Share Give

ONE morning, an economist went to buy a shirt. The one he chose was a marvel of global production. It was made in Malaysia using German machines. The cloth was woven from Indian cotton grown from seeds developed in America. The collar lining came from Brazil; the artificial fibre from Portugal. Millions of shirts of every size and colour are sold every day, writes Paul Seabright, the shirt-buying economist, in his 2004 book, "The Company of Strangers". No authority is in charge. The firms that make up the many links in the chain that supplied his shirt had merely obeyed market prices.

Throwing light on the magic of market co-ordination was a mainstay of the "classical" economics of the late-18th and 19th centuries. Then, in 1937, a paper published by Ronald Coase, a British economist, pointed out a glaring omission. The standard model of economics did not fit with what goes on within companies. When an employee switches from one division to another, for instance, he does not do so in response to higher wages, but because he is ordered to. The question posed by Coase was profound, if awkward for economics:

why are some activities directed by market forces and others by firms?

His answer was that firms are a response to the high cost of using markets. It is often cheaper to direct tasks by fiat than to negotiate and enforce separate contracts for every transaction. Such "exchange costs" are low in markets for standardised goods, wrote Coase. A well-defined task can easily be put out to the market, where a contractor is paid a fixed sum for doing it. The firm comes into its own when simple contracts of this kind will not suffice. Instead, an employee agrees to follow varied and changing instructions, up to agreed limits, for a fixed salary.

Coase had first set out his theory while working as a lecturer in Dundee, in 1932, having spent the prior academic year in America, visiting factories and businesses. "The nature of the firm", his paper, did not appear for another five years, in part because he was reluctant to rush into print. Though widely cited today, it went largely unread at first. But a second paper, "The problem of social cost", published in 1960, by which time he had moved to America, brought him to prominence. It argued that private bargaining could resolve social problems, such as pollution, as long as property rights are well defined and transaction costs are low (they rarely are). He had been asked to expound his new theory earlier that year to a sceptical audience of University of Chicago economists. By the end of the evening, he had won everyone around. Coase was invited to join the university's faculty in 1964; and there he remained until his death in 2013 at the age of 102.

In 1991 Coase was awarded the Nobel prize for economics, largely on the strength of these two papers. But as late as 1972, he lamented that "The nature of the firm" had been "much cited and little used". In a strange way, Coase himself was partly to blame. The idea of transaction costs was such a good catch-all explanation for tricky subjects that it was used to close down further inquiry. In fact, Coase's paper raised as many difficult questions as it answered. If firms exist to reduce transaction costs, why have market transactions at all? Why not further extend the firm's boundaries? In short, what decides how the economy as a whole is organised?

Almost as soon as Coase had wished for it, a body of more rigorous research on such questions began to flourish. Central to it was the idea that it is difficult to specify all that is required of a business relationship, so some contracts are necessarily "incomplete". Important figures in this field include Oliver Williamson, winner of the Nobel prize in economics in 2009, and Oliver Hart and Bengt Holmstrom, who shared the prize in 2016. These and other Coase apostles drew on the work of legal theorists in distinguishing between spot transactions and business relations that require longer-term or flexible contracts.

Spot markets cover most transactions. Once money is exchanged for goods, the deal is completed. The transaction is simple: one party wants, another supplies. There is little scope for dispute, so a written contract can be dispensed with. If one party is unhappy, he will take his business elsewhere next time. Spot markets are thus largely self-policing. They are well suited to simple, low-value transactions, such as buying a newspaper or taking a taxi.

Things become trickier when the parties are locked into a deal that is costly to get out of. Take a property lease, for instance. A business that is evicted from its premises might not quickly find a building with similar features. Equally, if a tenant suddenly quit, the landlord might not find a replacement straight away. Each could threaten the other in a bid for a better rent. The answer is a long-term contract that specifies the rent, the tenure and use of the property. Both parties benefit.

But for many business arrangements, it is difficult to set down all that is required of each party in all circumstances. In such cases, formal contracts are by necessity "incomplete" and sustained largely by trust. An employment contract is of this type. It has a few formal terms: job title, work hours, initial pay and so on, but many of the most important duties and obligations are not written down. It is thus like a "mini-society with a vast array of norms beyond those centred on the exchange and its immediate processes", wrote Mr Williamson. Such a contract stays in force mostly because its breakdown would hurt both parties. And because market forces are softened in such a contract, it calls for an alternative form of governance: the firm.

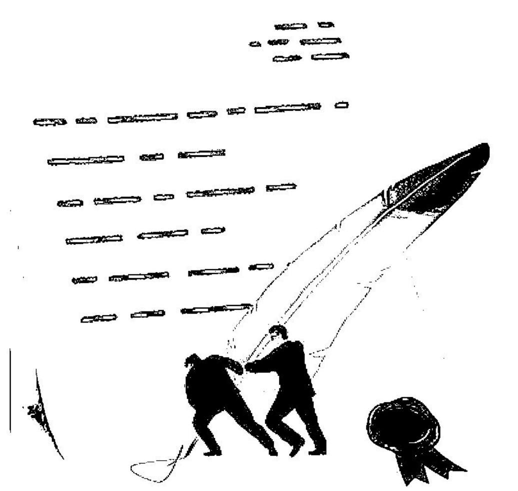
One of the first papers to elucidate these ideas was published in 1972 by Armen Alchian and Harold Demsetz. They defined the firm as the central contractor in a team-production process. When output is the result of a team effort, it is hard to put the necessary tasks out to the market. That is because it is tricky to measure the contribution of each member to the finished work and to then allocate their rewards accordingly. So the firm is needed to act as both co-ordinator and monitor of a team.

Chain tale

If a team of workers requires a firm as monitor, might that also be true for teams of suppliers? In some cases, firms are indeed vertically integrated, meaning that suppliers of inputs and producers of final goods are under the same ownership. But in other cases, suppliers and their customers are separate entities. When is one set-up right and not the other?

A paper published in 1986 by Sanford Grossman and Mr Hart sharpened the thinking on this. They distinguished between two types of rights over a firm's assets (its plant, machinery, brands, client lists and so on): specific rights, which can be contracted out, and residual rights, which come with ownership. Where it becomes costly for a company to specify all that it wants from a supplier, it might make sense to acquire it in order to claim the residual rights (and the profits) from ownership. But, as Messrs Grossman and Hart noted, something is also lost through the merger. The supplier's incentive to innovate and to control costs vanishes, because he no longer owns the residual rights.

To illustrate this kind of relationship, they used the example of an insurance firm that pays a commission to an agent for selling policies. To encourage the agent to find high-quality clients, which are more likely to renew a policy, the firm defers some portion of the agent's pay and ties it to the rate of policy renewals. The agent is thus induced to work hard to find good clients. But there is a drawback. The insurance firm now has an incentive of its own to shirk. While the agent is busting a gut to find the right sort of customers, the firm can take advantage by, say, cutting its spending on advertising its policies, raising their price or lowering their quality.



Luca D'Urbino

There is no set-up in which the incentives of firm and agent can be perfectly aligned. But Messrs Grossman and Hart identified a next-best solution: the party that brings the most to any venture in terms of "non-contractible" effort should own the key assets, which in this case is the client list. So the agent ought to own the list wherever policy renewals are sensitive to sales effort, as in the case of car insurance, for which people tend to shop around more. The agent would keep the residual rights and be rewarded for the effort to find the right sort of client. If the insurance firm shirks, the agent can simply sell the policies of a rival firm to his clients. But in cases where the firm brings more to the party than the sales agent—for example, when clients are "stickier" and the first sale is crucial, as with life insurance—a merger would make more sense.

This framework helps to address one of the questions raised by Coase's original paper: when should a firm "make" and when should it "buy"? It can be applied to vertical business ties of all kinds. For instance, franchises have to abide by a few rules that can be set down in a contract, but get to keep the residual profits in exchange for a royalty fee paid to the parent firm. That is because the important efforts that the parent requires of a franchisee are not easy to put in a contract or to enforce.

The management of ties between a firm and its "stakeholders" (its customers, suppliers, employees and investors) is another variation on this theme. A firm often wants to put restraints on the parties it does business with. Luxury-goods firms or makers of fancy sound equipment may ban retailers from discounting their goods as a way to spur them to compete with rivals on the quality of their shops, service and advice.

Inside the cubicle

If one of the challenges set by Coase was to explain where the boundary between firms and markets lies, another was for economic analysis not to cease once it reached the factory gate or office lobby. A key issue is how agreements are structured. Why, for instance, do employment contracts have so few formal obligations? One insight from the literature is that a tightly specified contract can have perverse outcomes. If teachers are paid according to test results, they will "teach to the test" and pay less regard to other tasks, such as inspiring pupils to think independently. If chief executives are paid to boost the firm's short-term share price, they will cut investment projects that may benefit shareholders in the long run.

Mr Holmstrom and Paul Milgrom established that where important tasks are hard to monitor, and where a balance of activities is needed, then a contract should shun strong incentives tied to any one task. The best approach is to pay a fixed salary and to leave the balance of tasks unspecified. A related idea developed by Mr Hart and John Moore is of a job contract as a "reference point" rather than as a detailed map. Another insight is that deferred forms of pay, such as company pension schemes and promotions based on seniority, help cement long-term ties with employees and reward them for investing in skills specific to the relationship.

Coase noted in 1937 that the degree to which the mechanism of price is superseded by the firm varies with the circumstances. Eighty years on, the boundary between the two might appear to be dissolving altogether. The share of self-employed contractors in the labour force has risen. The "gig economy" exemplified by Uber drivers is mushrooming.

Yet firms are unlikely to wither away. Prior to Uber, most taxi drivers were already self-employed. Spot-like job contracts are becoming more common, but flexibility comes at a cost. Workers have little incentive to invest in firm-specific skills, so productivity suffers. And even if Mr Seabright's shirt was delivered by a set of market-based transactions, the supply chains for complex goods, such as an iPhone or an Airbus A380 superjumbo, rely on long-term contracts that are often "incomplete". Coase was the first to spot an enduring truth. Successful economies need both the benign dictatorship of the firm and the invisible hand of the market.

Coase's theory of the firm: a reading list

1 "The Nature of the Firm" by R H Coase, Economica, 1937

2 "The Problem of Social Cost" by R H Coase, Journal of Law and Economics, 1960

3 "Industrial Organisation: A Proposal for Research" by R H Coase, NBER, 1972

4 "Production, Information Costs and Economic Organisation" by Armen A Alchian and Harold Demsetz, American Economic Review, 1972

Review, 1972
5 “Transaction-Cost Economics: The Governance of Contractural Relations” by Oliver E Williamson, Journal of Law and Economics, 1979
6 “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration” by Sanford Grossman and Oliver Hart, Journal of Political Economy, 1986
7 “Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design” by Bengt Holmstrom and Paul Milgrom, Journal of Law, Economics and Organisation, 1991
8 “The Firm as Sub-economy” by Bengt Holmstrom, Journal of Law Economics & Organisation, 1999
9 “The Theory of the the Firm as Governance Structure: From Choice to Contract” by Oliver E Williamson, 2002
10 “Contracts as Reference Points” by Oliver Hart and John Moore, Quarterly Journal of Economics, 2008

This article appeared in the Schools brief section of the print edition under the headline "Coase call"

From the July 28th 2017 edition

Discover stories from this section and more in the list of contents

Explore the edition

Saved Share Give Reuse this content

More from Schools brief

Finding living planets

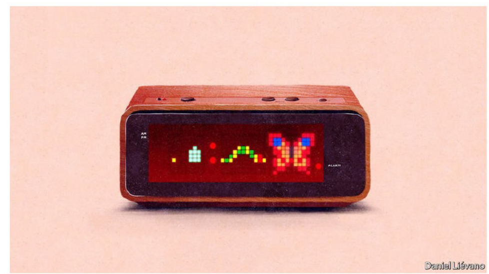
Life evolves on planets. And planets with life evolve

On the origin of "species"

The term, though widely used, is hard to define

Making your way in the world

An individual's life story is a dance to the music of time



How Do We Have 18.4% Earnings Growth In A 2.58% GDP Economy?

Apr. 19, 2018 1:04 PM ET10 comments | 1 Like
by: Matt Comer

Summary

- Growth rate of a stock equals dividend plus earnings growth.
- There are short-term factors that have boosted earnings.
- International growth makes up the largest portion of earnings growth.

Growing earnings have propelled the overall markets to new highs the last few years. With first-quarter earnings expected to grow at an 18.4% annual rate, it raises the question of how is this possible when GDP is only 2.58%? I researched this issue and found that many reputable sources have attacked this issue from different angles. My aim is to try and break down and simplify this issue and discuss why there is such a large discrepancy. There are many factors driving the discrepancy between earnings growth and GDP growth, but the primary factor is international earnings growth.

Going Over The Basics

Just in case some readers have forgotten, the growth rate of a stock should equal its dividend rate plus its earnings growth, assuming the PE ratio stays constant. A company late in its growth stage will usually pay a larger dividend than a younger company that is rapidly growing. The rapidly growing company has ample opportunity to re-invest earnings, which will allow it to grow its earnings at a faster rate than a company paying a large dividend.

The PE ratio is a little beyond the scope of this article, but it is worth noting the current S&P 500 trailing PE of 24.24 is high based on historical standards. A major reason this ratio is high is because analysts predict rapidly growing earnings and the forward PE ratio is 17.12. The question still remains, how do we have such rapidly growing earnings in a slow growth economic environment?

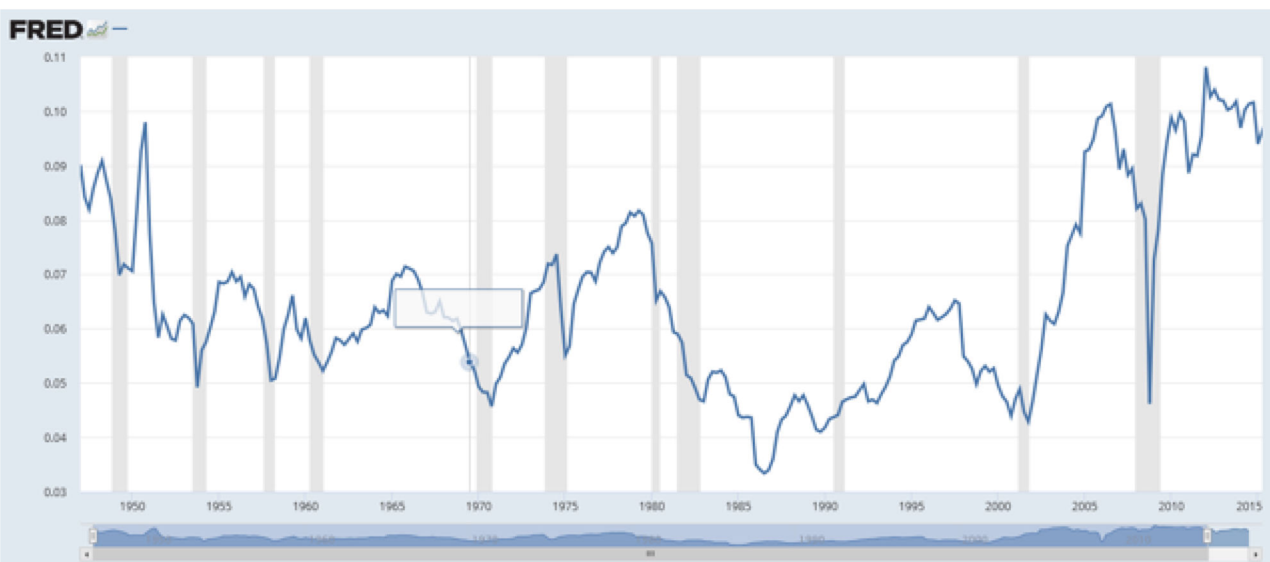
S&P Revenue is Not the Same as GDP but Should be Close in a Closed Economy

S&P revenue does not equal GDP and is calculated in a much different way. Gross domestic products and services are the total value of goods produced and services provided in a country during one year. Services make up a much larger portion of GDP, and manufacturing makes up a much larger portion of S&P 500 earnings. Consumer spending has a much larger impact on GDP and business spending impacts S&P 500 revenue more. There are other differences, but the biggest difference is obviously that GDP growth does not include the international sales of S&P 500 companies. With GDP, globalization has a negative impact since exports are subtracted from imports.

Despite the differences in how GDP growth and the S&P 500 earnings growth are calculated, historically, these numbers have been very close. We will discuss the impact of increasing globalization in another section, but in a closed economy, these two numbers are similar over an extended period of time. Over the short term, there are factors that can cause GDP and S&P earnings to differ in a closed economy.

There are Short-Term Factors Impacting Earnings

What if profit margins rise in a closed economy and earnings make up a greater percentage of revenue? Due to lower labor cost, a weak dollar, and low interest rates, that is exactly what have been occurring with the S&P 500. The following chart shows corporate profits as a percentage of GDP.



As the chart shows, corporate profits currently make up a much larger percentage than they have historically. With the labor market tightening, interest rates rising, and the dollar gaining strength, the portion of this trend related to U.S. earnings is likely to reverse. Over the long term, corporate profits as a percentage of GDP should drop to historical norms in a closed economy.

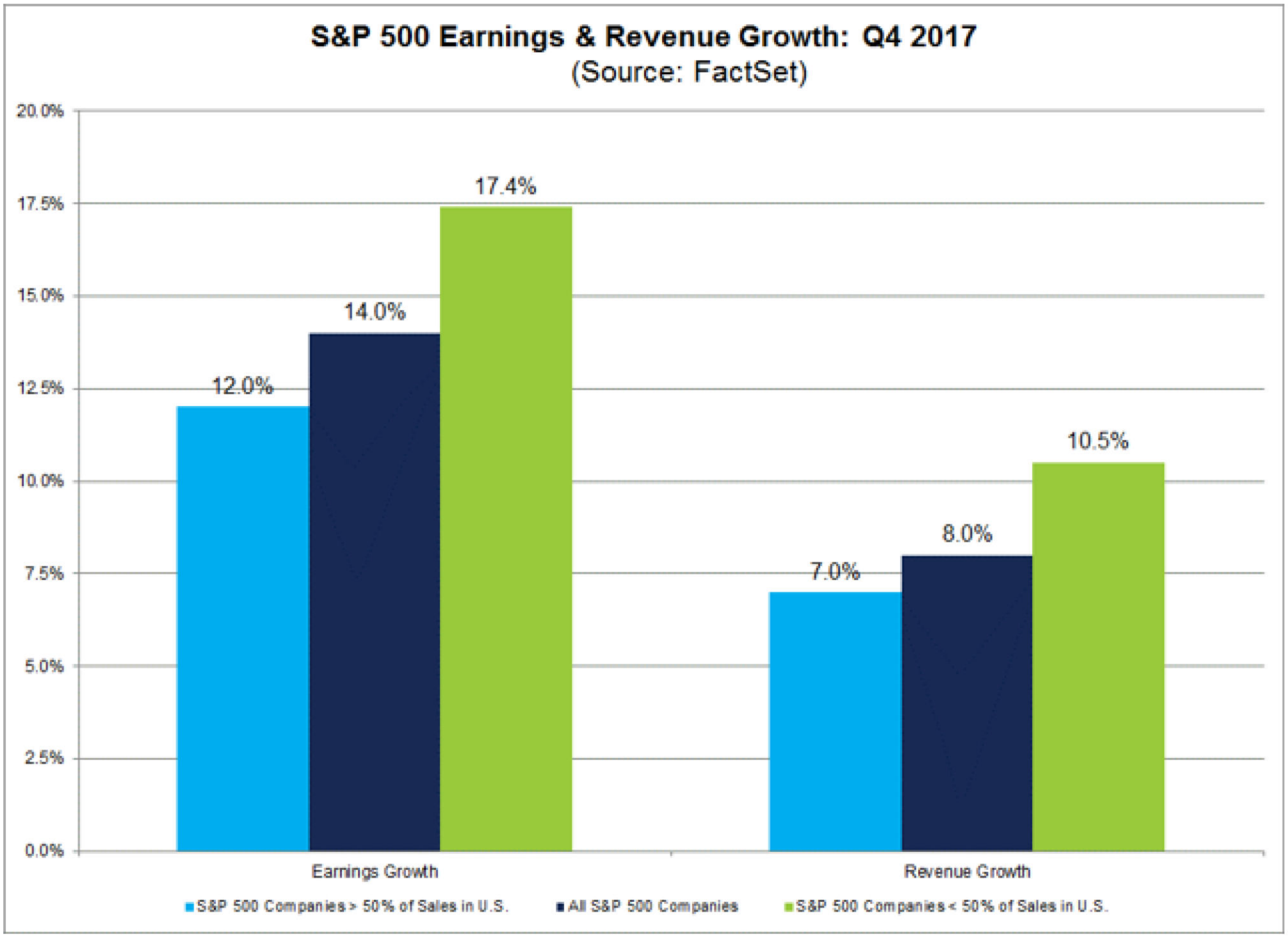
This is Not a Closed Economy

The most obvious difference between GDP growth and S&P 500 earnings growth is international earnings growth. GDP growth equals population growth plus productivity growth. With emerging economies like China and India rapidly increasing production as their population expands, it is no secret that their GDP growth rates are higher than developed countries like the U.S. This of course means if General Motors (GM) manufactures and sells cars in China, their growth rate for their Chinese business should roughly match China's relatively higher GDP growth rate. With China's GDP growth exceeding U.S. GDP growth, it is easy to see how a company with business in China might have an earnings growth rate that exceeds U.S. GDP.

I think varying GDP growth rates in different countries is well understood, but what about growth attributed to companies rapidly gaining market share within emerging market countries? For example, there is very little room for Walmart (WMT) to grow in the United States, as the market is clearly saturated with their stores. However, with far fewer stores in Mexico, Walmart's growing market share in Mexico could cause its growth rate to far exceed Mexico's GDP rate. Walmart could add stores in new cities and rapidly grow throughout Mexico, just as they did years ago in the United States. Earnings growth should roughly equal GDP growth in a closed market, but an established and dominant company like Walmart can far exceed the GDP growth rate when rapidly acquiring market share in a new market.

Understanding how international growth could impact S&P earnings is easier than quantifying it. This is because not all companies in the S&P 500 offer sales and earnings data for each region they do business in. Analysts have tried to organize data by either assuming all companies which didn't separate international sales had no international sales, or by only using companies which did offer complete international sales data. Both of these methods are imperfect, but it is certain that international sales make up an increasing amount of total S&P 500 revenue.

In the fourth quarter of 2017, companies with over 50 percent of international exposure reported higher earnings growth than companies with less than 50% international sales exposure. The following chart shows the earnings growth and revenue growth for S&P 500 companies based on their exposure to international business. This analysis assumes companies which didn't report international earnings had no international earnings.



As the chart shows, earnings growth for companies with over half of their sales in foreign countries was 5.4% greater than companies with less international exposure. The chart also shows international earnings growth exceeded international sales growth. These charts certainly show a direct correlation between international growth and earnings growth, but it is tough to tell the exact nature of this growth. We still don't know how much earnings growth was attributed to emerging markets such as China and developed international markets such as the UK and most of Europe. We also don't know how much of this growth can be attributed to normal GDP growth, and how much is attributed to the market share increases companies experience when they enter a new market.

Conclusion

Earnings growth has far exceeded GDP growth recently and it is difficult to understand exactly why. Short-term factors such as a weak dollar, low labor cost, and low interest rates should cause future earnings to decrease as they become aligned with historic levels. International growth's long-term impact on earnings is more difficult to determine. International growth is clearly the primary factor driving earnings growth, but it is difficult to analyze the international earnings data.

As international sales make up a greater and greater portion of S&P 500 sales, earnings growth and GDP growth will become increasingly disconnected. There are not a lot of books or articles written on international earnings growth of the S&P 500, because of the incomplete reporting by companies. Precise public data is simply not available, which leaves some guesswork for index fund investors.

Disclosure: I/we have no positions in any stocks mentioned, and no plans to initiate any positions within the next 72 hours. I wrote this article myself, and it expresses my own opinions. I am not receiving compensation for it (other than from Seeking Alpha). I have no business relationship with any company whose stock is mentioned in this article.

THE JOURNAL OF FINANCE • VOL. LVII, NO. 4 • AUGUST 2002

Rational Asset Prices

GEORGE M. CONSTANTINIDES*

ABSTRACT

The mean, covariability, and predictability of the return of different classes of financial assets challenge the rational economic model for an explanation. The unconditional mean aggregate equity premium is almost seven percent per year and remains high after adjusting downwards the sample mean premium by introducing prior beliefs about the stationarity of the price–dividend ratio and the (non)-forecastability of the long-term dividend growth and price–dividend ratio. Recognition that idiosyncratic income shocks are uninsurable and concentrated in recessions contributes toward an explanation. Also borrowing constraints over the investors' life cycle that shift the stock market risk to the saving middle-aged consumers contribute toward an explanation.

A central theme in finance and economics is the pursuit of a *unified* theory of the *rate of return* across different classes of financial assets. In particular, we are interested in the *mean*, *covariability*, and *predictability* of the return of financial assets. At the macro level, we study the short-term risk-free rate, the term premium of long-term bonds over the risk-free rate, and the aggregate equity premium of the stock market over the risk-free rate. At the micro level, we study the premium of individual stock returns and of classes of stocks, such as the small-capitalization versus large-capitalization stocks, the "value" versus "growth" stocks, and the past losing versus winning stocks.

The neoclassical rational economic model is a *unified* model that views these premia as the reward to risk-averse investors that *process information rationally* and *have unambiguously defined preferences over consumption* that typically (but not necessarily) belong to the von Neumann–Morgenstern class. Naturally, the theory allows for market incompleteness, market imperfections, informational asymmetries, and learning. The theory also allows for differences among assets for liquidity, transaction costs, tax status, and other institutional factors.

The cause of much anxiety over the last quarter of a century is evidence interpreted as failure of the rational economic paradigm to explain the price level and the rate of return of financial assets both at the macro and micro

*University of Chicago and NBER. I thank John Campbell, Gene Fama, Chris Geczy, Lars Hansen, John Heaton, Rajnish Mehra, L'uboš Pástor, Dick Thaler, and particularly Alon Brav and John Cochrane, for their insightful comments and constructive criticism. Finally, I thank Lior Menzly for his excellent research assistance and insightful comments throughout this project. Naturally, I remain responsible for errors.

George M. Constantinides
President of the American Finance Association
2001

levels. A celebrated example of such evidence, although by no means the only one, is the failure of the *representative-agent* rational economic paradigm to account for the large average premium of the aggregate return of stocks over short-term bonds and the small average return of short-term bonds from the last quarter of the 19th century to the present. Dubbed the "Equity Premium Puzzle" by Mehra and Prescott (1985), it has generated a cottage industry of rational and behavioral explanations of the level of asset prices and their rate of return.

Another example is the large increase in stock prices in the early and middle 1990s, which Federal Reserve Chairman Alan Greenspan decried as "Irrational Exuberance" even before the unprecedented further increase in stock prices and price–dividend ratios in the late 1990s.

My objective is to revisit some of this evidence and explore the extent to which the rational economic paradigm explains the price level and the rate of return of financial assets over the past 100+ years, both at the macro and micro levels.

In Section I, I reexamine the statistical evidence on the size of the unconditional mean of the aggregate equity return and premium. First, I draw a sharp distinction between *conditional*, *short-term forecasts* of the mean equity return and premium and *estimates of the unconditional mean*. I argue that the currently low conditional short-term forecasts of the return and premium do not lessen the burden on economic theory to explain the large unconditional mean equity return and premium, as measured by their sample average over the past 130 years. Second, I argue that even though one may introduce one's own strong prior beliefs and adjust downwards the sample-average estimate of the premium, the unconditional mean equity premium is at least 6 percent per year and the annual Sharpe ratio is at least 32 percent. These numbers are large and call for an economic explanation.

In Section II, I discuss limitations of the current theory to explain empirical regularities. I argue that per capita consumption growth covaries too little with the return of most classes of financial assets and this implies that the observed aggregate equity return, the long-term bond return, and the observed returns of various subclasses of financial assets are *too large*, *too variable*, and *too predictable*.

In the remaining sections, I revisit and examine the extent to which we can explain the asset returns by relaxing the assumptions of complete consumption insurance, perfect markets, and time-separable preferences. As the reader will readily observe—and I offer my apologies—my choice of issues is eclectic and mirrors in part my own research interests.

In Section III, I show that idiosyncratic income shocks concentrated in periods of economic recession play a key role in generating the mean equity premium, the low risk-free rate, and the predictability of returns. I argue that insufficient attention has been paid to the fact that the annual aggregate labor income exceeds annual dividends by a factor of over 20. Labor income is by far the single most important source of household savings and consumption. The shocks to labor income are uninsurable and persistent and arrive with greater frequency during economic contractions. Idiosyncratic income shocks go a long way toward explaining the unconditional moments of asset returns and the predictability of returns. The construct of per capita consumption is largely irrelevant in explaining the behavior of asset returns because idiosyncratic income shocks are averaged out in per capita consumption.

In Section IV, I show that borrowing constraints over the life cycle play an important role in simultaneously addressing the above issues *and* the demand for bonds. I argue that insufficient attention has been paid to the consumers' life cycle consumption and savings decisions in a market with borrowing constraints. These considerations are important in addressing the limited participation of consumers in the capital markets, the irrelevance of the construct of per capita consumption, and the demand for short-term bonds by consumers with moderate risk aversion, given that equities earn on average a large premium over short-term bonds.

In Section V, I discuss the role of limited market participation. In Section VI, I discuss the role of habit persistence in addressing the same class of issues. In Section VII, I conclude that the observed asset returns do not support the case for abandoning the rational economic theory as our null hypothesis. Much more remains to be done to fully exploit the ramifications of the rational asset-pricing paradigm.

I. How Large Is the Equity Premium?

The average premium of the arithmetic rate of return of the S&P Composite Index over the risk-free rate, measured over the last 130 years, is almost *7 percent* and the annual Sharpe ratio is 36 percent. If the equity premium is a stationary process, then the average premium is an unbiased estimate of the *unconditional* mean equity premium. One may introduce one's own prior beliefs and shave about 1 percent off the premium. The premium and the Sharpe ratio are still large and challenge economic theory for an explanation.

In Table I, I report the sample mean of the annual arithmetic aggregate equity return and of the equity premium. I proxy the aggregate equity return with the S&P Composite Index return. I proxy the annual risk-free rate with the rolled-over return on three-month Treasury bills and certificates. The reported real return is CPI-adjusted for inflation. Over the period 1872 to 2000, the sample mean of the real equity return is 8.9 percent and of the premium is 6.9 percent. Over the period 1926 to 2000, the sample mean of the equity return is 9.7 percent and that of the premium is 9.3 percent. Over the postwar period 1951 to 2000, the sample mean of the equity return is 9.9 percent and that of the premium is 8.7 percent. These sample means are large. Siegel (1998, 1999), Ibbotson Associates (2001), Ibbotson and Chen (2001), Dimson, Marsh, and Staunton (2002), Fama and French (2002), Mehra and Prescott (2002), and several others report the sample means of the equity return and premium in the United States and other countries and conclude that they are large. Some differences arise based on the proxy used for the risk-free rate.

Table I

The Equity Return and Premium

This table shows the sample mean and standard deviation of the annualized real arithmetic return on the S&P Composite Index total return series, the sample mean of the real risk-free rate, and the sample mean of the equity premium. The arithmetic rate of return on equity from the beginning to the end of year t is defined as $R_{t+1} = (P_{t+1} + D_{t+1} - P_t)/P_t$, where P_t is the real price of the aggregate equity at the beginning of year t and D_{t+1} is the aggregate real dividend from the beginning to the end of year t. All returns and premia are in percent. Real returns are CPI adjusted. The table also displays the mean annual growth, $(100/T)\{\ln(P_{T+1}/X_{T+1}) - \ln(P_1/X_1)\}$, of the price/$X$ ratio, where X is the dividends, earnings, book equity, or National Income. The pre-1926 S&P Index price series, the CPI series, the earnings series, and the dividends series are obtained from Shiller's database. The S&P Composite Index returns series post-1926 is obtained from the Ibbotson database. For years prior to 1926, the returns are calculated from the S&P 500 Index and dividend series, assuming no dividend reinvestment. The book equity series is obtained from Davis, Fama, and French (2000) and Vuolteenaho (2000). The National Income is obtained from the Bureau of Labor Statistics. The risk-free rate series is the one constructed by Mehra and Prescott (2002) and is based on an annual average nominal return on three-month Treasury certificates and bills.

	1872–2000	1872–1950	1951–2000	1926–2000
Sample mean S&P return	8.87	8.24	9.87	9.70
Std of return	18.49	19.28	17.32	20.33
Sample mean risk-free rate	2.00	2.54	1.15	0.40
Sample mean premium	6.87	5.69	8.72	9.30
Std of premium	19.19	20.23	17.45	20.50
Sharpe ratio	0.36	0.28	0.50	0.45
Mean annual growth of				
Price/dividends	1.18	−0.22	3.39	1.81
Price/earnings	0.71	−0.57	2.73	1.28
Price/book equity	1.18	−0.11	3.18	2.26
Price/national income	NA	NA	1.27	NA

I draw a sharp distinction between *conditional, short-term forecasts of the mean equity return and premium* and *estimates of the unconditional mean*. The conditional forecasts of the mean equity return and premium at the end of the 20th century and the beginning of the 21st are substantially lower than the estimates of the unconditional mean by at least three measures. First, based on evidence that price–dividend and price–earnings ratios forecast aggregate equity returns and that the values of these ratios prevailing at the beginning of the 21st century are well above their historic averages, Campbell and Shiller (1998) and Shiller (2000) forecast a *conditional* equity premium well below its sample average.[1] Second, Claus and Thomas (2001)

[1] Shiller (1984), Campbell and Shiller (1988a, 1988b), and Fama and French (1988) provide early evidence that the aggregate price–dividend and price–earnings ratios forecast aggregate equity returns. Goyal and Welch (1999) argue that the out-of-sample evidence is less convincing. I do not review here the debates and extensions relating to this literature. In the following paragraphs and in Appendix A, I argue that the forecastability results provide little, if any, guidance to my primary goal in this section, the estimation of the unconditional mean equity return.

calculate the expected aggregate equity premium to be a little above 3 percent in the period 1985 to 1998, based on analysts' earnings forecasts. Third, Welch (2001) reports that the mean forecast among finance and economics professors for the one-year conditional equity premium is 3.5 percent in 2001, down from 6 percent in 1997. These findings are important in their own right and relevant in asset allocation.

However, the currently low conditional, short-term forecasts of the equity premium do not necessarily imply that the unconditional estimate of the mean premium is lower than the sample average. Therefore, the low conditional forecasts do not necessarily lessen the burden on economic theory to explain the large sample average of the equity return and premium over the past 130 years.

The predictability of aggregate equity returns by the price–dividend and price–earnings ratios raises the possibility that use of these financial ratios may improve upon the estimates of the *unconditional* mean equity return (and premium) that are based on the sample mean, an approach pursued earlier by Fama and French (2002).[2] Over the period 1872 to 2000, the price–dividend ratio increased by a factor of 4.6 and the price–earnings ratio by a factor of 2.5. Over the period 1926 to 2000, the price–dividend ratio increased by a factor of 3.9 and the price–earnings ratio increased by a factor of 2.6.[3] One may consider adjusting downwards the sample-mean estimate of the unconditional mean return on equity, but it is unclear by how much.

The size of the adjustment ought to relate to the perceived cause of the increase of these financial ratios. In the year 1998, 52 percent of the U.S. adult population held equity either directly or indirectly, compared to 36 percent of the adult population in 1989. This equitization has been brought about by the increased accessibility of information on the stock market, electronic trading, the growth of mutual funds, the growth of defined-contribution pension plans, and demographic changes. Other regime shifts include the advent of the technology/media/telecoms "new economy" and changes in the taxation of dividends and capital gains. Explanations of the price increase that rely on economic models that are less than fully rational include cultural and psychological factors and tap into the rich and burgeoning literature on behavioral economics and finance.[4]

How does one process this information and adjust the sample mean estimate of the unconditional mean return and premium? To address this issue, I denote by $v_t \equiv \ln(P_t/X_t)$ the logarithm of the ratio of the price to the

[2] The estimators employed in Fama and French (2002) and in this section are discussed in Appendix A.

[3] The increase in these financial ratios should be interpreted with caution. The increase in the price–dividend ratio is due in part to an increase in share repurchases and a decrease in the fraction of dividend-paying firms.

[4] I do not provide a systematic review of the offered explanations. Heaton and Lucas (1999), Shiller (2000), and McGrattan and Prescott (2001) provide lucid accounts of a number of these explanations in the context of both rational economic models and models that deviate from full rationality.

normalizing variable X_t, where the normalizing variable stands for the aggregate dividends, earnings, book equity, National Income, or some combination of these and other economic variables.[5] I choose the normalizing variable X_t in a way that I can plausibly assert that the log financial ratio is stationary. Over the sample period of length T years, the mean annual (geometric) growth of the financial ratio P_t/X_t is given by $(v_{T+1} - v_1)/T$. I define the *adjusted* estimator of the unconditional mean of the annual aggregate real equity return as the sample mean return, less some fraction beta of the sample mean annual growth of the financial ratio, $\hat{R}_{SAMPLE} - \beta(v_{T+1} - v_1)/T$. If the equity return and the log financial ratio are stationary processes, then the adjusted estimator is unbiased for any value of beta.[6] However, the assumption of stationarity alone is insufficient to determine the value of beta.

The beta of the most efficient (mean squared error) adjusted estimator is equal to the slope coefficient of the regression of the sample mean return on the sample mean growth of the financial ratio, $(v_{T+1} - v_1)/T$. Since I have only one sample (of length T), I cannot run such a regression and must rely on information *outside the sample* and/or prior beliefs about the underlying economic model. In Appendix A, I present a set of sufficient conditions that imply that the beta of the most efficient estimator within this class of adjusted estimators is equal to one, when the adjustment is based on the price–dividend ratio. In addition to stationarity, the other main conditions are that the price–dividend ratio does not forecast the *long-run* growth in dividends *and* the *long-run* dividend growth does not forecast the price–dividend ratio. Adoption of the stationarity and (non)forecastability conditions requires strong prior beliefs.

In Table I, I report the mean annual growth of various financial ratios. Over the period 1951 to 2000, the mean annual growth of the price–dividend ratio is 3.4, the price–earnings ratio is 2.7, the price–book equity ratio is 3.2, and the price–National Income ratio is 1.3. Even if I subtract the entire mean annual growth of the price–earnings ratio from the sample mean, the adjusted estimate of the unconditional mean premium is 6.0 percent and is large. The corresponding estimate over the 1926 to 2000 period is 8.0 percent.

An alternative approach is to consider the longer sample period 1872 to 2000. Over this period, the mean annual growth of the price–dividend ratio and price–earnings ratio is 1.2 percent and 0.7 percent, respectively. Thus, this type of adjustment is largely a nonissue over the full sample. Essentially, the change in the financial ratios is "amortized" over 129 years and makes little difference in the estimate. Over the full period 1872 to 2000, the sample mean equity premium is 6.9 percent and the annual Sharpe ratio is

[5] The ratio of the stock market value to the National Income is discussed in Mehra (1998).

[6] A caveat is in order: Without additional assumptions, it is unclear what optimality properties (beyond unbiasedness) are associated with this class of estimators. Neither least squares, nor maximum likelihood, nor Bayesian methods motivate this class of estimators without further assumptions.

36 percent. Any adjustment with the average growth of the financial ratios still leaves the unconditional mean premium large and in need of an economic explanation.

II. Limitations of the Current Theory

The neoclassical rational-expectations economic model parsimoniously links the returns of all assets to the per capita consumption growth through the Euler equations of consumption (see Merton (1973), Rubinstein (1976), Lucas (1978), and Breeden (1979)). According to the theory, the risk premia of financial assets are explained by their covariance with per capita consumption growth. However, per capita consumption growth covaries too little with the returns of most classes of financial assets and this creates a whole class of asset-pricing puzzles: the aggregate equity return, the long-term bond return, and the returns of various subclasses of financial assets are too large, too variable, and too predictable. Attempts to leverage the low covariability typically backfire, implying that the observed risk-free rate is too low and has too low variance. I discuss in some depth the aggregate equity puzzle because it exemplifies many of the problems that arise in attempting to explain the premium of any subclass of financial assets.

The covariance of the per capita consumption growth with the aggregate equity return is *positive*. The rational model explains why the aggregate equity premium is positive. However, the covariance is typically one order of magnitude lower than what is needed to explain the premium. Thus, the equity premium is a *quantitative* puzzle.[7]

The equity premium puzzle is *robust*. One may address the problem by testing the Euler equations of consumption or by calibrating the economy. Either way, it is a puzzle. In calibrating an exchange economy, the model cannot generate the first and second unconditional moments of the equity returns. In testing and rejecting the Euler equations of consumption, one abstracts from the market clearing conditions. The rejections tell us that variations in the assumptions on the supply side of the economy do not resolve the puzzle.

The challenge is a *dual puzzle* of the equity premium that is too high and the risk-free rate that is too low relative to the predictions of the model. In calibrating an economy, the strategy of increasing the risk aversion coefficient in order to lever the effect of the problematic low covariance of consumption growth with equity returns increases the predicted risk-free rate

[7] Grossman and Shiller (1981), Hansen and Singleton (1982), Ferson and Constantinides (1991), Hansen and Jagannathan (1991), and many others test and reject the Euler equations of consumption. Mehra and Prescott (1985) calibrate an economy to match the process of consumption growth. They demonstrate that the unconditional mean annual premium of the aggregate equity return over the risk-free rate is, at most, 0.35 percent. This is too low, no matter how one estimates the unconditional mean equity premium. Weil (1989) stresses that the puzzle is a dual puzzle of the observed too high equity return and too low risk-free rate.

and aggravates the risk-free-rate puzzle. In testing the Euler equations of consumption, the rejections are strongest when the risk-free rate is included in the set of test assets.

Several generalizations of essential features of the model have been proposed to mitigate its poor performance. They include alternative assumptions on preferences,[8] modified probability distributions to admit rare but disastrous market-wide events,[9] incomplete markets,[10] and market imperfections.[11] They also include a better understanding of data problems such as limited participation of consumers in the stock market,[12] temporal aggregation,[13] and the survival bias of the U.S. capital market.[14] Many of these generalizations contribute in part toward our better understanding of the economic mechanism that determines the pricing of assets. I refer the reader to the excellent reviews in the textbooks by Campbell, Lo, and MacKinlay (1997) and Cochrane (2001), and in the articles by Cochrane and Hansen (1992), Kocherlakota (1996), Cochrane (1997), Campbell (2001, 2002), and Mehra and Prescott (2002).

III. Idiosyncratic Income Shocks and Incomplete Markets

A. The Role of Idiosyncratic Income Shocks

In economic recessions, investors are exposed to the double hazard of stock market losses and job loss. Investment in equities not only fails to hedge the risk of job loss but also accentuates its implications. Investors require a hefty equity premium in order to be induced to hold equities. In sum, this is the argument that I formalize below and address the predictability of asset returns and their unconditional moments.

The observed correlation of per capita consumption growth with stock returns is low. Over the years, I have grown skeptical of how meaningful an economic construct *aggregate* (as opposed to *disaggregate*) consumption is,

[8] For example, Abel (1990), Constantinides (1990), Epstein and Zin (1991), Ferson and Constantinides (1991), Benartzi and Thaler (1995), Campbell and Cochrane (1999), Anderson, Hansen, and Sargent (2000), Bansal and Yaron (2000), and Boldrin, Christiano, and Fisher (2001).

[9] The merits of this explanation are discussed in Mehra and Prescott (1988) and Rietz (1988).

[10] For example, Bewley (1982), Mehra and Prescott (1985), Mankiw (1986), Constantinides and Duffie (1996), Heaton and Lucas (1996), Storesletten, Telmer, and Yaron (2001), Brav, Constantinides, and Geczy (2002), and Krebs (2002).

[11] For example, Aiyagari and Gertler (1991), Danthine, Donaldson, and Mehra (1992), He and Modest (1995), Bansal and Coleman (1996), Heaton and Lucas (1996), Daniel and Marshall (1997), and Constantinides, Donaldson, and Mehra (2002a).

[12] Mankiw and Zeldes (1991), Brav and Geczy (1995), Attanasio, Banks, and Tanner (2002), Brav et al. (2002), and Vissing-Jorgensen (2002).

[13] Heaton (1995), Lynch (1996), and Gabaix and Laibson (2001).

[14] See Brown, Goetzmann, and Ross (1995). However, Jorion and Goetzmann (1999, Table 6) find that the average real capital gain rate of a U.S. equities index exceeds the average rate of a global equities index that includes both markets that have and have not survived by merely one percent per year.

and how hard we should push aggregate or per capita consumption to explain returns. At a theoretical level, aggregate consumption is a meaningful economic construct if the market is complete or effectively so.[15] In a complete market, heterogeneous households are able to equalize, state by state, their marginal rate of substitution. The equilibrium in a heterogeneous-household, full-information economy is isomorphic in its pricing implications to the equilibrium in a representative-household, full-information economy, if households have von Neumann–Morgenstern preferences.[16] The strong assumption of market completeness is indirectly built into asset pricing models in finance and neoclassical macroeconomic models through the assumption of the existence of a representative household.

Bewley (1982), Mehra and Prescott (1985), and Mankiw (1986) suggest the potential of enriching the asset-pricing implications of the representative-household paradigm, by relaxing the assumption of complete markets.[17] Constantinides and Duffie (1996) find that incomplete markets substantially enrich the implications of the representative-household model. Their main result is a proposition demonstrating, by construction, the existence of household income processes, consistent with given aggregate income and dividend processes, such that equilibrium equity and bond price processes match the given equity and bond price processes.

The theory requires that the idiosyncratic income shocks must have three properties in order to explain the returns on financial assets. First, they must be *uninsurable*. If the income shocks can be insured, then the household consumption growth is equal, state by state, to the aggregate consumption growth, and household consumption growth cannot do better than aggregate consumption growth in explaining the returns. Second, the income shocks must be *persistent*. If the shocks are transient, then households can smooth their consumption by borrowing or by drawing down their savings.[18] Third, the income shocks must be *heteroscedastic*, with *countercyclical* conditional variance.

A good example of a major uninsurable income shock is job loss. Job loss is *uninsurable* because unemployment compensation is inadequate. Layoffs have *persistent* implications on household income, even though the laid-off

[15] The market is effectively complete when all households have preferences that imply one-fund or two-fund separation.

[16] See Negishi (1960), Constantinides (1982), and Mehra and Prescott (1985, an unpublished earlier draft).

[17] There is an extensive literature on the hypothesis of complete consumption insurance. See Cochrane (1991), Mace (1991), Altonji, Hayashi, and Kotlikoff (1992), and Attanasio and Davis (1997).

[18] Aiyagari and Gertler (1991) and Heaton and Lucas (1996) find that consumers facing *transient* shocks come close to the complete-markets rule of complete risk sharing even with transaction costs and/or borrowing costs, *provided that the supply of bonds is not restricted to an unrealistically low level.*

workers typically find another job quickly.[19] Layoffs are *countercyclical* as they are more likely to occur in recessions.

The first implication of the theory is an explanation of the countercyclical behavior of the equity risk premium: The risk premium is highest in a recession because the stock is a poor hedge against the uninsurable income shocks, such as job loss, that are more likely to arrive during a recession.

The second implication is an explanation of the unconditional equity premium puzzle: Even though per capita consumption growth is poorly correlated with stocks returns, investors require a hefty premium to hold stocks over short-term bonds because stocks perform poorly in recessions, when the investor is most likely to be laid off.

Since the proposition demonstrates the existence of equilibrium in frictionless markets, it implies that the Euler equations of household (but not necessarily of per capita) consumption must hold. Furthermore, *since the given price processes have embedded in them whatever predictability of returns by the price–dividend ratios, dividend growth rates, and other instruments that the researcher cares to ascribe to returns, the equilibrium price processes have this predictability built into them by construction.*

B. Empirical Evidence and Generalizations

Brav et al. (2002) provide empirical evidence of the importance of uninsurable idiosyncratic income risk on pricing. They estimate the RRA coefficient and test the set of Euler equations of *household* consumption on the premium of the value-weighted and the equally weighted market portfolio return over the risk-free rate, and on the premium of value stocks over growth stocks.[20] They do not reject the Euler equations of *household* consumption with RRA coefficient between two and four, although they reject the Euler equations of per capita consumption with any value of the RRA coefficient. A RRA coefficient between two and four is economically plausible.

Open questions remain that warrant further investigation. According to the theory in Constantinides and Duffie (1996), periods with frequent and large uninsurable idiosyncratic income shocks are associated with both dispersed cross-sectional distribution of the household consumption growth and low stock returns. An interesting empirical question is *which moments* of the

[19] The empirical evidence is sensitive to the model specification. Heaton and Lucas (1996) model the income process as *univariate* and provide empirical evidence from the Panel Study on Income Dynamics (PSID) that the idiosyncratic income shocks are transitory. Storesletten et al. (2001) model the income process as *bivariate* and provide empirical evidence from the PSID that the idiosyncratic income shocks have a highly persistent component that becomes more volatile during economic contractions. Storesletten, Telmer, and Yaron (2000) corroborate the latter evidence by studying household consumption over the life cycle.

[20] In related studies, Jacobs (1999) studies the PSID database on food consumption; Cogley (2002) and Vissing-Jorgensen (2002) study the CEX database on broad measures of consumption; Jacobs and Wang (2001) study the CEX database by constructing synthetic cohorts; and Ait-Sahalia, Parker, and Yogo (2001) instrument the household consumption with the purchases of certain luxury goods.

cross-sectional distribution of the household consumption growth capture the dispersion. Brav et al. (2002) find that, in addition to the mean and variance, the *skewness* of the cross-sectional distribution is important in explaining the equity premium.

Krebs (2002) provides a theoretical justification as to why it is possible that neither the variance nor the skewness, but higher moments of the cross-sectional distribution are important in explaining the equity premium. He extends the Constantinides and Duffie (1996) model that has only *lognormal* idiosyncratic income shocks by introducing *rare* idiosyncratic income shocks that drive consumption close to *zero*. In his model, the conditional variance and skewness of the idiosyncratic income shocks are nearly constant over time. Despite this, Krebs demonstrates that the original proposition of Constantinides and Duffie remains valid, that is, there exist household income processes, consistent with given aggregate income and dividend processes, such that equilibrium equity and bond price processes match the given equity and bond price processes. Essentially, he provides a theoretical justification as to why it may be hard to empirically detect the rare but catastrophic shocks in the low-order cross-sectional moments of household consumption growth. In Appendix B, I present an example based on Krebs (2002).

A promising direction for future research is to address the relation between the equity return and the higher-order cross-sectional moments of household consumption with Monte Carlo methods. Another promising direction is to instrument the hard-to-observe time-series changes in the cross-sectional distribution with Labor Bureau statistics.

IV. The Life Cycle and Borrowing Constraints

A. Borrowing Constraints over the Life Cycle

Borrowing constraints provide an endogenous partial explanation for the limited participation of young consumers in the stock market. Constantinides et al. (2002a) construct an overlapping-generations exchange economy in which consumers live for three periods. In the first period, a period of human capital acquisition, the consumer receives a relatively low endowment income. In the second period, the consumer is employed and receives wage income subject to large uncertainty. In the third period, the consumer retires and consumes the assets accumulated in the second period. The key feature is that the bulk of the *future* income of the young consumers is derived from their wages forthcoming in their middle age, while the *future* income of the middle-aged consumers is derived primarily from their savings in equity and bonds.

The young would like to invest in equity, given the observed large equity premium. However, they are unwilling to decrease their current consumption in order to save by investing in equity, because the bulk of their lifetime income is derived from their wages forthcoming in their middle age. They would like to borrow, but the borrowing constraint prevents them from doing so. Human capital alone does not collateralize major loans in modern economies for reasons of moral hazard and adverse selection. The model explains why many consumers do not participate in the stock market in the early phase of their life cycle.

The future income of the middle-aged consumers is derived from their current savings in equity and bonds. Therefore, the risk of holding equity and bonds is concentrated in the hands of the middle-aged saving consumers. This concentration of risk generates the high equity premium and the demand for bonds, in addition to the demand for equity, by the middle-aged.[21] The model recognizes and addresses *simultaneously*, at least in part, the *equity premium*, the *limited participation* in the stock market, and the *demand for bonds*.

The model serves as a useful laboratory to address a range of economic issues. Campbell et al. (2001), and Constantinides, Donaldson, and Mehra (2001) address the cost of Social Security reform. Storesletten et al. (2001) explore the interaction of life-cycle effects and the uninsurable wage income shocks and find that the interaction plays an important role in explaining asset returns. Heaton and Lucas (1999) explore whether changes in market participation patterns account for the recent rise in stock prices and find that they do not.

B. Utility of Wealth—An Old Folks' Tale

The low covariance of the growth rate of aggregate consumption with equity returns is a major stumbling block in explaining the mean aggregate equity premium and the cross section of the asset returns, in the context of a representative-consumer economy with time separable preferences. Mankiw and Shapiro (1986) find that the market beta often explains asset returns better than the consumption beta does. Over the years, a number of different economic models have been proposed that effectively increase the covariance of equity returns with the growth rate of aggregate consumption, by proxying the growth rate of aggregate consumption with the aggregate stock market return in the Euler equations of consumption.[22]

I present an old folks' tale, introduced in Constantinides, Donaldson, and Mehra (2002a, 2002b), that accomplishes this goal without introducing Epstein–Zin (1991) preferences or preferences defined directly over wealth.

[21] See also the discussion in the related papers by Bodie, Merton, and Samuelson (1992), Jagannathan and Kocherlakota (1996), Bertaut and Haliassos (1997), Cocco, Gomes, and Maenhout (1999), and Storesletten et al. (2001).

[22] Friend and Blume (1975) explain the mean equity premium with low RRA coefficient by assuming a single-period economy in which the end-of-period consumption inevitably equals the end-of-period wealth. Epstein and Zin (1991) introduce a recursive preference structure that emphasizes the timing of the resolution of uncertainty. Even though the preferences are defined over consumption alone, the stock market return enters directly in the Euler equations of consumption. Bakshi and Chen (1996) introduce a set of preferences defined over consumption and wealth—the spirit of capitalism—that also have the effect of introducing the stock market return in the Euler equations of consumption.

Old folks who are rich enough to be nontrivial investors in the capital markets care about their wealth just as much as younger folks do, even though the state of their health and their medical expenses account for their consumption patterns better than fluctuations of their wealth do. This simple observation takes us a long way toward understanding why the stock market return does a better job than the growth of aggregate consumption does in explaining asset returns.

In the context of an overlapping-generations economy, the major investors in the market are the middle-aged households at the saving phase of their life cycle. These households save with the objective to maximize the utility of their "consumption" in their middle and old age. The insight here is that "consumption" of the old consists of two components, direct consumption, c_D; and the "joy of giving," c_B, in the form of *inter vivos* gifts and *post mortem* bequests. Since the old households' direct consumption is constrained by the state of their health, the correlation between the direct consumption of the old and the stock market return is *low*, a prediction that is borne out empirically. Therefore, the balance of the old households' wealth, c_B, is *a fortiori* highly correlated with the stock market return. In terms of a utility function of consumption at the old age, $u(c_D) + v(c_B)$, that is separable over direct consumption and bequests, the model predicts an Euler equation of consumption with marginal utility at the old age given by $v'(c_B)$ and not by $u'(c_D)$, where c_B is proxied by the stock market value.

This model remains to be tested. Nevertheless, it reinforces the general point that per capita consumption measures neither the total consumption of the marginal investor in the stock market nor that part of the marginal investor's consumption that is unconstrained by health and medical considerations.

V. Limited Stock Market Participation

Limited stock market participation is another potential culprit in understanding why models of per capita consumption do a poor job in explaining returns. Whereas we understood all along that many households whose consumption is counted in the measure of per capita consumption do not hold stocks, it took a paper by Mankiw and Zeldes (1991) to point out that the emperor has no clothes.[23] Even though 52 percent of the U.S. adult population held stock either directly or indirectly in 1998, compared to 36 percent in 1989, stockholdings remain extremely concentrated in the hands of the wealthiest few. Furthermore, wealthy entrepreneurs may be inframarginal in the stock market if their wealth is tied up in private equity.

[23] Since then, several papers have studied the savings and portfolio composition of households, stratified by income, wealth, age, education, and nationality. See Blume and Zeldes (1993), Haliassos and Bertaut (1995), Heaton and Lucas (1999, 2000), Poterba (2001), and the collected essays in Guiso, Haliassos, and Jappelli (2001).

Mankiw and Zeldes (1991) calculate the per capita food consumption of a subset of households, designated as *asset holders* according to a criterion of asset holdings above some threshold. They find that the implied RRA coefficient decreases as the threshold is raised. Brav and Geczy (1995) confirm their result by using the nondurables and services per capita consumption, reconstructed from the Consumer Expenditure Survey (CEX) database. Attanasio et al. (2002), Brav et al. (2002), and Vissing-Jorgensen (2002) find some evidence that per capita consumption growth can explain the equity premium with a relatively high value of the RRA coefficient, once we account for limited stock market participation. However, Brav et al. point out that the statistical evidence is weak and the results are sensitive to experimental design.

Limited stock market participation is a fact of life and empirical tests of the Euler equations of consumption should account for it. However, my interpretation of the empirical results is that recognition of limited stock market participation alone is insufficient to explain the returns on assets. Essentially, *the subset of households that are marginal in the stock market are still subject to uninsurable idiosyncratic income risk and we should take that into account also in attempting to explain asset returns.*

VI. Habit Persistence

Habit persistence has a long tradition in economic theory, dating back to Marshall (1920) and Duesenberry (1949). It is the property of preferences that an increase in consumption increases the marginal utility of consumption at adjacent dates relative to the marginal utility of consumption at distant ones. Building on earlier work by Ryder and Heal (1973) and Sundaresan (1989), I demonstrate in Constantinides (1990) that habit persistence can, in principle, reconcile the high mean equity premium with the low variance of consumption growth and with the low covariance of consumption growth with equity returns. Habit persistence lowers the intertemporal elasticity of substitution in consumption, given the risk aversion. The mean equity premium is equal to the covariance of consumption growth with equity returns, divided by this elasticity. Therefore, given the risk aversion, habit persistence lowers the elasticity and raises the mean equity premium.[24]

There are several interesting variations of the above class of preferences. Pollak (1970) discusses a model of *external* habit persistence in which the consumer does not take into account the effect of current consumption on future preferences. Abel (1990) and Campbell and Cochrane (1999) address

[24] Ferson and Constantinides (1991) test the special case of the linear habit model in which the habit depends only on the *first* lag of own consumption and report that the habit model performs better than the time-separable model and that the habit persistence parameter is economically and statistically significant. See also Hansen and Jagannathan (1991) and Heaton (1995).

the equity premium in the context of models with external habit persistence. In particular, the latter introduce a *nonlinear* specification of habit, reverse-engineered to keep the variability of the interest rate low. The large average equity premium, the predictability of long-horizon returns, and the behavior of equity prices along the business cycle are induced by a volatile RRA coefficient that has the value of 80 in the steady state and much higher still in economic recessions. Calibrated with the actual history of aggregate consumption, the model hits the aggregate price–dividend ratio in a number of periods but misses it in the 1950s and 1990s.

A promising direction for future research is to endogenize the currently ad hoc specification of the nonlinear habit. Another direction is to address the predictability of asset returns and their behavior along the business cycle in a model that benefits from the added flexibility of the nonlinear specification of habit but keeps risk aversion low and credible with the specification of habit to be *internal*.

Empirical tests of consumption-based models that incorporate habit persistence and aimed at explaining asset returns produce mixed results.[25] It is hardly surprising that the results on both the habit and the external habit persistence models are mixed. The National Income and Product Accounts (NIPA) per capita consumption series is an imperfect proxy of the consumption of investors that are marginal in the capital markets, given the earlier-identified problems of incomplete consumption insurance, limited participation of households in the capital markets, borrowing constraints, and the exclusion of bequests from the definition of consumption. Both NIPA per capita consumption and consumption surplus over habit have low covariance with asset returns. Nonlinear refinements in the definition of habit do not remedy the problem of low covariance with asset returns. Habit persistence may well gain in empirical relevance in explaining asset returns, once we correctly measure the consumption of the unconstrained marginal investors in the capital markets.

Habit persistence is already gaining ground as an ingredient of economic models addressing a diverse set of economic problems beyond asset pricing, including the consumption–saving behavior and the home-equity puzzle. Habit persistence is a sensible property of preferences. It is also a property that allows for the separate specification of the RRA coefficient and the intertemporal elasticity of substitution within the class of von Neuman–Morgenstern preferences.

[25] Ferson and Harvey (1992) report positive results for the linear external habit model. Wachter (2001) reports that long lags of consumption growth predict the short-term interest rate, as implied by the nonlinear external habit model. Li (2001) reports that in both the linear and the nonlinear external habit models, the surplus consumption over habit has limited success in explaining the time series of the premia of stock and bond portfolios. Menzly, Santos, and Veronesi (2001) develop an external habit model and report that it helps explain the cross section of asset returns.

VII. Concluding Remarks

I examine the observed asset returns and conclude that the evidence does not support the case for abandoning the rational economic model. I argue that the standard model is greatly enhanced by relaxing some of its assumptions. In particular, I argue that we go a long way toward addressing market behavior by recognizing that consumers face uninsurable and idiosyncratic income shocks, for example, the loss of employment. The prospect of such events is higher in economic downturns and this observation takes us a long way toward understanding both the unconditional moments of asset returns and their variation along the business cycle.

I also argue that life-cycle considerations are important and often overlooked in finance. Borrowing constraints become important when placed in the context of the life cycle. The fictitious representative consumer that holds all the stock market and bond market wealth does not face credible borrowing constraints. Young consumers, however, do face credible borrowing constraints. I trace their impact on the equity premium, the demand for bonds—Who holds bonds if the equity premium is so high?—and on the limited participation of consumers in the capital markets.

Finally, I argue that relaxing the assumption of convenience that preferences are time separable drives a wedge between the preference properties of risk aversion and intertemporal elasticity of substitution, within the class of von Neumann–Morgenstern preferences. Further work along these lines may enhance our understanding of the price behavior along the business cycle with credibly low risk-aversion coefficient.

I believe that the integration of the notions of *incomplete markets*, the *life cycle*, *borrowing constraints*, and other sources of *limited stock market participation* is a promising vantage point from which to study the prices of asset and their returns both theoretically and empirically within the class of rational asset-pricing models.

At the same time, I believe that specific deviations from rationality in the agents' choices and in the agents' processing of information potentially enhance the realism and economic analysis of certain phenomena on a case-by-case basis.[26] However, several examples of apparent deviation from rationality may be reconciled with the rational economic paradigm, once we recognize that rational investors have incomplete knowledge of the fundamental structure of the economy and engage in learning.[27] In any case, the collection of these deviations from rationality does not yet amount to a new economic paradigm that challenges the rational economic model.

It has been more than 60 years since Keynes (1936) wrote about animal spirits, and 15 since Shiller (1984) wrote about noise traders and DeBondt and Thaler (1985) wrote about stock market overreaction. I have yet to see an *unambiguously articulated* set of principles that emerges from the kalei-

[26] Barberis and Thaler (2002) and Hirshleifer (2001) provide excellent reviews of this literature.

[27] Brav and Heaton (2002) provide excellent discussion of these issues.

doscope of these clinical investigations and that is put forth as an alternative to the rational economic paradigm. Serious scholars are keenly aware of this criticism and hard at work to address it. Until such a paradigm is put forth *and* is empirically vindicated, the rational economic paradigm remains our principal guide to economic behavior.

Appendix A. Estimation of the Unconditional Mean Return on Equity

I define the adjusted estimator of the unconditional mean of the annual aggregate arithmetic real return on equity as

$$\hat{R}_x \equiv T^{-1}\sum_{t=1}^{T} R_{t+1} - \beta T^{-1}(v_{T+1} - v_1) = \hat{R}_{SAMPLE} - \beta T^{-1}(v_{T+1} - v_1). \quad \text{(A1)}$$

The term $v_t \equiv \ln(P_t/X_t)$ is the logarithm of the price of aggregate equity, normalized with the variable X_t, where X_t stands for the aggregate dividends, earnings, book equity, National Income, or some other economic variable.

I assume that R_t and v_t are stationary processes. Then $E[v_{T+1} - v_1] = 0$ and $\hat{R}_x$ is an *unbiased* estimator of the unconditional mean equity return. Note that the assumption of stationarity alone does not determine the value of the parameter beta that provides the most efficient estimator of the unconditional mean equity return. The variance of the estimator $\hat{R}_x$ is

$$\begin{aligned}\operatorname{var}(\hat{R}_x) = {} & \operatorname{var}(\hat{R}_{SAMPLE}) - 2\beta \operatorname{cov}(\hat{R}_{SAMPLE}, T^{-1}(v_{T+1} - v_1)) \\ & + \beta^2 \operatorname{var}(T^{-1}(v_{T+1} - v_1))\end{aligned} \quad \text{(A2)}$$

and is minimized when beta is set equal to

$$\beta^* = \frac{\operatorname{cov}(\hat{R}_{SAMPLE}, T^{-1}(v_{T+1} - v_1))}{\operatorname{var}(T^{-1}(v_{T+1} - v_1))}. \quad \text{(A3)}$$

The beta of the most efficient (mean squared error) estimator is equal to the slope coefficient of the regression of $\hat{R}_{SAMPLE}$ on $T^{-1}(v_{T+1} - v_1)$.

Since I have only one sample of length T, I cannot run such a regression and must rely on information *outside the sample* and/or prior beliefs about the underlying economic model. Essentially, within the sample of length T, I can examine the high-frequency behavior of the joint time series R_t and v_t, but I need to assert my prior beliefs on how these findings relate to the behavior of the joint time series at the T-year frequency.

For example, consider the case in which v_t stands for the log price–dividend ratio. Since a high price–dividend ratio forecasts *in-sample* low long-horizon returns, it is a plausible prior belief that it also forecasts low T-horizon returns, $\operatorname{cov}(\hat{R}_{SAMPLE}, v_1) < 0$, for $T = 50$ years (1951 to 2000) or

$T = 129$ years (1872 to 2000). It is also a plausible prior belief that periods of high returns are not followed by low price–dividend ratios, that is, it is plausible to believe that $\text{cov}(\hat{R}_{SAMPLE}, v_{T+1}) \gtrsim 0$. Then equation (A3) implies that the beta of the most efficient estimator is positive.

I present a set of sufficient (but not necessary) conditions that imply that the beta of the most efficient estimator in the class $\hat{R}_x$ equals one. Let v_t stand for the log price–dividend ratio and assume the following: (1) the returns and the price–dividend ratio are stationary, (2) the price–dividend ratio does not forecast the growth in dividends, (3) dividend growth does not forecast the price–dividend ratio, (4) the price–dividend ratio does not forecast the difference in the conditional variance of the capital gain rate and the dividend growth rate, and (5) the difference in the conditional variance of the capital gain rate and the dividend growth rate does not forecast the price–dividend ratio. To prove the claim, I use a Taylor-series expansion:

$$\Delta v_{t+1} = \Delta P_{t+1}/P_t - \Delta D_{t+1}/D_t - k_{t+1} \tag{A4}$$

where

$$k_{t+1} \equiv (\Delta P_{t+1}/P_t)^2/2 - (\Delta D_{t+1}/D_t)^2/2$$

and write the sample mean of the arithmetic return as

$$\begin{aligned}\hat{R}_{SAMPLE} &\equiv T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta P_{t+1}/P_t\}\\ &= T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta D_{t+1}/D_t + k_{t+1} + \Delta v_{t+1}\} \qquad \text{(A5)}\\ &= T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta D_{t+1}/D_t + k_{t+1}\} + T^{-1}(v_{T+1} - v_1).\end{aligned}$$

I substitute the value of $\hat{R}_{SAMPLE}$ from equation (A5) into equation (A3) and obtain the result that the variance of the estimator is minimized when the value of beta is one:

$$\begin{aligned}\beta^* = {} & \frac{\text{cov}\left(\sum_{t=1}^{T} D_{t+1}/P_t, (v_{T+1} - v_1)\right)}{\text{var}(v_{T+1} - v_1)} + \frac{\text{cov}\left(\sum_{t=1}^{T} \Delta D_{t+1}/D_t, v_{T+1}\right)}{\text{var}(v_{T+1} - v_1)}\\ & - \frac{\text{cov}\left(\sum_{t=1}^{T} \Delta D_{t+1}/D_t, v_1\right)}{\text{var}(v_{T+1} - v_1)} + \frac{\text{cov}\left(\sum_{t=1}^{T} k_{t+1}, (v_{T+1} - v_1)\right)}{\text{var}(v_{T+1} - v_1)} + 1 \qquad \text{(A6)}\\ = {} & 1.\end{aligned}$$

The first term in equation (A6) is approximately zero because the stationarity of the price–dividend ratio implies

$$\operatorname{cov}\left(\sum_{t=1}^{T} D_{t+1}/P_t, v_{T+1}\right) \approx \operatorname{cov}\left(\sum_{t=1}^{T} D_{t+1}/P_t, v_1\right). \tag{A7}$$

The second term in equation (A6) is zero because, by assumption, the dividend growth rate does not forecast the price–dividend ratio. The third term is zero because, by assumption, the price–dividend ratio does not forecast the dividend growth. Finally, the fourth term is zero because, by assumption, the price–dividend ratio does not forecast and is not forecasted by the difference of the conditional variance of the capital gain rate and the dividend growth rate.

Thus, when X_t stands for the dividends and conditions (1)–(5) hold, the minimum variance estimator in the class of estimators given by equation (A1) is

$$\begin{aligned}\hat{R}_D &= \hat{R}_{SAMPLE} - T^{-1}(v_{T+1} - v_1) \\ &= T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta D_{t+1}/D_t\} + T^{-1}\sum_{t=1}^{T} k_{t+1}.\end{aligned} \tag{A8}$$

Fama and French (2002) report adjusted estimates of the unconditional mean return (and premium) based on the fundamentals dividends and earnings. Specifically, their estimate of the expected stock return based on the dividend growth model is equivalent to $T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta D_{t+1}/D_t\}$ and their biased-adjusted estimate is equivalent to $T^{-1}\sum_{t=1}^{T}\{D_{t+1}/P_t + \Delta D_{t+1}/D_t\} + T^{-1}\sum_{t=1}^{T} k_{t+1}$. Ibbotson and Chen (2001) also report adjusted estimates of the unconditional mean return (and premium) based on dividends, income, earnings, payout ratio, book equity, and National Income.

Appendix B. Extension of the Constantinides and Duffie (1996) Model

I illustrate an extension of the Constantinides and Duffie (1996) model along the lines of Krebs (2002). The extension provides theoretical justification as to why it may be hard to detect empirically in the low-order cross-sectional moments of household consumption growth the rare but catastrophic shocks that play a major role in driving asset prices.

The ith household's consumption, $c_{i,t}$, follows the process

$$\frac{c_{it}}{c_{i,t-1}} = \frac{c_t}{c_{t-1}} X_{i,t}\eta_{i,t}. \tag{B1}$$

The random variables $\{\eta_{i,t}\}$ have the following properties: Distinct subsets of $\{\eta_{i,t}\}$ are independent; for all i and t; $\eta_{i,t}$ is independent of $c_{t-1}, c_t, c_{i,t-1}, X_{i,t}$, and the asset prices; and $E[\eta_{i,t}] = 1$. Since the random variables $\{\eta_{i,t}\}$ are independent of the asset prices, they do not contribute to the equity premium. One may choose to view them as observation error, but does not have to.

In the Constantinides and Duffie (1996) model, the idiosyncratic income shocks are lognormal: $X_{i,t} = e^{b_t \epsilon_{i,t} - b^2/2}$ with $\epsilon_{i,t}$ normal and $\eta_{i,t} \equiv 1$. The conditional variance, b_t^2, explains the risk premia because it is modeled as countercyclical and correlated with the stock returns. Whereas Brav et al. (2002) find that the pricing kernel $I^{-1} \sum_{i=1}^{I} (c_{i,t}/c_{i,t-1})^{-\alpha}$ goes a long way toward explaining the equity premium and the value-versus-growth premium, they also find little evidence that the conditional variance, b_t^2, is correlated with stock returns, or indeed whether the time series of this variance has any discernible pattern relative to the business cycle. I build this feature in the model by choosing a binomial distribution for $X_{i,t}$.

I assume that the random variables $\{X_{it}\}$ have the following properties: Distinct subsets of $\{X_{it}\}$ are independent; for all i and t, X_{it} is independent of $c_{t-1}, c_t, c_{i,t-1}$ and $X_{i,t-1}$; and X_{it} has the following binomial distribution:

$$\begin{aligned} X_{i,t} &= \frac{1 - y_t^{-\alpha^{-1}} \pi^{1+\alpha^{-1}}}{1 - \pi}, \quad \text{with probability } 1 - \pi \\ &= y_t^{-\alpha^{-1}} \pi^{\alpha^{-1}}, \quad \text{with probability } \pi, \end{aligned} \tag{B2}$$

where $0 < \pi \ll 1$, and α is the constant RRA coefficient. The variable $y_t, y_t > 0$ is defined shortly. Since

$$E\left[\frac{c_{it}}{c_{i,t-1}} \middle| y_t, \frac{c_t}{c_{t-1}} \right] = \frac{c_t}{c_{t-1}}, \tag{B3}$$

arguments along the lines in Constantinides and Duffie (1996) identify c_t as the per capita consumption.

The time-t expectation of the ith household's marginal rate of substitution, conditional on $\{c_t/c_{t-1}, y_t\}$, is

$$\begin{aligned} &E\left[e^{-\rho} \left(\frac{c_{it}}{c_{i,t-1}} \right)^{-\alpha} \middle| \frac{c_t}{c_{t-1}}, y_t \right] \\ &\quad = e^{-\rho} \left(\frac{c_t}{c_{t-1}} \right)^{-\alpha} \{(1 - \pi)^{1+\alpha} (1 - y_t^{-\alpha^{-1}} \pi^{1+\alpha^{-1}})^{-\alpha} + y_t\} E[\eta_{i,t}^{-\alpha}] \\ &\quad \approx e^{-\rho} \left(\frac{c_t}{c_{t-1}} \right)^{-\alpha} (1 + y_t) E[\eta_{i,t}^{-\alpha}], \quad \text{for } \pi \ll 1. \end{aligned} \tag{B4}$$

I define the variable y_t implicitly with the equation

$$e^{-\rho}\left(\frac{c_t}{c_{t-1}}\right)^{-\alpha}(1+y_t)E[\eta_{i,t}^{-\alpha}] = M_t, \tag{B5}$$

where M_t is the pricing kernel that supports the given joint process of aggregate income, asset prices, and dividends. By construction, it follows that any individual household's marginal rate of substitution, $e^{-\rho}(c_{it}/c_{i,t-1})^{-\alpha}$, supports the given joint process of aggregate income, asset prices, and dividends.

Finally, I demonstrate that the variance, skewness, and higher moments of the cross-sectional distribution of the households' consumption growth need not bear any relationship to asset returns and the business cycle. This is despite the fact that each individual household's marginal rate of substitution supports the given joint process of aggregate income, asset prices, and dividends.

The Nth central moment, $N \geq 1$, of the households' logarithmic consumption growth is the sum of the Nth central moments of $\ln(c_t/c_{t-1})$, $\ln(X_{i,t})$, and $\ln(\eta_{i,t})$, given the assumed independence of $c_t/c_{t-1}, X_{i,t}$, and $\eta_{i,t}$. It is easily shown that

$$\lim_{\pi\to 0} E[(\ln X_{i,t})^N] = 0, \quad N \geq 1. \tag{B6}$$

If the probability of the idiosyncratic consumption shocks is sufficiently low, $\pi \ll 1$, the central moments of the households' consumption growth are driven by the corresponding central moments of the per capita consumption growth and $\eta_{i,t}$. These moments need not bear any pattern relating to the business cycle and need not be correlated in any particular way with the asset returns. Despite this, each individual household's marginal rate of substitution supports the given joint process of aggregate income, asset prices, and dividends. The illustration explains why it may be empirically difficult or infeasible to detect the idiosyncratic consumption shocks in the cross-sectional moments of household consumption growth.

REFERENCES

Abel, Andrew B., 1990, Asset prices under habit formation and catching up with the Joneses, *American Economic Review Papers and Proceedings* 80, 38–42.

Ait-Sahalia, Yacine, Jonathan A. Parker, and Motohiro Yogo, 2001, Luxury goods and the equity premium, Working paper 8417, NBER.

Aiyagari, Rao S., and Mark Gertler, 1991, Asset returns with transactions costs and uninsured individual risk, *Journal of Monetary Economics* 27, 311–331.

Altonji, Joseph G., Fumio Hayashi, and Laurence J. Kotlikoff, 1992, Is the extended family altruistically linked? *American Economic Review* 82, 1177–1198.

Anderson, Evan W., Lars Peter Hansen, and Thomas J. Sargent, 2000, Robustness, detection and the price of risk, Working paper, University of Chicago.

Attanasio, Orazio P., James Banks, and Sarah Tanner, 2002, Asset holding and consumption volatility, *Journal of Political Economy* 110, forthcoming.

Attanasio, Orazio P., and Steven J. Davis, 1997, Relative wage movements and the distribution of consumption, *Journal of Political Economy* 104, 1227–1262.

Bakshi, Gurdip, and Zhiwu Chen, 1996, The spirit of capitalism and stock market prices, *American Economic Review* 86, 133–157.

Bansal, Ravi, and John W. Coleman, 1996, A monetary explanation of the equity premium, term premium, and risk-free rate puzzles, *Journal of Political Economy* 104, 1135–1171.

Bansal, Ravi, and Amir Yaron, 2000, Risks for the long run: A potential resolution of asset pricing puzzles, Working paper 8059, NBER.

Barberis, Nicholas, and Richard Thaler, 2002, A survey of behavioral finance, in George M. Constantinides, Milton Harris, and Rene Stulz, eds.: *Handbook of the Economics of Finance* (North-Holland Elsevier, Amsterdam), forthcoming.

Benartzi, Shlomo, and Richard H. Thaler, 1995, Myopic loss aversion and the equity premium puzzle, *Quarterly Journal of Economics* 110, 73–92.

Bertaut, Carol C., and Michael Haliassos, 1997, Precautionary portfolio behavior from a life-cycle perspective, *Journal of Economic Dynamics and Control* 21, 1511–1542.

Bewley, Truman F., 1982, Thoughts on tests of the intertemporal asset pricing model, Working paper, Northwestern University.

Blume, Marshall E., and Stephen P. Zeldes, 1993, The structure of stock ownership in the U.S., Working paper, University of Pennsylvania.

Bodie, Zvi, Robert C. Merton, and William F. Samuelson, 1992, Labor supply flexibility and portfolio choice in a life cycle model, *Journal of Economic Dynamics and Control* 16, 427–449.

Boldrin, Michel, Lawrence J. Christiano, and Jonas D. M. Fisher, 2001, Habit persistence, asset returns, and the business cycle, *American Economic Review* 91, 149–166.

Brav, Alon, George M. Constantinides, and Christopher C. Geczy, 2002, Asset pricing with heterogeneous consumers and limited participation: Empirical evidence, *Journal of Political Economy* 110, forthcoming.

Brav, Alon, and Christopher C. Geczy, 1995, An empirical resurrection of the simple consumption CAPM with power utility, Working paper, University of Chicago.

Brav, Alon, and J. B. Heaton, 2002, Competing theories of financial anomalies, *Review of Financial Studies* 15, 475–506.

Breeden, Douglas T., 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 265–296.

Brown, Stephen J., William N. Goetzmann, and Stephen Ross, 1995, Survival, *Journal of Finance* 50, 853–873.

Campbell, John Y., 2001, Asset pricing at the millennium, *Journal of Finance* 55, 1515–1567.

Campbell, John Y., 2002, Consumption-based asset pricing, in George M. Constantinides, Milton Harris, and Rene Stulz, eds.: *Handbook of the Economics of Finance* (North-Holland Elsevier, Amsterdam), forthcoming.

Campbell, John Y., Joao F. Cocco, Francisco J. Gomes, and Pascal J. Maenhout, 2001, Investing retirement wealth: A life-cycle model, in John Y. Campbell and Martin Feldstein, eds.: *Risk Aspects of Social Security Reform* (University of Chicago Press, Chicago).

Campbell, John Y., and John H. Cochrane, 1999, By force of habit: A consumption-based explanation of aggregate stock market behavior, *Journal of Political Economy* 107, 205–251.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, NJ).

Campbell, John Y., and Robert J. Shiller, 1988a, The dividend–price ratio and expectations of future dividends and discount factors, *Review of Financial Studies* 1, 195–228.

Campbell, John Y., and Robert J. Shiller, 1988b, Stock prices, earnings, and expected dividends, *Journal of Finance*, 43, 661–676.

Campbell, John Y., and Robert J. Shiller, 1998, Valuation ratios and the long-run stock market outlook, *Journal of Portfolio Management* 24, 11–26.

Claus, James, and Jacob Thomas, 2001, Equity premia as low as three percent? Evidence from analysts' earnings forecasts for domestic and international stock markets, *Journal of Finance* 56, 1629–1666.

Cocco, Joao F., Francisco J. Gomes, and Pascal J. Maenhout, 1999, Portfolio choice over the life cycle, Working paper, Harvard University.

Cochrane, John H., 1991, A simple test of consumption insurance, *Journal of Political Economy* 99, 957–976.

Cochrane, John H., 1997, Where is the market going? Uncertain facts and novel theories, *Economic Perspectives* (Federal Reserve Bank of Chicago) 21, 3–37.

Cochrane, John H., 2001, *Asset Pricing* (Princeton University Press, Princeton, NJ).

Cochrane, John H., and Lars Peter Hansen, 1992, Asset pricing explorations for macroeconomics, in Olivier J. Blanchard and Stanley Fischer, eds.: *NBER Macroeconomics Annual* (MIT Press, Cambridge, MA).

Cogley, Timothy, 2002, Idiosyncratic risk and the equity premium: Evidence from the consumer expenditure survey, *Journal of Monetary Economics* 49, 309–334.

Constantinides, George M., 1982, Intertemporal asset pricing with heterogeneous consumers and without demand aggregation, *Journal of Business* 55, 253–267.

Constantinides, George M., 1990, Habit formation: A resolution of the equity premium puzzle, *Journal of Political Economy* 98, 519–543.

Constantinides, George M., John B. Donaldson, and Rajnish Mehra, 2001, Junior must pay: Pricing the implicit put in privatizing social security, Working paper, University of Chicago.

Constantinides, George M., John B. Donaldson, and Rajnish Mehra, 2002a, Junior can't borrow: A new perspective on the equity premium puzzle, *Quarterly Journal of Economics* 117, 269–296.

Constantinides, George M., John B. Donaldson, and Rajnish Mehra, 2002b, Junior is rich: Bequests as consumption, Working paper, University of Chicago.

Constantinides, George M., and Darrell Duffie, 1996, Asset pricing with heterogeneous consumers, *Journal of Political Economy* 104, 219–240.

Daniel, Kent, and David Marshall, 1997, The equity premium puzzle and the risk-free rate puzzle at long horizons, *Macroeconomic Dynamics* 1, 452–484.

Danthine, Jean-Pierre, John B. Donaldson, and Rajnish Mehra, 1992, The equity premium and the allocation of income risk, *Journal of Economic Dynamics and Control* 16, 509–532.

Davis, James L., Eugene F. Fama, and Kenneth R. French, 2000, Characteristics, covariances, and average returns, *Journal of Finance* 55, 389–406.

DeBondt, Werner F. M., and Richard Thaler, 1985, Does the stock market overreact? *Journal of Finance* 40, 793–805.

Dimson, Elroy, Paul Marsh, and Michael Staunton, 2002, *Triumph of the Optimists: 101 Years of Global Investment Returns* (Princeton University Press, Princeton, NJ).

Duesenberry, James S., 1949, *Income, Saving, and the Theory of Consumer Behavior* (Harvard University Press, Cambridge, MA).

Epstein, Larry G., and Stanley E. Zin, 1991, Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis, *Journal of Political Economy* 99, 263–286.

Fama, Eugene F., and Kenneth R. French, 1988, Dividend yields and expected stock returns, *Journal of Financial Economics* 22, 3–25.

Fama, Eugene F., and Kenneth R. French, 2002, The equity premium, *Journal of Finance* 57, 637–660.

Ferson, Wayne E., and George M. Constantinides, 1991, Habit persistence and durability in aggregate consumption, *Journal of Financial Economics* 29, 199–240.

Ferson, Wayne E., and Campbell R. Harvey, 1992, Seasonality and consumption-based asset pricing, *Journal of Finance* 47, 511–552.

Friend, Irwin, and Marshall E. Blume, 1975, The demand for risky assets, *The American Economic Review* 65, 900–922.

Gabaix, Xavier, and David Laibson, 2001, The 6D bias and the equity premium puzzle, in Ben Bernanke and Ken Rogoff, eds.: *NBER Macroeconomics Annual 2001* (MIT Press, Cambridge, MA).

Goyal, Amit, and Ivo Welch, 1999, Predicting the equity premium, Working paper, UCLA.

Grossman, Sanford J., and Robert J. Shiller, 1981, The determinants of the variability of stock market prices, *American Economic Review Papers and Proceedings* 71, 222–227.

Guiso, Luigi, Michael Haliassos, and Tullio Jappelli, 2001, *Household Portfolios* (MIT Press, Cambridge, MA).

Haliassos, Michael, and Carol C. Bertaut, 1995, Why do so few hold stocks? *The Economic Journal* 105, 1110–1129.

Hansen, Lars Peter, and Ravi Jagannathan, 1991, Implications of security market data for models of dynamic economies, *Journal of Political Economy* 99, 225–262.

Hansen, Lars Peter, and Kenneth J. Singleton, 1982, Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* 50, 1269–1288.

He, Hua, and David M. Modest, 1995, Market frictions and consumption-based asset pricing, *Journal of Political Economy* 103, 94–117.

Heaton, John C., 1995, An empirical investigation of asset pricing with temporally dependent preference specifications, *Econometrica* 63, 681–717.

Heaton, John C., and Deborah J. Lucas, 1996, Evaluating the effects of incomplete markets on risk sharing and asset pricing, *Journal of Political Economy* 104, 443–487.

Heaton, John C., and Deborah J. Lucas, 1999, Stock prices and fundamentals, in Ben Bernanke and Julio Rotemberg, eds.: *NBER Macroeconomics Annual 1999* (MIT Press, Cambridge, MA).

Heaton, John, C., and Deborah J. Lucas, 2000, Portfolio choice and asset prices: The importance of entrepreneurial risk, *Journal of Finance* 55, 1163–1198.

Hirshleifer, David, 2001, Investor psychology and asset pricing, *Journal of Finance* 56, 1533–1597.

Ibbotson Associates, 2001, *Stocks, Bonds, Bills, and Inflation Yearbook* (Ibbotson Associates, Chicago).

Ibbotson, Roger G., and Peng Chen, 2001, Stock market returns in the long run: Participating in the real economy, Working paper, Ibbotson Associates.

Jacobs, Kris, 1999, Incomplete markets and security prices: Do asset-pricing puzzles result from aggregation problems? *Journal of Finance* 54, 123–163.

Jacobs, Kris, and Kevin Q. Wang, 2001, Idiosyncratic consumption risk and the cross section of asset returns, Working paper, McGill University and University of Toronto.

Jagannathan, Ravi, and Narayana R. Kocherlakota, 1996, Why should older people invest less in stocks than younger people? *Federal Bank of Minneapolis Quarterly Review* 20, 11–23.

Jorion, Philippe, and William N. Goetzmann, 1999, Global stock markets in the twentieth century, *Journal of Finance* 54, 953–980.

Keynes, John M., 1936, *The General Theory of Employment, Interest, and Money* (Hartcourt, Brace and Company, New York).

Kocherlakota, Narayana R., 1996, The equity premium: It's still a puzzle, *Journal of Economic Literature* 34, 42–71.

Krebs, Tom, 2002, Testable implications of consumption-based asset pricing with incomplete markets, *Journal of Mathematical Economics*, forthcoming.

Li, Yuming, 2001, Expected returns and habit persistence, *Review of Financial Studies* 14, 861–899.

Lucas, Robert, Jr., 1978, Asset prices in an exchange economy, *Econometrica* 46, 1429–1446.

Lynch, Anthony W., 1996, Decision frequency and synchronization across agents: Implications for aggregate consumption and equity returns, *Journal of Finance* 51, 1479–1497.

Mace, Barbara J., 1991, Full insurance in the presence of aggregate uncertainty, *Journal of Political Economy* 99, 928–956.

Mankiw, N. Gregory, 1986, The equity premium and the concentration of aggregate shocks, *Journal of Financial Economics* 17, 211–219.

Mankiw, N. Gregory, and Matthew D. Shapiro, 1986, Risk and return: Consumption beta versus market beta, *Review of Economics and Statistics* 68, 452–459.

Mankiw, N. Gregory, and Stephen P. Zeldes, 1991, The consumption of stockholders and nonstockholders, *Journal of Financial Economics* 29, 97–112.

Marshall, Alfred, 1920, *Principles of Economics: An Introductory Volume* (Macmillan, London).

McGrattan, Ellen R., and Edward C. Prescott, 2001, Taxes, regulations, and asset prices, Working paper, Federal Reserve Bank of Minneapolis.
Mehra, Rajnish, 1998, On the volatility of stock prices: An exercise in quantitative theory, *International Journal of Systems Science* 29, 1203–1211.
Mehra, Rajnish, and Edward C. Prescott, 1985, The equity premium: A puzzle, *Journal of Monetary Economics* 15, 145–161.
Mehra, Rajnish, and Edward C. Prescott, 1988, The equity premium: A solution? *Journal of Monetary Economics* 22, 133–136.
Mehra, Rajnish, and Edward C. Prescott, 2002, The equity premium in retrospect, in George M. Constantinides, Milton Harris, and Rene Stulz, eds.: *Handbook of the Economics of Finance* (North-Holland Elsevier, Amsterdam), forthcoming.
Menzly, Lior, Tano Santos, and Pietro Veronesi, 2001, Habit formation and the cross section of stock returns, Working paper, University of Chicago.
Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.
Negishi, Takashi, 1960, Welfare economics and existence of an equilibrium for a competitive economy, *Metroeconomica* 12, 92–97.
Pollak, Robert A., 1970, Habit formation and dynamic demand functions, *Journal of Political Economy* 78, 745–763.
Poterba, James M., 2001, The rise of the "equity culture": U.S. stockownership patterns, 1989–1998, Working paper, MIT.
Rietz, Thomas A., 1988, The equity risk premium: A solution, *Journal of Monetary Economics* 22, 117–131.
Rubinstein, Mark, 1976, The valuation of uncertain income streams and the pricing of options, *Bell Journal of Economics* 7, 407–425.
Ryder, Harl E., and Geoffrey M. Heal, 1973, Optimum growth with intertemporally dependent preferences, *Review of Economic Studies* 40, 1–43.
Shiller, Robert, 1984, Stock prices and social dynamics, *Brookings Papers on Economic Activity* 2, 457–498.
Shiller, Robert, 2000, *Irrational Exuberance* (Princeton University Press, Princeton, NJ).
Siegel, Jeremy J., 1998, *Stocks for the Long Run* (McGraw-Hill, New York).
Siegel, Jeremy J., 1999, The shrinking equity premium, *Journal of Portfolio Management* 26, 10–17.
Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, 2000, Consumption and risk sharing over the life cycle, Working paper, Carnegie Mellon University.
Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, 2001, Asset pricing with idiosyncratic risk and overlapping generations, Working paper, Carnegie Mellon University.
Sundaresan, Suresh M., 1989, Intertemporally dependent preferences and the volatility of consumption and wealth, *Review of Financial Studies* 2, 73–89.
Vissing-Jorgensen, Annette, 2002, Limited asset market participation and the elasticity of intertemporal substitution, *Journal of Political Economy* 110, forthcoming.
Vuolteenaho, Tuomo, 2000, Understanding the aggregate book-to-market ratio and its implications to current equity-premium expectations, Working paper, Harvard University.
Wachter, Jessica, 2001, Habit formation and returns on bonds and stocks, Working paper, New York University.
Weil, Philippe, 1989, The equity premium puzzle and the risk-free rate puzzle, *Journal of Monetary Economics* 24, 401–421.
Welch, Ivo, 2001, The equity premium consensus forecast revisited, Working paper, Yale University.

FAJ | Financial Analysts Journal
Volume 66 • Number 1
©2010 CFA Institute

Economic Growth and Equity Investing

Bradford Cornell

The performance of equity investments is inextricably linked to economic growth. Nonetheless, few studies on investing have explicitly taken research on economic growth into account. This study bridges that gap by examining the implications for equity investing of both theoretical models and empirical results from growth theory. The study concludes that over the long run, investors should anticipate real returns on common stock to average no more than about 4 percent.

The performance of equity investments is inextricably linked to economic growth. Earnings, the source of value for equity investments, are themselves driven by economic activity. Unless corporate profits rise as a percentage of GDP, which cannot continue indefinitely, earnings growth is constrained by GDP growth. This dynamic means that the same factors that determine the rate of economic growth also place bounds on earnings growth and, thereby, the performance of equity investments. Despite these well-known facts, few studies on equity investing have explicitly taken the literature on economic growth into account. This observation is not meant to imply that research connecting economic growth with equity returns is sparse. Numerous contributions in that area include several provocative pieces by Arnott and Bernstein (2002), Arnott and Asness (2003), and Bernstein and Arnott (2003). Nonetheless, rarely has this research been expressly tied to the literature on the theory of economic growth. By bridging that gap, further insight can be gained into the relationship between economic growth and equity returns and forecasts regarding future returns can be placed on a more solid foundation.

Economic Growth: Theory and Data

The focus of economic growth theory is explaining expansion in the standard of living as measured by real per capita GDP. In the neoclassical model of economic growth, originally developed by Solow (1956), per capita GDP growth over the long run is entirely attributable to exogenous technological innovation.[1] This conclusion may surprise those not steeped in growth theory, given the intuitive thinking that output per capita can always be increased by simply adding more capital. Although adding capital does increase output per capita, it does so at a declining rate. Consequently, rational producers stop adding capital when the marginal product of capital drops to its marginal cost. When the economy reaches that point, it is said to be in a steady state. Once the economy reaches the steady state growth path, the ratio of capital to labor (C/L) remains constant and per capita GDP growth ceases unless the production function changes so as to increase the marginal product of capital.

The source of change in the production function is technological innovation. By increasing the marginal product of capital, technological progress breaks the deadlock imposed by diminishing returns and makes further growth in per capita output profitable. So long as the technological innovation continues, so too does the growth in per capita GDP.

This conclusion is not limited to such early models as Solow's, in which the rate of technological change is exogenous. Following Romer (1990), a variety of growth models have been developed in which the amount of investment in R&D—and thus the rate of technological progress—is endogenous. Even in these more sophisticated models, however, the declining marginal product of capital ensures that long-run per capita growth is bounded by the rate of technological progress. The word "bounded" is important because the ability of a society to exploit modern technology effectively is not a foregone conclusion. For example, from 1960 to 2005, all the countries of sub-Saharan Africa, with the exception of South Africa, experienced little or no growth. This failure of certain poor countries to grow is one of the fundamental mysteries of economics, but it is not a relevant consideration here.[2] Virtually the entire global stock

Bradford Cornell is professor of financial economics at California Institute of Technology, Pasadena.

market capitalization is concentrated in a relatively few highly developed countries. For those countries, the impediments to effective adoption of technology have proved to be minor, at least to date.

Before turning to the data on economic growth, I need to address one remaining issue. The conclusion that growth is attributable exclusively to technological innovation is based on the assumption that the economy has reached the steady state. If the capital stock is below the steady state—and thus the marginal product of capital exceeds its marginal cost—room still exists for the deepening of capital. In that situation, a country's growth rate can exceed the steady state growth rate because it is spurred by capital deepening, as well as by technological innovation. As C/L rises toward its steady state value, the growth rate converges to the steady state level that is attributable to technological change.

The capital stock of a country may be below its steady state level for a variety of reasons. An obvious example is warfare. Another is the opening of a previously closed society. Whatever the reason, growth theory predicts that a country with a C/L below the steady state level will grow more rapidly during a period of capital deepening. Growth theorists refer to this "catch-up" as convergence.

Convergence is important to bear in mind when analyzing historical growth rates with the goal of forecasting future growth. If the historical sample includes growth rates of countries that are in the process of converging to a steady state, the historical growth rates will exceed the future rates that will apply once the steady state has been achieved.

Convergence also helps explain why long-run growth rates for a particular country are remarkably constant. To illustrate, **Figure 1** plots the log of real per capita GDP in the United States from 1802 through 2008. The long-run average growth rate of 1.8 percent is also shown. Over this period, even the largest downturns (associated with the U.S. Civil War and the Great Depression) appear only as temporary dips in a remarkably smooth progression. That smooth progression is attributable in part to the fact that accelerations in economic growth, associated with capital accumulation, followed the dips, which were tied to a drop in the capital stock below its steady state level.

With that background, **Table 1** presents Barro and Ursúa's (2008) update of Maddison's (2003) compilation of information on world economic growth from 1923 to 2006. The starting point in Table 1 is 1923, the first year for which Barro and Ursúa had data for all the countries in their sample. Extending the sample backward for those countries with longer time series available does not affect the essential nature of the findings. Table 1 also reports growth rates for a shorter sample period (beginning in 1960) to take into account the possibility of nonstationarity in the data.

Figure 1. Logarithm of Real per Capita GDP, 1802–2008

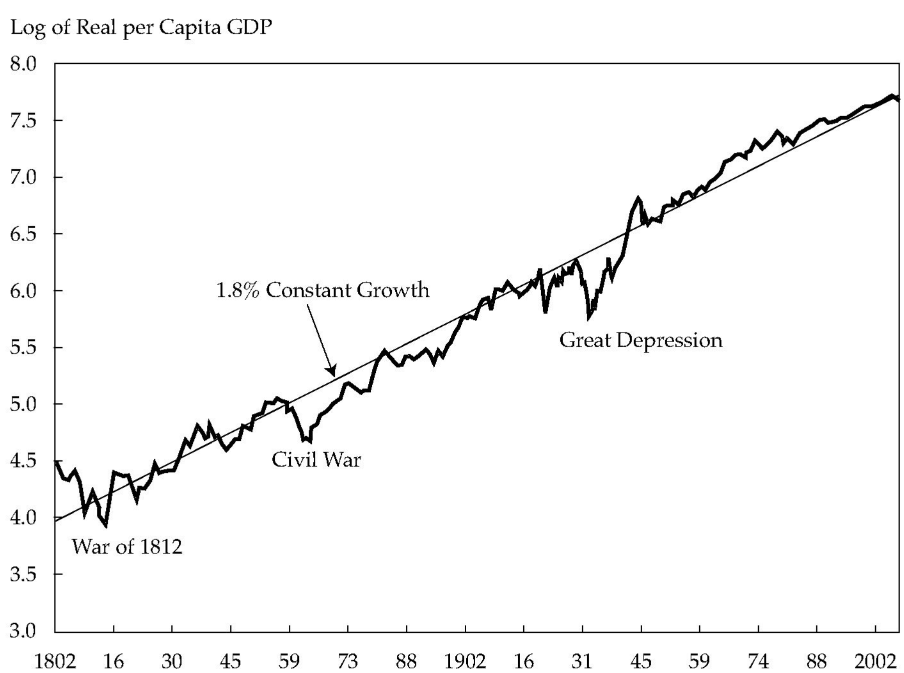


Table 1. Real Growth Rates in per Capita GDP, 1923–2006

Country	1923–2006	1960–2006
A. Mature Economies		
Australia	1.85%	2.16%
Austria	2.53	2.76
Belgium	2.11	2.62
Canada	2.22	2.27
Denmark	1.97	2.11
France	2.28	2.51
Germany	2.41	2.23
Italy	2.57	2.98
Japan	3.11	3.86
Netherlands	2.01	2.35
Spain	2.30	3.42
Sweden	2.50	2.25
Switzerland	1.63	1.51
United Kingdom	1.95	2.15
United States	1.42	1.14
Average	2.19%	2.42%
B. Developing and More Recently Developed Economies		
Argentina	1.10%	1.16%
Brazil	2.68	2.34
Chile	1.95	2.47
Colombia	2.18	2.24
Egypt	1.45	3.09
Finland	2.91	2.92
Greece	2.77	3.23
Iceland	3.24	2.87
India	1.74	2.88
Indonesia	1.81	3.08
S. Korea	3.55	5.72
Malaysia	1.91	2.14
Mexico	2.70	4.16
New Zealand	1.51	1.36
Norway	2.86	3.01
Peru	1.44	0.97
Philippines	1.32	1.46
Portugal	2.75	3.43
S. Africa	1.53	1.01
Singapore	3.33	5.72
Sri Lanka	1.93	3.06
Taiwan	3.78	6.24
Turkey	2.75	2.40
Uruguay	2.19	2.24
Venezuela	2.54	0.45
Average	2.32%	2.79%

Source: Barro and Ursúa (2008).

The results are reported in terms of compound growth rates. The following example illustrates why using compound growth rates is preferable to using averages of annual growth rates. Suppose that the ratio of corporate profits to GDP is stationary but not constant. In particular, assume (as the data will later show) that corporate profits are more variable than GDP. In that case, even though the compound growth rates of the two variables must converge in the long run, the arithmetic mean of annual growth rates for corporate profits will exceed that for GDP because of the variance effect.[3] The higher mean growth rate in earnings is illusory, however, because it fails to take into account the mean reversion in earnings growth that must occur for the ratio to be stationary.

The results reported in Table 1 are divided into two groups. The first group comprises mature economies that were already developed before World War II. These countries, which account for virtually the entire global stock market capitalization, are the focus of this study. The second group consists of economies that were developed more recently or are still considered developing. Results for the second group are presented for completeness and to provide perspective on the impact of convergence.

Consistent with the hypothesis that a common rate of technological advance is driving growth in all the developed countries, the results for the first group are remarkably homogeneous. Virtually all the growth rates for the full sample are close to the average of 2.19 percent. The exceptions are the United States, on the low end, and Japan, on the high end. The former's rate of 1.42 percent reflects the fact that the United States was the closest to steady state growth in 1923, after emerging from World War I relatively unscathed. The higher growth rate for Japan reflects convergence. At the start of the sample period, Japan was a relatively undeveloped country whose capital stock was below the steady state level. Convergence is also evident in the shorter sample period, beginning in 1960. The European countries and Japan, whose capital stocks were damaged in World War II, grew more rapidly than the United States, Switzerland, and Australia, all of which avoided war-related domestic destruction.

The results for the second group are more heterogeneous, reflecting the fact that growth in some countries (e.g., Peru and Venezuela) has stalled for reasons not fully understood whereas others (e.g., South Korea and Taiwan) have experienced rapid convergence. Despite the heterogeneity, however, the average growth rates of 2.32 percent for the sample period beginning in 1923 and 2.79 percent for the sample period beginning in 1960 are close to the averages for the first group of countries.

The averages reported in Table 1 are simple averages. If the growth rates for the first group of countries are weighted by market capitalization, the average falls to about 2 percent in both periods because of the predominant role of the United States. Giving the United States a higher weight is reasonable not only because of its large market capitalization but also because its economy is closest to steady state growth. Given the long period of time since World War II, to assume that all the countries in the first group will eventually converge to steady state growth is reasonable. Therefore, they are more likely to grow at rates comparable to the U.S. historical rate than at their own historical rates. This likelihood suggests that 2 percent real per capita growth, which exceeds the recent U.S. growth rate by 0.5 percent, is the most that investors can reasonably expect in the long run. Furthermore, although growth could be stalled by a catastrophe, such as another world war, the speed of technological innovation has proved almost impossible to accelerate meaningfully. In the remainder of this article, therefore, I will use 2 percent as the estimate of future per capita GDP growth. This number should be thought of as an achievable, but not necessarily expected, outcome.

In addition to the possibility of a catastrophe are two other reasons why 2 percent may prove to be an optimistic growth forecast. First, national income accounting does not deduct costs associated with pollution and environmental degradation in the calculation of GDP. Although these costs have been a tiny fraction of GDP in the past, concern that they are growing rapidly is widespread. If that concern is justified, properly accounting for these costs will reduce the future growth rate of per capita GDP. Second, whether the historical rate of technological innovation is sustainable is far from clear. Weil (2009, p. 260) noted that the rate of growth of real per capita GDP attributable to technological progress remained largely constant from 1950 to 2005, but over the same period, the number of researchers in the G–20 countries grew from 251,000 to 2.6 million. This finding suggests a declining marginal product of research as making and applying new discoveries become more difficult. If this trend continues, it could lead to falling rates of growth in per capita GDP.

Population Growth

Business opportunities depend on total economic activity, not per capita output. To see why, consider a hypothetical example of an economy for which technological innovation—and thus productivity growth—is zero but which is experiencing 5 percent population growth. Companies that provide goods and services in this economy will, on average, experience 5 percent growth in real revenues. Assuming that their margins remain constant, this rate translates into 5 percent growth in real earnings. Of course, in a dynamic economy, existing companies could lose business to start-ups, which could result in dilution for existing investors (which is a separate issue addressed later in the article). For companies in the aggregate, real earnings should be tied to real GDP, as data presented later in the article reveal to be the case.

Converting per capita growth to aggregate growth requires an estimate of population growth. Fortunately, population growth rates change even more slowly and are more predictable than growth rates of real per capita GDP.

Data on population growth for the sample countries are reported in **Table 2**. The first column presents historical growth rates from 2000 to 2007 taken from the U.S. Central Intelligence Agency's *2008 World Fact Book*. The second column presents United Nations (2007) forecasts of population growth rates from 2005 to 2010. That the two columns are very similar reflects the slowly changing nature of population growth.

The data in Table 2 are consistent with the widely documented fact that population growth is negatively correlated with per capita GDP.[4] The average population growth rate for the first group of countries is less than half that for the second group. Even for the second group, however, both the average historical growth rate and the average projected growth rate are less than 1 percent. Presumably, as per capita GDP continues to rise, these growth rates will continue to decline.

On the basis of the data presented in Table 2, population growth can be expected to add no more than 1 percent to the growth rate in per capita GDP. In fact, an assumption of a zero long-run future growth rate for the developed countries would not be unreasonable. Given real per capita growth of 2 percent, this assumption implies that investors cannot reasonably expect long-run future growth in real GDP to exceed 3 percent.

Earnings and GDP

The fundamental source of value for equity investors is earnings, not GDP. That long-run real GDP growth is reasonably bounded at 3 percent does not necessarily mean that the same is true of earnings, which depends on whether the ratio of earnings to GDP (E/GDP) is stationary. To test that hypothesis requires data on aggregate earnings.

Two primary measures of aggregate earnings are used in the United States. The first measure is derived from the national income and product accounts (NIPAs), produced by the U.S. Department

Table 2. Historical and Projected Population Growth Rates, 2000–2010

Country	Historical 2000–2007	Projected 2005–2010
A. Mature Economies		
Australia	1.22%	1.01%
Austria	0.06	0.36
Belgium	0.11	0.24
Canada	0.83	0.90
Denmark	0.30	0.90
France	0.57	0.49
Germany	–0.04	–0.07
Italy	0.00	0.13
Japan	–0.14	–0.02
Netherlands	0.44	0.21
Spain	0.10	0.77
Sweden	0.16	0.45
Switzerland	0.33	0.38
United Kingdom	0.28	0.42
United States	0.88	0.97
Average	0.34%	0.48%
B. Developing and More Recently Developed Economies		
Argentina	1.07%	1.00%
Brazil	1.23	1.26
Chile	0.91	1.00
Colombia	1.41	1.27
Egypt	1.68	1.76
Finland	0.11	0.29
Greece	0.15	0.21
Iceland	0.78	0.84
India	1.58	1.46
Indonesia	0.18	1.16
S. Korea	0.27	0.33
Malaysia	1.74	1.69
Mexico	1.14	1.12
New Zealand	0.97	0.90
Norway	0.35	0.62
Peru	1.26	1.15
Philippines	1.99	1.72
Portugal	0.31	0.37
S. Africa	0.83	0.55
Singapore	1.14	1.19
Sri Lanka	0.94	0.47
Taiwan	0.24	0.36
Turkey	1.01	1.26
Uruguay	0.49	0.29
Venezuela	1.50	1.67
Average	0.94%	0.96%

Sources: Central Intelligence Agency (2008) and the United Nations (2007).

of Commerce's Bureau of Economic Analysis. The NIPAs contain an estimate of aggregate corporate profits that is based on data collected from corporate income tax returns. The second measure of aggregate earnings is derived by Standard & Poor's from data collected from corporate financial reports. Because the two measures are not identical, distinguishing what is included in each measure before using the data is important.

The NIPA profit measure is designed to provide a time series of the income earned from the current production of all U.S. corporations. The sample is not limited to publicly traded companies. The tax rules on which the NIPAs are based are designed to expedite the timely and uniform completion of corporate tax returns. For that reason, all corporations use a highly uniform set of rules for tax accounting.

Because the NIPAs are designed to measure economic activity connected with current production, the NIPA definition of corporate profits includes only receipts arising from current production less associated expenses. The NIPA definition, therefore, excludes transactions that reflect the acquisition or sale of assets or liabilities. Dividend receipts from domestic corporations are excluded to avoid a double counting of profits. For the same reason, bad-debt expenses and capital losses are also excluded.

The Standard & Poor's estimate of aggregate earnings is derived from reported financial statements. Rather than being based on a unified set of tax rules, financial accounting is based on GAAP, which is designed to allow management to tailor financial statements so as to reveal information that is useful to a particular company. Furthermore, financial accounting provides for depreciation and amortization schedules that allow companies to attempt to match expenses with the associated stream of income.

The aggregate earnings data available from Standard & Poor's are for the companies in the S&P 500 Index. Each year's data consist of the aggregate GAAP after-tax earnings for the 500 companies in the S&P 500 for that year. Thus, the sample of companies in the aggregate is constantly changing as the index is updated. Because the S&P 500 earnings reflect a shifting sample of corporations, the series of reported earnings can be discontinuous over time. Fortunately, given the size of the index, these discontinuities are small and have little impact on estimated earnings growth.

The differences between financial and tax accounting create two dissimilarities between the measures of earnings for the same company.[5] First, intertemporal differences arise because of the timing of revenue, and expense recognition often differs between the two systems. The best example is depreciation because tax rules generally allow for more rapid depreciation than companies choose to report under GAAP. Second, permanent differences exist because the revenues and expenses recognized under the two systems are not the same. Although important in the short run, these differences tend to cancel out over long horizons, and thus, the long-run growth rates in the two measures are similar. For example, the average growth rate in NIPA real corporate profits from 1947 to 2008 was 3.23 percent, as compared with a growth rate of 3.17 percent in S&P 500 real aggregate earnings.

As an aid in examining the behavior of E/GDP, **Figure 2** plots after-tax corporate profits from the NIPAs as a fraction of GDP for 1947–2008. The figure reveals no overall trend. The fraction is approximately the same at the end as at the beginning, and thus, the growth rate of corporate profits is almost identical to that of GDP. The same is largely true of S&P 500 aggregate earnings as a fraction of GDP, which is plotted in **Figure 3** (normalized to start at 8.23 percent to facilitate comparison with Figure 2). The fraction for the S&P 500 earnings is smaller because the S&P 500 measure is less comprehensive than the NIPA measure. Unlike the NIPA data, the S&P 500 ratio exhibits a slight downward trend, reflecting the fact that as the economy has grown, the S&P 500 companies have become a progressively smaller fraction of total earnings. Therefore, the data are generally consistent with the hypothesis that over the long run, aggregate earnings are a stationary fraction of GDP. Certainly, no evidence exists of a persistent increase in the ratio, no matter which measure of earnings is chosen. This observation implies that the long-run growth rates of GDP place a limit on the long-run growth rates of earnings.

Although the data largely support the hypothesis that E/GDP is stationary, it is far from constant. Figure 2 shows that corporate profits vary between 3 percent and 11 percent of GDP. The variability of the ratio for S&P 500 earnings is even greater. This variability suggests that when earnings are low relative to GDP, they grow more quickly; the reverse is true when earnings are relatively high. This mean reversion in the growth rate of earnings maintains the stationarity of E/GDP.

Note that in an efficient market, the mean reversion in earnings growth would have no impact on stock returns because it would be impounded into current prices. Campbell and Shiller (1998), however, provided evidence that long-run average earnings are, in fact, predictive of future stock returns. Specifically, when the ratio of price to average earnings over the previous 10 years is high, future stock returns tend to be low; the reverse is true when the ratio is low. This finding suggests that the market does not fully account for the mean-reverting nature of long-run earnings growth.

Figure 2. Corporate Profits as a Percentage of GDP, 1947–2008

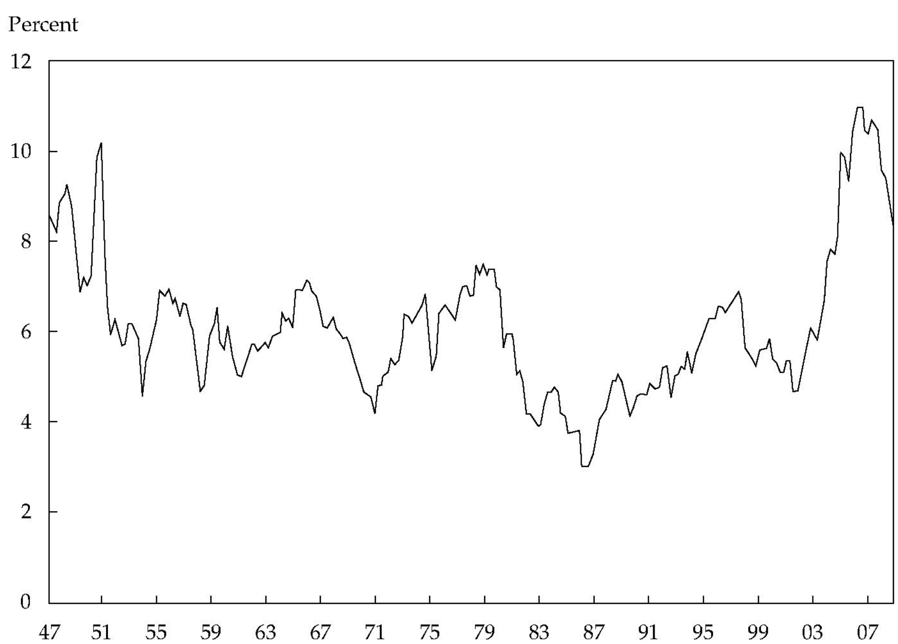
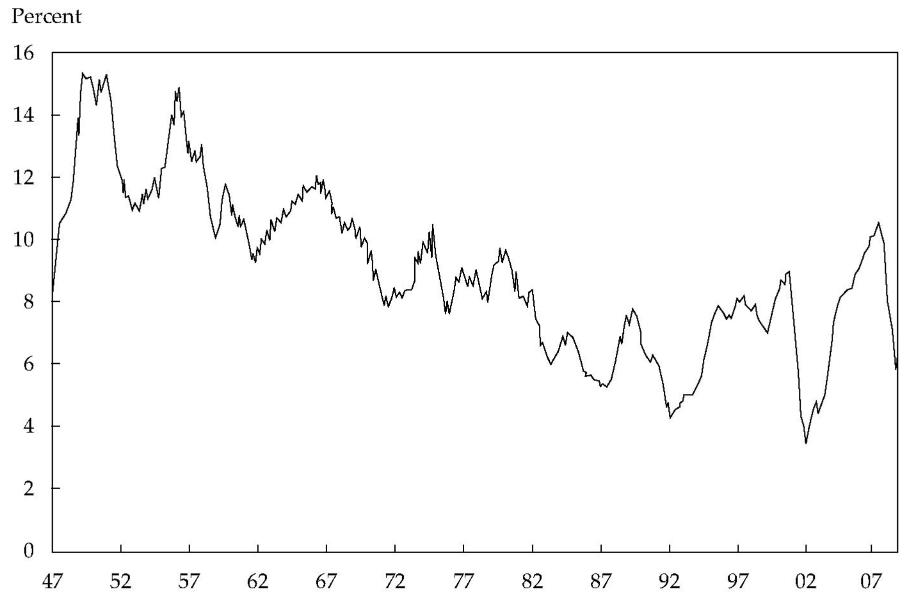


Figure 3. S&P 500 Earnings as a Percentage of GDP, 1947–2008



That the ratio of aggregate earnings to GDP is stationary implies that investors can expect aggregate real earnings growth to match, but not exceed, real GDP growth in the long run. Unfortunately, the same is not true of the earnings to which current investors have a claim. Two reasons explain this discrepancy. First, an investor's pro rata portion of a company's earnings will be affected by the company's share issuances and repurchases. If this dilution (or accretion) is ongoing, growth in aggregate earnings and earnings per share will diverge. Second and more important, current investors do not participate in the earnings of new businesses unless they dilute their current holdings to purchase shares in start-ups. Therefore, start-ups drive a wedge between the growth in aggregate earnings and the growth in the earnings to which current investors have a claim.

To illustrate the second effect, consider a simple example in which all companies in the economy are identical and earn $10 a share per period. Furthermore, assume that each company has a market value of $100 a share and has 1,000 shares outstanding. All earnings are paid out, so the values of the companies remain constant. Finally, assume that at the outset only two companies are in the economy, so aggregate earnings are $20,000. A current investor who holds 1 percent of each company has a pro rata share of aggregate earnings of $200. Now assume that the economy grows and a third company is started. As a result, aggregate earnings rise to $30,000, but the current investor does not participate in that growth and thus still holds 1 percent of the first two companies with rights to earnings of $200. If the current investor wanted to add the third company to the portfolio without investing new cash, the investor would have to dilute the portfolio's holdings in the first two companies. After the dilution, the investor would hold 0.67 percent of each of the three companies and would thus still have rights to earnings of $200. Therefore, the growth in earnings experienced by the current investor does not match the growth in aggregate earnings.

Bernstein and Arnott (2003) suggested an ingenious procedure for estimating the combined impact of both effects on the rate of growth of earnings to which current investors have a claim. They noted that total dilution on a marketwide basis can be measured by the ratio of the proportionate increase in market capitalization to the value-weighted proportionate increase in stock price. More precisely, net dilution for each period is given by the equation

$$\text{Net dilution} = \frac{1+c}{1+k} - 1, \qquad (1)$$

where c is the percentage capitalization increase and k is the percentage increase in the value-weighted price index. Note that this dilution measure holds exactly only for the aggregate market portfolio. For narrower indices, the measure can be artificially affected if securities are added to or deleted from the index.

To account for the impact of dilution, the Bernstein–Arnott measure was estimated by using monthly data for the entire universe of CRSP stocks from 1926 to 2008. Using CRSP data for this purpose presents one problem. The CRSP universe was expanded twice during the sample period: in

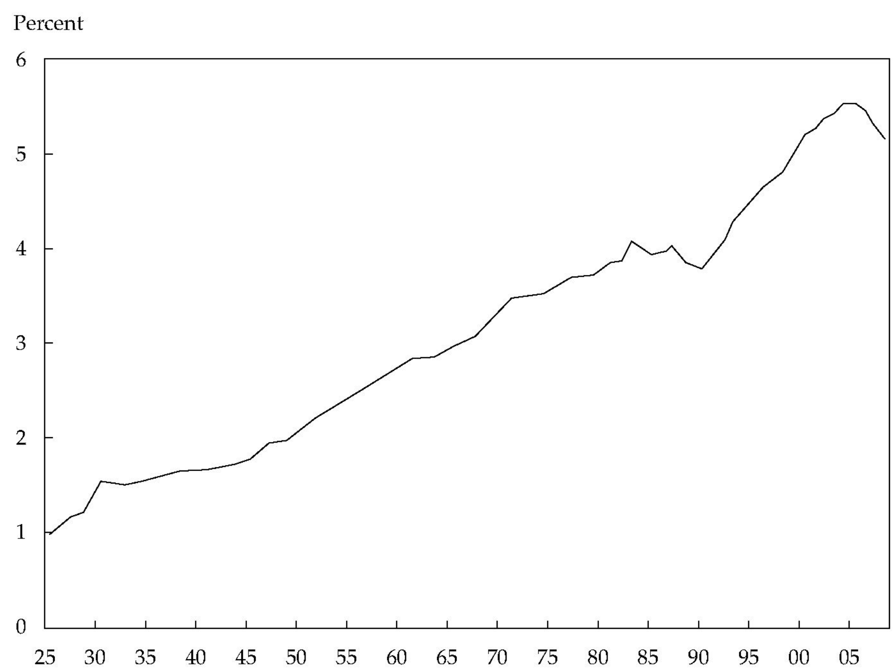
July 1962, when Amex stocks were added, and in July 1972, when NASDAQ stocks were added. Both these additions caused a significant increase in market capitalization unaccompanied by a corresponding increase in the value-weighted price. To eliminate the impact of these artificial discontinuities, I set the estimate of net dilution at zero for both July 1962 and July 1972.

Figure 4 plots the compounded estimate of net dilution from 1926 to 2008. It rises continuously except for downturns in the early 1990s and in 2006–2008. The average rate of dilution over the entire period is 2 percent. The primary source of dilution is the net creation of new shares as new companies capitalize their businesses with equity. The impact of start-ups is not surprising in light of the fact that more than half of U.S. economic growth comes from new enterprises, not from the growth of established businesses. Given the continuing importance of start-ups, the rate of dilution is highly unlikely to subside unless the rate of innovation slows. If the rate of innovation slows, however, GDP growth will also decline. Consequently, to conclude that the rate of growth of earnings, net of dilution, will remain largely constant is reasonable. Therefore, to estimate the growth rate of earnings to which current investors have a claim, approximately 2 percent must be deducted from the growth rate of aggregate earnings.

Putting the pieces together, we can see that growth theory predicts that current investors should count on long-run growth in real earnings of no more than 1 percent. This rate equals real growth of 3 percent in aggregate earnings, adjusted downward by 2 percent to account for dilution.

Arnott and Bernstein (2002) and Bernstein and Arnott (2003, p. 49) observed that "earnings and dividends grow at a pace very similar to that of per capita GDP." This observation correctly summarizes U.S. economic history, but it may not be true for other countries and it may not hold for the United States in the future. In terms of my analysis, the reason that earnings and dividends mirror per capita GDP is that population growth and dilution have both been about 2 percent between 1870 and 2008. Consequently, these two terms cancel each other out when we move from estimated growth in real per capita GDP to estimated growth in real earnings per share. But there is no theoretical reason why this cancellation should necessarily occur. For instance, population growth in Western Europe has fallen essentially to zero. If the United States were to follow suit but dilution were to continue at about 2 percent a year, growth in real earnings would be 2 percentage points *less* than growth in per capita GDP. In short, the Arnott–Bernstein observation is a shortcut that has historically held in the United States but is not a necessary condition. Therefore, a more complete analysis that takes into

Figure 4. The Impact of Dilution on Investor Earnings, 1926–2008



account both population growth and dilution is generally preferable. I do not present that analysis here because of limitations on dilution data for countries other than the United States.

Implications of Economic Growth Theory for Expected Stock Returns

The story thus far is that economic growth places a limit on the long-run growth of real earnings per share available to investors. On the basis of the data I have analyzed here, that limit is what many investors might consider a relatively anemic 1 percent. The next step is to explore the implications of that limitation for future returns on common stocks.

By definition, the rate of return on stock in period t is given by

$$R_t = \frac{D_t}{P_{t-1}} + GP_t, \tag{2}$$

where D_t is the dividend for year t, P_{t-1} is the price at the end of year $t - 1$, and $GP_t = (P_t - P_{t-1})/P_{t-1}$. Following Fama and French (2002), we can write Equation 2 in terms of long-run average values, denoted by A(), as

$$A(R_t) = A\left(\frac{D_t}{P_{t-1}}\right) + A(GP_t). \tag{3}$$

Equation 3 states that the long-run average return equals the average dividend yield plus the average capital gain.

Equation 3 holds *ex ante* as well as *ex post*. It implies that the long-run future average return equals the future average dividend yield plus the future average capital gain. Assuming that the earnings-to-price ratio is stationary, the long-run average earnings growth rate, $A(GE_t)$, can be substituted for the average capital gain rate, giving

$$A(R_t) = A\left(\frac{D_t}{P_{t-1}}\right) + A(GE_t). \tag{4}$$

My preceding analysis implies that $A(GE_t)$ in Equation 4 should be no more than about 1 percent in the future. In addition, as of December 2008, the current dividend yield was 3.1 percent and the previous 50-year average was 3.3 percent. Because the two are nearly equal, substituting either into Equation 4 as a proxy for the future average yield suggests that investors should not expect long-run real returns on common stocks to go much beyond 4 percent. Note that this calculation does not need to be adjusted for repurchases because the impact of repurchases is already accounted for in the dilution calculation. An adjustment is required only if future repurchases are expected to exceed their past average.

Equation 4 can also be used to approximate the equity risk premium. Because the real return on short-term government securities has averaged about 1 percent over the last 80 years, Equation 4 implies that the equity risk premium measured with respect to short-term government securities is approximately equal to the expected average dividend yield. Using either the current yield or the past average yield translates this number into a long-run average equity risk premium of just more than 3 percent. If the premium is measured with respect to longer-maturity government securities with greater expected real returns, the equity premium is commensurately less. This result is markedly less than the average historical risk premium measured over the 1926–2008 period that is commonly referenced. It is consistent, however, with a long-running body of empirical work that shows the *ex ante* risk premium to be significantly smaller than the historical average.[6]

Thus far, all the results have been stated in terms of compound growth rates. For many purposes, however, the object of interest is the annual expected return. For example, discounted-cash-flow valuations typically require annual estimates of the discount rate. To convert compound growth rates, which are geometric averages, into arithmetic averages requires taking the variance effect into account. This step can be well approximated by adding one-half of the annual variance of returns to the compound growth rate.

Because earnings are volatile, the variance effect adds about 1 percent to the compound growth rates. This result means that growth theory predicts that future annual real returns on common stocks should average no more than about 5 percent and that the annual equity risk premium for short-term government securities is about 4 percent.

Using annual data, we can tie the growth theory analysis to the long-run performance of company investments. If a company retains a fraction, b, of its earnings and invests those funds at a real rate of return, k, then basic finance theory teaches that the earnings per share will grow at the rate $(b)(k)$. Growth theory predicts that the annual long-run average growth in real earnings per share is about 2 percent, taking into account both dilution and the variance effect. From 1960 to 2008, companies in the S&P 500 retained, on average, 54 percent of their earnings. Solving for k, this retention ratio implies a real return on corporate investments of about 4 percent.

One possible adjustment might be made to the foregoing results. Recall that the dilution calculation was based on the assumption of a stable repurchase rate throughout the sample period. In fact, repurchases accelerated following the passage, in 1982, of U.S. SEC Rule 10b-18, which greatly reduced the legal risk associated with repurchases. More specifically, a pronounced trend toward repurchases as the preferred form of marginal payout to shareholders took place. Brav, Graham, Harvey, and Michaely (2005) reported that following the SEC ruling, managers began behaving as if a significant capital market penalty were associated with cutting dividends but not with reducing repurchases. Accordingly, dividends are set conservatively and repurchases are used to absorb variations in total payout. To the extent that this reliance on repurchases is expected to continue, the estimated 2 percent dilution effect might be too large and growth rates would have to be adjusted upward. Most of the 2 percent dilution, however, is associated not with the actions of existing companies but with start-ups that finance their businesses with new equity. Therefore, the adjustment in the overall rate of future dilution should not be large.

International Considerations

Thus far, I have limited my analysis to the United States. This restriction is an obvious shortcoming because most major corporations are becoming increasingly global. Although a detailed examination of international data is beyond the scope of this article, several general conclusions can be drawn. First, the data presented in Table 2 suggest that real per capita GDP growth rates for the other developed countries should be comparable to the U.S. growth rate in the future. Second, for the other developed countries, population growth rates are forecasted to be lower. As a result, the implied limitations on earnings growth remain largely unchanged and are perhaps even lower when other developed countries are included in the sample. Third, with respect to the developing countries—particularly India and China, which are the most important by virtue of their size—convergence predicts that they will experience higher growth rates in real per capita GDP than the United States. In addition, most developing countries are forecasted to have comparable or higher population growth rates than the United States. These forecasts suggest that companies doing business in the developing world will experience higher rates of earnings growth than they achieve in the developed world. Nonetheless, as those countries develop, both real GDP and population growth rates should decline. Furthermore, the fraction of total earnings attributable to business in the developing world is relatively small for most companies. Therefore, if a complete analysis were done on a global basis, the earnings bounds derived from U.S. data and the related predictions regarding stock returns would be unlikely to be markedly affected.

Conclusion

The long-run performance of equity investments is fundamentally linked to growth in earnings. Earnings growth, in turn, depends on growth in real GDP. This article demonstrates that both theoretical research and empirical research in development economics suggest relatively strict limits on future growth. In particular, real GDP growth in excess of 3 percent in the long run is highly unlikely in the developed world. In light of ongoing dilution in earnings per share, this finding implies that investors should anticipate real returns on U.S. common stocks to average no more than about 4–5 percent in real terms. Although more work needs to be done before equally definitive predictions can be made with respect to international equities, the basic outlook appears to be quite similar.

I thank Rob Arnott, Eugene Fama, Kenneth French, John Haut, John Hirshleifer, Jason Hsu, and Brian Palmer for helpful comments on earlier versions of this article. Data were graciously provided by Robert Barro and by Research Associates, LLC.

This article qualifies for 1 CE credit.

Notes

1. For details on the Solow model and more recent elaborations, see Barro and Sala-i-Martin (2004).
2. Hall and Jones (1999) described the problem in detail and offered an intriguing solution.
3. As a first-order approximation, the annual arithmetic mean equals the compound growth rate plus one-half the standard deviation of the annual growth rates.
4. See, for example, Weil (2009, ch. 4).
5. For further details on the relationship between reported earnings and NIPA profits, see Mead, Moulton, and Petrick (2004).
6. Contributions in this area include those of Rozeff (1984); Ross, Brown, and Goetzmann (1995); Claus and Thomas (2001); Fama and French (2002); and Cornell and Moroz (forthcoming).

References

Arnott, Robert D., and Clifford S. Asness. 2003. "Surprise! Higher Dividends = Higher Earnings Growth." *Financial Analysts Journal*, vol. 59, no. 1 (January/February):70–87.

Arnott, Robert D., and Peter L. Bernstein. 2002. "What Risk Premium Is 'Normal'?" *Financial Analysts Journal*, vol. 58, no. 2 (March/April):64–85.

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

Barro, Robert J., and José F. Ursúa. 2008. "Macroeconomic Crises since 1870." NBER Working Paper 13940 (April).

Bernstein, William J., and Robert D. Arnott. 2003. "Earnings Growth: The Two Percent Dilution." *Financial Analysts Journal*, vol. 59, no. 5 (September/October):47–55.

Brav, Alon, John R. Graham, Campbell R. Harvey, and Roni Michaely. 2005. "Payout Policy in the 21st Century." *Journal of Financial Economics*, vol. 77, no. 3 (September):483–527.

Campbell, John Y., and Robert J. Shiller. 1998. "Valuation Ratios and the Long-Run Stock Market Outlook." *Journal of Portfolio Management*, vol. 24, no. 2 (Winter):11–26.

Central Intelligence Agency. 2008. *The 2008 World Fact Book*. Washington, DC: Central Intelligence Agency.

Claus, James, and Jacob K. Thomas. 2001. "Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets." *Journal of Finance*, vol. 56, no. 5 (October):1629–1666.

Cornell, Bradford, and Max Moroz. Forthcoming. "The Equity Premium Revisited." *Journal of Portfolio Management*.

Fama, Eugene F., and Kenneth R. French. 2002. "The Equity Premium." *Journal of Finance*, vol. 57, no. 2 (April):637–659.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output per Worker Than Others?" *Quarterly Journal of Economics*, vol. 114, no. 1 (February):83–116.

Maddison, Angus. 2003. *The World Economy: Historical Statistics*. Paris: OECD.

Mead, Charles Ian, Brent R. Moulton, and Kenneth Petrick. 2004. "NIPA Corporate Profits and Reported Earnings: A Comparison and Measurement Issues." Washington, DC: Bureau of Economic Analysis.

Romer, Paul M. 1990. "Endogenous Technical Change." *Journal of Political Economy*, vol. 98, no. S5 (October):S71–S102.

Ross, Stephen A., Stephen J. Brown, and William N. Goetzmann. 1995. "Survival." *Journal of Finance*, vol. 50, no. 3 (July):853–873.

Rozeff, Michael S. 1984. "Dividend Yields Are Equity Risk Premiums." *Journal of Portfolio Management*, vol. 11, no. 1 (Fall):68–75.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics*, vol. 70, no. 1 (February):65–94.

United Nations, Department of Economic and Social Affairs. 2007. "World Population Prospects: The 2006 Revision, Highlights." United Nations Working Paper No. ESA/P/WP.202 (www.un.org/esa/population/publications/wpp2006/WPP2006_Highlights_rev.pdf).

Weil, David N. 2009. *Economic Growth*. 2nd ed. Boston: Addison Wesley.

Historical Results II

Bradford Cornell

University of California
Los Angeles

The basic investment and constant-growth models, used with some justifiable simplifying assumptions about the U.S. market, indicate that the earnings growth rate cannot be greater than the GNP growth rate because of political forces and that the expected return, or cost of capital, in the long run should unconditionally be about 1.5 times the dividend-to-price ratio plus GNP growth. Adding reasonable assumptions about inflation produces a finding that equity risk premiums cannot be more than 3 percent (300 bps) because earnings growth is constrained by the real growth rate of the economy, which has been in the 1.5–3.0 percent range. In a consideration of today's market valuation, three reasons for the high market valuations seem possible: (1) stocks are simply seen as less risky, (2) valuation of equities is fundamentally determined by taxation, or (3) equity prices today are simply a mistake. A research question that remains and is of primary interest is the relationship between aggregate stock market earnings and GNP.

The very basic investment and constant-growth models from introductory finance courses can be used to interpret the long-run unconditional historical data on returns. So, let's begin with the basic model:

$$\frac{E_{t+1}}{E_t} = 1 + [(b)(ROE)],$$

where

E = earnings
b = the retention rate
ROE = return on equity

So that, with investment at time t denoted by I_t,

$$ROE = \frac{E_{t+1} - E_t}{I_t}$$

and

$$b = \frac{I_t}{E_t};$$

therefore, the growth rate of earnings is

$$(b)(ROE) = \frac{E_{t+1} - E_t}{E_t}.$$

This model implies that the growth rate in earnings is the retention rate times the return on equity, $(b)(ROE)$. In discussing the models, I would like to stress an important point: If you are interpreting the growth in earnings as being the retention rate times the return on equity, you have to be very careful when you are working with historical data. For example, does the retention rate apply only to dividends or to dividends and other payouts, such as share repurchases? The distinction is important because those proportions change in the more recent period. And if you make that distinction, you have to make a distinction between aggregate dividends and per share dividends because the per share numbers and the aggregate numbers will diverge. In working with the historical data, I have attempted to correct for that aspect.

Figure 1. S&P 500 Earnings and Dividends to GNP, 1950–July 2001

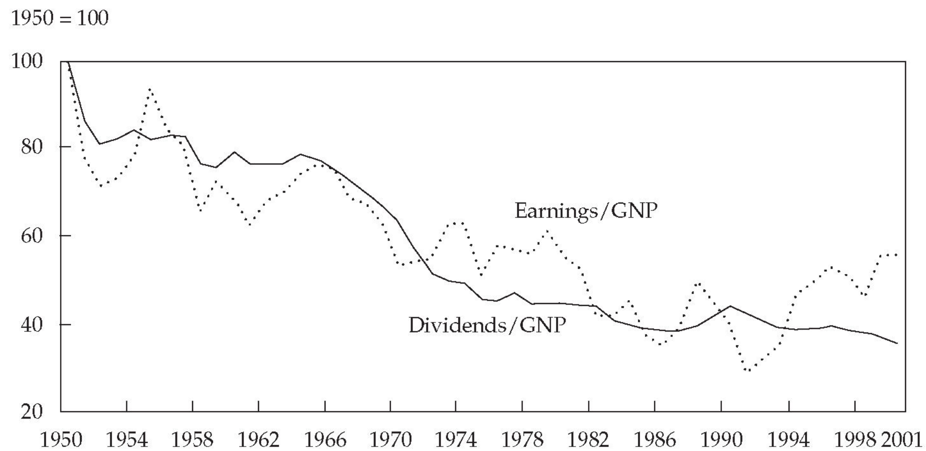


Table 1 gives the arithmetic average data for growth rates in GNP, earnings, and dividends for two periods: 1951–2000 and 1972–2000. (I used the 1972–2000 period because it mirrors the same period shown in Figure 1.) The earnings growth rates are so much more volatile than the dividend growth rates. And because of the volatility effect on arithmetic averages, GNP and earnings exhibit very similar growth rates from the early 1970s to the present. Dividends (and Table 1 shows the growth rate of actual dividends, not payouts) have grown much less than earnings for two reasons: First, dividends are less volatile, and second, dividend substitution is occurring. Corporations are not providing shareholders the same constant fraction of earnings (in the form of dividends) that they were in the past.

Despite the 1972–2000 data, it seems to me that earnings are not going to grow as fast as or faster than GNP in the future. This notion seems to be consistent with long-term historical data, and it fits my view of how politics works on the economy. If you accept that notion, it has immediate implications for the future.

Table 1. Historical Growth Rates of GNP, Earnings, and Dividends: Two Modern Periods

Period/Measure	GNP	Earnings	Dividends
1951–2000			
Mean	3.21%	2.85%	1.07%
Standard deviation	2.89	14.29	4.13
1972–2000			
Mean	2.62%	3.79%	0.96%
Standard deviation	2.94	15.72	3.58

Note: Growth rates for earnings and dividends are based on aggregate data.

First, under any reasonable underlying assumptions about inflation, equity risk premiums cannot be much more than 3 percent (300 bps) because the earnings growth rate is constrained unconditionally in the long run by the real growth rate of the economy, which has been in the range of 1.5–3.0 percent. Second, as **Table 2** shows, for an S&P level of about 1,000, you simply cannot have an equity risk premium any higher than 2 percent, 2.5 percent, or (at most) 3 percent.

Table 2. Value of the S&P 500 Index Given Various Real (Earnings or GNP) Growth Rates and Equity Risk Premiums

Real Growth Rate	Equity Risk Premium						
	2.0%	2.5%	3.0%	4.0%	5.0%	6.0%	7.0%
1.5%	845	724	634	507	423	362	317
2.0%	1,014	845	724	563	461	390	338
2.5%	1,268	1,014	845	634	507	423	362
3.0%	1,690	1,268	1,014	724	563	461	390

Assumptions: Inflation = 3 percent; long-term risk-free rate = 5.5 percent; payout = 1.5(S&P 500 dividend). The S&P 500 dividend used in the calculation was $16.90, so $P = 1.5(\$16.90)/(k - g)$, where k = 5.5 percent (the risk-free rate minus 3 percent inflation plus the risk premium) and g = real growth rate.

What simplifying assumptions can be made to work with the unconditional data? I have made some relatively innocuous simplifying assumptions. First, that b should adjust until the cost of capital equals the ROE at the margin. To be very conservative, therefore, I will assume that the ROE equals the cost of capital, or expected returns, in the aggregate. The problem that arises is: What if the retention rate times the cost of capital (that is, the minimal expected return on equity), bk, is greater than GNP growth? The second assumption deals with this possibility: I assume bk cannot be greater than GNP growth because political forces will come into play that will limit the ROE if earnings start to rise as a fraction of GNP.

The relationship between aggregate earnings and GNP is one of the research questions that I have been unable to find interesting papers on—perhaps because I have not searched well enough—but I want to bring up the subject to this group. It seems to me that if aggregate earnings start to rise, and Robert Shiller mentioned several reasons why it can happen [see the "Current Estimates and Prospects for Change" session], then tax rates can change, antitrust regulation can change (one of Microsoft's problems probably was that it was making a great deal of money, which is an indication that some type of regulation may be necessary), labor regulation can change, and so forth. And these variables can change *ex post* as well as *ex ante*. So, once a company starts making superior returns using a particular technology, the government may step in *ex post* and limit those returns. The critical research question is how earnings relate to GNP.

The constant-growth model is

$$P = \frac{D}{k-g}$$

or

$$k = \frac{D}{p} + g,$$

where

P = price
D = dividends
k = cost of capital
g = growth rate

What I am going to do is just an approximation because I am going to work with aggregate, not per share, data. I am going to assume that total payouts are 1.5 times dividends.[1] Payouts will probably be lower in the future, but if I work with aggregate payouts, then g should be the growth rate in aggregate potential payouts, which I will characterize as earnings.

One of the implications of the simplifying assumptions I have made, and it relates to the data that Jeremy Siegel just produced ["Historical Results I"], is that the expected returns on stocks should be equal to the earnings-to-price ratio. (In the more complicated equations, you have situations in which the ROE is not exactly equal to expected returns, but for my long-run data, the simplifying assumption that earnings yield equals the expected ROE is fine.) So, with these assumptions,

$$\begin{aligned} P &= \frac{D}{k-g} \\ &= \frac{D}{k-bk} \\ &= 1-(b)\left(\frac{E}{1-b}\right)(k) \\ &= \frac{E}{k} \end{aligned}$$

or

$$k = \frac{E}{P}.$$

A further implication is that if g is constrained to be close to the growth of GNP, then it is reasonable to substitute GNP growth for g in the constant-growth model. The implication of this conclusion is that the expected return, or cost of capital, in the long run should unconditionally be about 1.5 times the dividend-to-price ratio plus GNP growth:

$$k = 1.5\frac{D}{P} + \text{GNP growth}.$$

With this background, we can now look at some of the data.

Earnings and GNP

Figure 1 allows a comparison of dividends/GNP and (after-tax) earnings/GNP for 1950 through July 2001.[2] The data begin in 1950 because Fama believed that the data before then were unreliable. Figure 1 shows that, historically, earnings have declined as a fraction of GNP in this period. My assumption that earnings keep up with GNP works from about 1970 on, but I am looking at the picture in Figure 1 in order to make that conclusion. The ratio of earnings to GNP depends on a lot of things: the productivity of labor, capital, the labor-to-capital ratio, taxes, and (as I said earlier) a host of political forces. Figure 1 shows that earnings have, at best, kept up with GNP.

[1] This choice is based on recent findings by Jagannathan, Stephens, and Weisbach (2000) that we are seeing significant payouts today.

[2] These data were provided by Eugene Fama, who attributed them to Robert Shiller.

Valuation

Why is the market so high? As an aside, and this concern is not directed toward our topic today of the equity risk premium, but I think it is an interesting question: Why is the market where it is today relative to where it was on September 10 or September 9 or just before the events of September 11, 2001? The market then and now is at about the same level. Almost every economist and analyst has said that the September 11 attacks accelerated a recession, that they changed perceptions of risk, and so forth. It is curious to me that such a situation does not seem to be reflected in market prices.

But in general, why is the market so high? I believe three possible explanations exist. One idea, and I consider it a "rational" theory, is that stocks are simply seen as less risky than in the past. I do not know whether the behavioral theories are rational or not, in the sense that prices are high because of behavioral phenomena that are real and are going to persist. If so, then those phenomena—as identified by Jeremy Siegel and Richard Thaler [see the "Theoretical Foundations" session]—are also rational. In that case, the market is not "too high"; it is not, in a sense, a mistake. It is simply reflecting characteristics of human beings that are not fully explained by economic theories.

Another rational explanation has been given less attention but is the subject of a recent paper by McGrattan and Prescott (2001). It is that the valuation of equities is fundamentally determined by taxation. McGrattan and Prescott argue that the move toward holding equities in nontaxable accounts has led to a drop in the relative tax rate on dividends. Therefore, stock prices should rise relative to the valuation of the underlying capital and expected returns should fall. This effect is a rational tax effect.

Both this theory and the theory that stocks are now seen as less risky say that the market is high because it should be high and that, looking ahead, equities are going to have low expected returns, or low risk premiums—about 2 percent—but that investors have nothing to worry about.

The final explanation, which I attribute to John Campbell and Robert Shiller, focuses on the view that equity prices today are simply a mistake. (I suppose mistakes are a behavioral phenomenon, but presumably, they are not as persistent as an underlying psychological condition.) Now, when people realize they have made a mistake, they attempt to correct the behavior. And those corrections imply a period of *negative* returns from the U.S. equity market before the risk premium can return to a more normal level.

Closing

To close, I want to repeat that, to me, the fundamental historical piece of data that needs more explanation is the relationship between the aggregate behavior of earnings and GNP—what it has been in the past and what it can reasonably be going forward. This relationship is interesting, and I look forward to hearing what all of you have to say about it. In my view, it is the key to unlocking the mystery of the equity risk premium's behavior.

Chapter 2

Evaluating the Historical Record

Primitive peoples, with no knowledge of modern science, express confidence in the proposition that the sun will rise tomorrow. The reason is that the historical record is unambiguous on this point. Ask whether it will rain tomorrow, though, and doubt arises. Because of random variation in weather, the historical record is a good deal more ambiguous. Rain today does not necessarily mean rain tomorrow.

With respect to the equity premium, the confidence that can be placed in the assumption that the future will be like the past depends on two related characteristics of the historical data: how accurately the historical premium can be measured and the extent to which the measured premium depends on the choice of the sample period. Before those questions can be addressed, however, there is the issue of how the average returns that go into the premium should be computed in the first place.

Computing the Average Premium: Arithmetic versus Geometric

The historical equity risk premium equals the difference between the average return on equities and the average return on treasury

securities calculated over a specified time period. It can be seen in Table 1.2, for instance, that over the full sample period between 1926 and 1997, the average return on stocks was 13.0% and the average return on treasury bills was 3.8%, so the equity risk premium over bills was 9.2%. Those are arithmetic averages. They are computed in the standard way: Add up all the annual returns and divide by the numbers of years (in this case, 72).

Although it is familiar, the arithmetic average has a peculiar property. As an illustration, suppose that an investor earns returns of 10%, 20%, –25%, and 15% in 4 consecutive years. The arithmetic average of the four returns is 5%. Now consider an investor who starts with $100. If he or she earns 10%, 20%, –25%, and 15% in each of 4 years, his or her ending wealth will be $113.85. However, if that investor earns 5% per year for 4 years, he or she will end up with $121.55. This is a general problem. Investors who earn the arithmetic average of a series of returns wind up with more money than investors who earn the series of returns that are being averaged.

The geometric average solves this problem. By definition, the geometric average is the constant return an investor must earn every year to arrive at the same final value that would be produced by a series of variable returns. The geometric average is calculated using the formula

$$\text{Geometric Average} = (\text{Final Value}/\text{Initial Value})^{1/n} - 1$$

where n is the number of periods in the average. When the formula is applied to the preceding example, the results are as follows:

$$\text{Geometric Average} = (113.85/100)^{1/4} - 1 = 3.29\%$$

An investor who earns 3.29% for 4 years will end up with $113.85.

There are four properties of arithmetic and geometric averages that are worth noting:

- The geometric average is always less than or equal to the arithmetic average. For instance, in Table 1.2 the arithmetic average stock return is 13.0%, but the geometric average is only 11.0%. (The geometric averages are reported at the bottom of the path of wealth columns in Table 1.2.)
- The more variable the series of returns, the greater the difference between the arithmetic and geometric average. For example, the returns for common stock are highly variable. As a result, the arithmetic average exceeds the geometric average by 200 basis points. For treasury bonds, whose returns are less variable, the difference between the two averages is only 40 basis points.
- For a given sample period, the geometric average is independent of the length of the observation interval.[1] The arithmetic average, however, tends to rise as the observation interval is shortened. For instance, the arithmetic average of monthly returns for the S&P 500 (calculated on an annualized basis by compounding the monthly arithmetic average) over the period between 1926 and 1997 is 13.1%, compared with the 13.0% average of annual returns.
- The difference between the geometric averages for two series does not equal the geometric average of the difference. Consider, for instance, stock returns and inflation. Table 1.2 reveals that the geometric average stock return is 11.0% and the average inflation rate is 3.1%, for a difference of 7.9%. However, Table 1.3 shows that the geometric average real return on common stock was 7.7%. This discrepancy does not arise for arithmetic averages, where the mean difference always equals the difference of the means.

With respect to the equity risk premium, the manner in which the average is calculated makes a significant difference. When compared with treasury bills over the full 1926-to-1997 period, the arithmetic average risk premium is 9.2%, whereas the geometric average premium is only 7.2%. Which average is the more appropriate choice? That depends on the question being asked. Assuming that the returns being averaged are largely independent and that the future is like the past, the best estimate of expected returns over a given future holding period is the arithmetic average of past returns over the same holding period. For instance, if the goal is to estimate future stock-market returns on a year-by-year basis, the appropriate average is the annual arithmetic risk premium. On the other hand, if the goal is to estimate what the average equity risk premium will be over the next 50 years, the geometric average is a better choice. Because the ultimate goal in this book is to arrive at reasonable forward-looking estimates of the equity risk premium, both arithmetic and geometric averages are employed where they are useful.

It is worth reiterating that projection of any past average is based on the implicit assumption that the future will be like the past. If the assumption is not reasonable, both the arithmetic and geometric averages will tend to be misleading.

How Accurately Can the Historical Risk Premium Be Measured?

The accuracy with which the historical risk premium can be measured depends on the variability of the observations from which the average is calculated. In an assessment of the impact of that variability, the best place to start is with an expanded version of Table 1.2 that includes monthly returns for the four asset classes over the period between 1926 and 1997. Given this expanded data set, one way to assess the variability of the ex-post risk premium, defined as the difference between the observed returns for stocks and the related treasury securities, is to plot one histogram for stocks versus bonds and another for stocks versus bills. Each bar on the histogram represents the fraction of the 864 monthly

[1] This follows immediately from the fact that the geometric average depends only on the initial and final values of the investment.

WILEY FRONTIERS IN FINANCE

THE EQUITY RISK PREMIUM

The Long-Run Future of the Stock Market

BRADFORD CORNELL

Chapter 6

The Equity Risk Premium and the Long-Run Outlook for Common Stocks

So that there is no suspense, here is the bottom line: The future will not be as bright as the past. The data of Ibbotson Associates showed that over the period from 1926 to 1997, the average equity risk premium was 7.4% over treasury bonds and 9.2% over treasury bills. Investors cannot reasonably expect equities to produce such large premiums going forward. Instead, premiums are much more likely to be on the order of 300 to 400 basis points lower. Reasonable forward-looking ranges for the future equity risk premium in the long run are 3.5% to 5.5% over treasury bonds and 5.0% to 7.0% over treasury bills.

This relatively pessimistic conclusion is based on two considerations. The first is an overall assessment of the empirical data and theoretical arguments presented in Chapters 1 through 4. The second is the analysis of the level of stock prices presented in Chapter 5. Although forecasting future stock returns, even over the long run, is hazardous at best, when all the evidence is taken into account, the conclusion that the future will be less rosy than the past has strong support.

Investment Banking Relationships and Analyst Affiliation Bias: The Impact of Global Settlement on Sanctioned and Non-Sanctioned Banks

Shane A. Corwin*
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556
scorwin@nd.edu

Stephannie Larocque
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556
larocque.1@nd.edu

Mike Stegemoller
Hankamer School of Business
Baylor University
Waco, TX 76798
michael_stegemoller@baylor.edu

January 2015

Abstract

We examine the impact of the Global Analyst Research Settlement on analyst affiliation bias in stock recommendations. Using a comprehensive measure of investment bank-firm relationships, including equity and debt underwriting and M&A advising, we find that affiliation bias is substantially reduced, but not eliminated, for analysts employed by banks named in the settlement. In contrast, we find strong evidence of analyst affiliation bias both before and after the Global Settlement for analysts at non-sanctioned banks. The results hold after controlling for shifts in the recommendation schemes used by investment banks and are robust to alternative empirical specifications.

JEL classification: G10, G24, G34, L14

Keywords: Analysts, Recommendations, Investment Banking, Investment Banking Relationships

* We thank Robert Battalio, Larry Brown, Gus De Franco, Marcus Kirk, Tim Loughran, and Hai Lu for helpful comments. Steven Carroll, Brian Ford, and Travis Johnson provided excellent research assistance. Any remaining errors are the responsibility of the authors.

1. INTRODUCTION

Conflicts of interest within investments banks and other financial institutions have been the subject of numerous academic studies (see Mehran and Stulz 2007 for a discussion). One particular conflict that has received significant attention from both regulators and academics is analyst affiliation bias. Specifically, prior research provides strong evidence that analysts are overly optimistic when their employers have equity underwriting relationships with the covered firms. Early in the 2000s, several attempts to reduce conflicts of interest were implemented in the securities industry, culminating in the 2003 Global Analyst Research Settlement (Global Settlement). In particular, a major purpose of the Global Settlement reached between the SEC, NYSE, NASD, New York Attorney General, and North American Securities Administrators Association and 12 of the largest investment banks was to reduce the conflicts of interest between the investment banking and research departments within the major banks.[1] Subsequent research suggests that investment banks changed their behavior following the Global Settlement[2], but provides little evidence on affiliation bias for analysts employed by sanctioned and non-sanctioned banks nor on relationships beyond the well-studied equity underwriting relationship. In this study, we use a broad measure of investment banking relationships, including equity and debt underwriting and mergers and acquisitions (M&A) advising, to examine analyst affiliation bias for a large sample of sanctioned and non-sanctioned investment banks (IBs) in the periods before and after the Global Settlement and contemporaneous regulatory changes.

Sell-side financial analysts provide buy/sell recommendations and earnings forecasts for a set of covered firms. In general, analysts are compensated and earn a reputation based on the quality of the information they provide. Despite these incentives to produce accurate information, however, analysts can also face pressure to issue optimistic or biased coverage. In particular, the financial services firms that employ analysts also compete for lucrative underwriting and M&A advisory mandates and may seek to

[1] The original settlement in April 2003 named ten investment banks, including Bear Stearns, CSFB, Goldman Sachs, Lehman Brothers, J.P. Morgan, Merrill Lynch, Morgan Stanley, Citigroup (Salomon Smith Barney), UBS Warburg, and U.S. Bancorp Piper Jaffray. Similar settlements with Deutsche Bank and Thomas Weisel were added later. We refer to these banks (including other name variations of the same banks) as "sanctioned" banks.

[2] See, for example, Kadan, Madureira, Wang, and Zach (2009).

use biased coverage as one means of winning potential clients. As a result, analysts face a conflict between their role in providing quality information to financial markets (and the associated reputational concerns) and the motivations of their employers to win future investment banking business.

Following prior research, we define an affiliated analyst as one whose employer also has an investment banking relationship with the covered firm. Existing research suggests that affiliated analysts tend to produce optimistic (i.e., upward biased) recommendations and earnings forecasts relative to unaffiliated analysts (see, for example, Dugar and Nathan 1995, Lin and McNichols 1998). This research focuses primarily on affiliation through equity underwriting relationships, with a particular emphasis on affiliation at the time of an equity issue.[3] However, equity underwriting is only one of many services that investment banks provide to firms. In the fourth quarter of 2013, for example, equity underwriting accounted for only 36% of total investment banking revenues at Goldman Sachs, compared to 34% for financial advising and 30% for debt underwriting. This suggests that investment banking relationships may have an impact beyond that evidenced through equity underwriting.[4]

To better understand the impact of investment banking relationships on analyst behavior, we examine the individual equity, debt, and M&A components of the relationship, as well as the overall investment banking relationship. We expect the results to be strongest for the overall relationship for two reasons. First, since equity, debt, and M&A transactions are discrete observations of the firm-bank relationship, viewing all of these transactions together allows us to observe the relationship at more points in time, better capturing the ongoing nature of the relationship. Second, we expect investment banking relationships that span multiple functional areas to put more pressure on analysts than narrow relationships.

To analyze affiliation bias, we study recommendations on a large sample of U.S. non-financial firms between 1998 and 2009 by analysts whose employers are either sanctioned investment banks or top

[3] One exception is Ljungqvist, Marston, Starks, Wei, and Yan (2007) who control for both equity and debt underwriting affiliations. This study is discussed in more detail below.

[4] The importance of firm-wide relationships may also change over time. For example, Corwin and Stegemoller (2014) find that the tendency of firms to use the same investment bank in multiple functional areas (i.e., equity underwriting, debt underwriting, or M&A advising) has increased significantly over time.

non-sanctioned banks. Our main variable of interest is the analyst's relative recommendation, defined as the difference between the analyst recommendation (with strong buy=5 and strong sell=1) and the median recommendation across all analysts covering the stock. Following Ljungqvist et al. (2007), we construct this variable at the end of each quarter, using the most recent recommendation by each analyst during the preceding twelve months. In our main tests, we regress this variable on proxies for investment banking relationships and a set of control variables shown in prior literature to have an association with analyst recommendations. Our primary relationship variable is an indicator variable equal to one if, during the prior three years, the firm hired the investment bank as a lead or co-manager on an equity or debt deal or as an advisor on an M&A transaction. However, we also provide tests using a continuous measure of relationships, defined as the proportion of a firm's total transaction value during a three-year window for which the investment bank acted as a lead manager, co-manager, or advisor. We define these relationship variables separately for equity, debt, and M&A transactions, as well as for the combined set of transactions across all types.

Consistent with prior research, we find strong evidence of analyst affiliation bias prior to the Global Settlement in 2003. For banks named in the Global Settlement (sanctioned banks), this bias is evident for all individual transaction types and for the overall relationship measure. For non-sanctioned banks in the period prior to the Global Settlement, we find mixed evidence of an affiliation bias based on individual transaction type relationship measures, but strong evidence of an affiliation bias based on the overall relationship measure. This evidence is consistent with our prediction that the overall measure better captures the ongoing nature of the investment banking relationship. The more striking results appear during the period following the Global Settlement. During this period, there remains evidence of an affiliation bias for sanctioned banks, but the bias is substantially reduced from the pre-Global Settlement effect. In contrast, non-sanctioned banks continue to exhibit strong analyst affiliation bias even after the Global Settlement. This bias is evident across all types of transactions and for the overall relationship measure. These results suggest that while the Global Settlement was successful at reducing analyst affiliation bias for the banks named in the settlement, conflicts of interest persist, especially for non-sanctioned investment banks.

Our results are robust to several alternative specifications and robustness checks. While our main results are based on relationship indicator variables, we find similar results based on continuous measures of relationships. The results are also robust to alternative fixed effects specifications, including firm, analyst, and investment bank fixed effects. Most importantly, our results are not driven by the shift of many investment banks from a five-tier to a three-tier recommendation scheme following the Global Settlement (Kadan et al. 2009). We find similar results when we repeat our analysis on a relative recommendation variable based on a three-tier recommendation scheme.

As an alternative specification, we use logistic regressions to examine the impact of investment banking relationships on the likelihood of issuing a buy or strong buy and the likelihood of issuing a sell or strong sell. Consistent with the relative recommendation results, this analysis suggests that prior to the Global Settlement, analysts at both sanctioned and non-sanctioned banks were significantly more likely to issue a buy or strong buy recommendation and significantly less likely to issue a sell or strong sell recommendation when affiliated with the firm through an investment banking relationship. After the Global Settlement, the bias for sanctioned banks is reduced, but remains significant. For non-sanctioned banks, the bias is significant both before and after the Global Settlement. For both groups of banks, the logit results suggest that a significant affiliation bias remains following the Global Settlement, with the effect being substantially larger for non-sanctioned banks.

As a final test, we examine whether incorporating lending data has an impact on the measurement of analyst affiliation bias. We find only weak evidence that lending relationships have an incremental effect on the measurement of analyst affiliation bias. Thus, affiliation bias appears to be best captured through the equity, debt, and M&A relationships. We assert that an overall measure, incorporating equity underwriting, debt underwriting, and M&A advising, is better able to capture investment banking relationships and their effects than measures based on any one type of transaction.

In summary, our findings suggest that conflicts of interest within investment banks have not been completely eliminated by the Global Settlement and contemporary regulatory changes. Our results suggest that the Global Settlement reduced, but did not eliminate, analyst affiliation bias in recommendations from banks named in the Global Settlement. Further, for large banks not named in the Global Settlement, we find strong evidence of a continued affiliation bias in the post-Global Settlement period. This suggests that our findings are driven by the punitive and bank-specific requirements imposed by the Global Settlement, rather than the broader regulatory changes that accompanied the settlement.

The remainder of the paper is organized as follows. Section 2 summarizes the literature related to analyst affiliation bias, provides background information on the Global Settlement, and describes our main hypothesis. In Section 3, we describe our data and sample construction. Section 4 presents our main results related to analyst affiliation bias and Section 5 examines the incremental impact of lending relationships. Section 6 concludes.

2. BACKGROUND AND HYPOTHESIS DEVELOPMENT

2.1. Analyst Affiliation Bias

Sell-side financial analysts have been widely studied as proxies for the market's expectations. At the same time, however, analysts' recommendations, target prices, and forecasts have been shown to be optimistic (Beneish 1991; Bradshaw 2004; La Porta 1996). In particular, prior research provides strong evidence of a link between analyst optimism (or bias) and investment banking relationships between covered firms and the banks that employ analysts. Dugar and Nathan (1995) find that recommendations and earnings forecasts are more optimistic for analysts who also have an investment banking relationship with the covered firm than for non-affiliated analysts and Lin and McNichols (1998) show that analysts employed by lead and co-managing underwriters issue growth forecasts and recommendations on the issuing firms that are significantly more favorable than those made by unaffiliated analysts. Further, Dechow, Hutton, and Sloan (2000) provide evidence that analysts employed by lead managers of equity offerings make more optimistic long-term growth forecasts around equity offerings and O'Brien, McNichols, and Lin (2005) conclude that investment banking relationships increase analysts' reluctance to reveal negative news.

Prior studies also point to factors that appear to mitigate analyst affiliation bias. Cowen, Groysberg, and Healy (2006) find that the bias is lower for bulge bracket investment banks than for lower-tier banks, suggesting that the reputational concerns of bulge bracket banks outweigh the benefits of biased analyst coverage. Ljungqvist et al. (2007) argue that, because analysts rely on institutional investors for trading commissions and ratings, they will be less likely to produce biased coverage on affiliated stocks that are also highly visible to institutional investors. Their results confirm that relative recommendations are negatively related to the presence of institutional investors.

Other research examines the impact of analyst bias on investors and the post-recommendation performance of covered firms. De Franco, Lu, and Vasvari (2007) examine the investor consequences of analysts' misleading behavior in the period prior to the Global Settlement. Using a sample of 50 firm-events identified in the Global Settlement in which analysts' private beliefs differed from their public disclosures, they provide evidence that these events are associated with selling by sophisticated investors and a wealth transfer from individuals to institutions. Michaely and Womack (1999) report that in the month following the post-IPO quiet period, affiliated analysts issue more buy recommendations for the IPO firm than do unaffiliated analysts, and the IPOs recommended by affiliated analysts substantially under-perform IPOs recommended by unaffiliated analysts. Similarly, Barber, Lehavy, and Trueman (2007) find that the "buy" and "strong buy" ratings of IB-employed analysts tend to underperform those of other analysts.

Research also examines whether analyst coverage affects the investment bank's ability to win future business from the covered firm. Bradshaw, Richardson, and Sloan (2006) surmise that all analysts bias their recommendations and forecasts in an attempt to win underwriting business. Ljungqvist, Marston, and Wilhelm (2006) find little evidence that optimistic analyst coverage affects an investment bank's likelihood of winning future lead underwriting mandates. However, Ljungqvist, Marston, and

Wilhelm (2009) show that optimistic analyst coverage does increase the likelihood of winning future co-managing appointments, which in turns leads to an increased likelihood of future lead mandates.

Existing research focuses primarily on affiliation through equity underwriting relationships. However, some recent research extends the analysis of affiliation bias to other areas. Ljungqvist et al. (2007) examine both equity and debt underwriting relationships and find that affiliation bias is stronger with respect to equity relationships. Kolasinski and Kothari (2006) investigate affiliation bias in analyst recommendations issued around M&A deals. They find that analysts affiliated with acquirer advisors upgrade acquirer stocks around M&A deals and target-affiliated analysts issue optimistic coverage on acquirers after exchange ratios (for all-stock deals) have been set.

2.2. The Global Settlement

During 2000, the securities industry attempted to reduce investment banking conflicts of interest, with the Securities Industries Association endorsing best practices around research and investment banking and the Association for Investment Management and Research (since renamed CFA Institute) releasing a white paper titled "Preserving the Integrity of Research." In 2002, the Sarbanes-Oxley Act (SOX) amended the Securities and Exchange Act of 1934 with the creation of Section 15D, which required the NYSE and the NASD to adopt rules designed to address research analysts' conflicts of interest. To comply with SOX, in 2002 the NYSE amended its Rule 351 (Reporting Requirement) and Rule 472 (Communication with the Public), while the NASD released Rule 2711 (Research Analysts and Research Report).[5] These rules were approved by the SEC in May 2002.

In 2001, following allegations of research tainted by investment banking conflicts of interest, the

[5] NYSE Rule 472 (Communication with the Public) requires that research reports be approved by a supervisory analyst, that research analysts not be subject to the supervision of any member of the investment banking department, that research analysts not purchase issuer securities prior to an IPO, that an IB not distribute research regarding an issuer 40 calendar days following an IPO offering in which the IB acted as a manager or co-manager, that an IB not issue a favorable research report in return for business, that analysts not receive compensation for investment banking business, and that the above be disclosed in the analyst's research reports. NYSE Rule 351(f) requires an annual letter of attestation by the investment bank that it is in compliance with Rule 472. Similarly, NASD Rule 2711 (Research Analysts and Research Report) restricts relationships between investment banking and research departments and restricts the review of research reports by the subject company. It also prohibits analyst compensation based upon investment banking services, prohibits the promise of favorable research, imposes a 40 (10) day quiet period for research following an IPO (SEO), restricts personal trading by analysts in their covered stocks, and requires additional disclosures in research reports as well as additional supervisory procedures at the investment bank.

New York Attorney General began investigating Merrill Lynch and, subsequently, several other large investment banks. This investigation culminated in April 2003 with the Global Analyst Research Settlement reached by the SEC, NYSE, NASD, New York Attorney General, and North American Securities Administrators Association with ten of the largest investment banks – Bear Stearns, CSFB, Goldman, Lehman, J.P. Morgan, Merrill Lynch, Morgan Stanley, Citigroup (Salomon Smith Barney), UBS Warburg, and U.S. Bancorp Piper Jaffray (with Deutsche Bank and Thomas Weisel added later).[6] The Global Settlement required the payment of $875 million in penalties and disgorgement, $432.5 million to fund independent research, and $80 million to fund investor education. In addition, the settlement made numerous structural reforms including the physical separation of investment banking and research departments, the inability to compensate research analysts based upon investment banking revenues, and the prohibition of research analysts taking part in investment banking pitches and roadshows.

Subsequent research suggests that these regulatory changes affected the behavior of analysts within investment banks. Kadan, Madureira, Wang, and Zach (2009) find that the overall informativeness of recommendations (measured using absolute price reactions) declined following the Global Settlement. They also document that sanctioned banks shifted their stock recommendations from a 5-tier scale to a 3-tier scale. Barniv, Hope, Myring, and Thomas (2009) and Chen and Chen (2009) both document that the mapping between analysts' forecasts and target prices improved following the regulatory changes of the early 2000s. Clarke, Khorana, Patel, and Rau (2011) investigate market reactions to independent, affiliated, and unaffiliated analysts before and after the Global Settlement. They find that affiliated (independent) analysts issued fewer (more) strong buys following the settlement, with recommendation upgrades by affiliated analysts being more informative in the post-period. Moreover, Guan, Lu, and Wong (2012) find that forecasts by research firms are more optimistic than those of brokerage firms, syndicate firms, and investment banks following the regulatory changes in the early 2000s, but that forecast

[6] See http://www.sec.gov/news/press/2003-54.htm for the April 2003 press release and http://www.sec.gov/news/press/2003-144.htm for the SEC's October 2003 approval of Global Settlement.

accuracy and recommendation profitability for research firms are not significantly different from those of investment banks after the reforms.

Despite these behavior changes, there is some evidence that the Global Settlement may not have eliminated analyst affiliation bias. Using data from 1994 through 2008, Malmendier and Shanthikumar (2014) distinguish between strategic and non-strategic distortions in analyst behavior. Consistent with their expectations for strategic behavior, they find that affiliated analysts tend to issue more positive recommendations, but similar or more negative forecasts, than unaffiliated analysts.[7] In a recent survey of sell-side analysts, Brown, Call, Clement, and Sharp (2014) report that analysts view the generation of investment banking business as an important driver of their compensation and feel pressure from their research management to issue optimistic forecasts and/or recommendations. Recent actions by FINRA against Citigroup and Goldman Sachs also provide evidence of analyst involvement in IPO road shows and of analysts tipping selected clients, even after the Global Settlement.

2.3. Hypothesis

We contribute to the literature on analyst affiliation bias by examining the differential impact of the Global Settlement and contemporaneous regulatory changes on affiliation bias for sanctioned and non-sanctioned banks. We also provide a detailed analysis of the link between affiliation bias and the equity, debt, and M&A components of investment banking relationships. Our primary hypothesis is that analyst affiliation bias was eliminated following the Global Settlement. However, by separating sanctioned and non-sanctioned banks, we are able to examine two variations of this hypothesis. If the Global Settlement and concurrent regulatory changes imposed on the industry eliminated the conflicts of interest within investment banks that lead to analyst affiliation bias, we expect the bias to be eliminated for both sanctioned and non-sanctioned banks. However, if the principal effects of the Global Settlement

[7] Although not the main subject of our analysis, we also examined the relation between investment banking relationships and the bias and accuracy of analyst earnings forecasts. We define bias and accuracy by comparing each analyst's most recent forecast to actual earnings, where bias and accuracy are scaled by the standard deviation of forecasts across all analysts following the stock and normalized by subtracting the consensus (median) level of bias/accuracy. We find some evidence of optimistic forecasts by GS banks in the period prior to Global Settlement, but little evidence of a link between investment banking relationships and forecasts for GS banks in the post period or for non-sanctioned banks in either the pre or post period. We find little evidence of a consistent relation between analyst affiliation and forecast accuracy for either class of banks.

result from the punitive aspects or bank-specific requirements of the settlement, we expect affiliation bias to be eliminated only for sanctioned banks. We test these alternative versions of the hypothesis below.

3. Data and Sample Characteristics

To construct our sample, we use two main data sources. First, we use SDC to identify all equity, debt, and M&A activity by a large sample of U.S. firms, allowing us to measure the relationships between firms and their investment banks. Second, we use I/B/E/S data to identify the stock recommendations of sell-side analysts and the brokerage firms for which the analysts work. Together, these two datasets allow us to provide a detailed examination of the link between analyst recommendations and investment banking relationships both before and after the Global Settlement.

3.1. Sample Firms and Investment Banking Activity

We begin with the sample of all U.S. firms with listed common stock (CRSP share codes 10 or 11) between 1996 and 2009. After eliminating financials, utilities, and government agencies, the resulting sample includes 8,322 unique firms. For these firms, we then use the Securities Data Company (SDC) database to collect information on all public and private issues of equity and debt by the firm and any M&A transactions in which the firm is either the acquirer or the target. Firms are identified based on PERMCO in the CRSP data and based on CIDGEN in the SDC data. Firms are matched between the two databases using Cusip and, where possible, Ticker. To provide meaningful analysis of investment banking relationships, we exclude transactions for which either the transaction value or the identity of the underwriter/advisor is missing.

To identify affiliation through investment banking relationships, we focus on the most important investment banks in the sample. To identify these banks, we begin with the full sample of banks identified as lead or co-managing underwriters in the equity and debt samples or as advisors in the M&A sample.[8] We then compute market share ranks on an annual basis for each transaction type (equity, debt, and M&A). Finally, we compute each bank's average market share rank in each transaction type category

[8] Investment bank names are cleaned to eliminate multiple variations of the same investment bank name and to adjust for mergers and acquisitions among investment banks.

across all years during which the bank appears in the sample and limit our analysis to those investment banks with an average market share rank of 25 or higher in at least one transaction type category. In cases where one of the top 25 banks reflects the merger of two or more predecessor banks, all predecessor banks are also included. As shown in Table A2 in the Appendix, the resulting sample includes 57 different investment bank names during the sample period, with 48 active at the beginning of the sample period and 28 active at the end of the sample period.[9]

3.2. Analyst Recommendations

To test analyst affiliation bias, we focus on analyst stock recommendations, one of the analysts' primary and most visible outputs. We collect recommendations data, including the identity of the broker employing the analyst, from I/B/E/S. We then link the recommendations to the sample of CRSP firms using CUSIPs and hand-match the broker names in I/B/E/S to the sample investment banks using the I/B/E/S broker translation file.

Following Ljungqvist et al. (2007) we examine recommendations at a quarterly frequency. For each calendar quarter end and each firm in our sample, we select the most recent recommendation issued during the preceding 12 months by each analyst covering the stock. We code recommendations as 1 (strong sell) through 5 (strong buy). We then define each analyst's relative recommendation, *RelRec*, by subtracting the consensus (i.e., median) recommendation across all analysts covering the firm in the same one-year window.[10] Finally, we limit our sample to stocks covered by at least one analyst employed by a sample investment bank. The resulting sample includes 216,242 quarterly observations, involving 4,628 analysts and 5,111 sample stocks.

3.3 Variable Construction and Sample Characteristics

Our main empirical tests examine the relation between the relative recommendations of analysts

[9] For clarity following large investment bank mergers, we assign a new name to the combined bank. For example, we refer to the combination of Citibank and Salomon Smith Barney as Citigroup Salomon Smith Barney and the combination of UBS Warburg and Paine Webber as UBS Paine Webber. The 28 ultimate banks considered here compares to 16 studied in Ljungqvist et al. (2006) and Ljungqvist et al. (2007). Lehman and Merrill Lynch are eliminated from the sample because their recommendations are excluded from the I/B/E/S database for all or part of our sample period.

[10] In order to compute relative recommendations, our sample is restricted to firms that are followed by two or more analysts. As discussed in Section 3 below, we also provide robustness tests based on a redefined three-point recommendation scale. Our main conclusions are robust to this alternative specification.

(*RelRec*) and investment banking relationships between the analyst's firm and the covered stock, after controlling for firm, analyst, and investment bank characteristics that have been shown to affect recommendations. Our empirical model closely follows that in Ljungqvist et al. (2007), with several important differences. First, we examine investment banking relationships across a wider set of transaction types, including equity, debt, and M&A transactions. Second, we define relationships both within specific functional areas and across all functional areas. Finally, we examine affiliation bias both before and after the Global Settlement, allowing for differences between investment banks named in the Global Settlement and other banks. Table A1 in the Appendix contains all variable definitions.

Summary statistics for our sample of quarterly observations are provided in Panel A of Table 1. Consistent with previous research, we find that analysts primarily issue "buy" or "strong buy" recommendations, giving a mean (median) analyst recommendation across our sample of 3.6 (4.0). As noted earlier, our main variable of interest is the relative recommendation of the analyst (*RelRec*), defined as the difference between the analyst's recommendation and the consensus (i.e., median) recommendation across all analysts following the stock. *RelRec* has a range from -4 to +3, with a mean (median) of 0.0025 (0.0000) across our sample observations.

To proxy for investment banking relationships, we examine each firm's equity, debt, and M&A transactions during the 36 months preceding each quarter end. We then define relationship dummy variables (*IBRel*) for each investment bank-firm pair that equal one if the investment bank acted as lead or co-managing underwriter on an equity or debt issue, or as an advisor on an M&A transaction. While the majority of our tests are based on these relationship dummy variables, we also analyze continuous relationship variables based on the proportion of each firm's equity, debt, and M&A transaction value for which the bank acted as lead or co-managing underwriter, or advisor.

We define relationship measures both by transaction type (equity, debt, or M&A) and across all combined transactions (overall relationship).[11] We expect affiliation bias to be better captured by overall relationships than by type-specific relationships for two reasons. First, equity, debt, and M&A transactions are discrete measures of what is likely an ongoing relationship. Thus, the use of multiple transaction types will better capture the ongoing nature of any underlying relationship. Second, if there is any pressure placed on the analyst to produce optimistic coverage, then this pressure will only be magnified when the investment banking relationship spans multiple functional areas.

To illustrate the potential benefits of the overall relationship measure, Figure 1 plots the time series of relationships between Convergys Corp. and Citi-Salomon-Smith, based on 36-month windows. Convergys used this bank as a lead equity underwriter on their August 1998 IPO, as a lead debt underwriter in September 2000 and December 2004, and as an M&A advisor in April 2001. When we incorporate all three transaction types, we are able to capture the ongoing nature of the relationship between Convergys and Citi-Salomon-Smith over the entire period from 1998 through 2007. However, when we define relationships based on any individual transaction type (equity, debt, or M&A) the relationship measure is spotty and only covers sub-periods from August 1998 through December 2007.

Summary statistics for our type-specific and overall relationship measures are provided in the second section of Table 1. Across all quarterly observations, the mean transaction type-specific relationship ranges from 2.43% for M&A transactions to 3.24% for equity transactions. Incorporating all transaction types, the mean overall relationship is 5.90%. In untabulated results, we find that the proportion of quarterly observations with no relationship equals 87.2% for the overall relationship measure, compared to 93.5% for equity, 93.6% for debt, and 96.3% for M&A. This provides one indication that the overall relationship measure may better identify ongoing relationships in cases where type-specific relationship measures do not.

Our remaining control variables are motivated by prior literature and closely follow the specification in Ljungqvist et al. (2007). To control for investment bank characteristics, we define two

[11] For the overall relationship variable, we measure at each quarter end date the proportion of a firm's combined equity, debt, and M&A transaction value during the preceding 36 months for which each investment bank acted as lead underwriter, co-managing underwriter, or adviser, and an indicator variable for whether this value is greater than zero.

continuous variables and a set of indicator variables. We define investment bank size (*IB_Size*), as the number of analysts employed by the investment bank during quarter *t*, based on I/B/E/S recommendations.[12] Investment bank market share, *IB_MktShare*, is the proportion of total deal value across all firms during the previous 12 months for which the investment bank acted as a lead or co-managing underwriter or M&A advisor. Like the relationship measures, *IB_MktShare* is defined by transaction type (equity, debt, or M&A) and across all combined transactions (overall). As shown in Table 1, the mean (median) number of analysts employed by an investment bank is 89 (85) and investment bank market shares average 4.55%, 4.77%, and 4.38% for equity, debt, and M&A, respectively. We also define two indicator variables, *IB_GS* and *IB_NonGS*, to distinguish between those investment banks sanctioned in the Global Settlement (including subsequent name variations of the same banks) and other non-sanctioned banks, respectively. Based on this categorization, 57% of our quarterly observations are from sanctioned banks and 43% from non-sanctioned banks. Appendix Table A2 lists the sample investment banks in each category.

We define six analyst-level characteristics. Four of these variables are defined directly from the I/B/E/S recommendations data. *Seniority* is the number of years since the analyst first appeared in I/B/E/S and *Seasoning* is the number of years since the analyst initiated coverage on the particular stock. *NFollow* is the number of firms followed by the analyst during the quarter and *JobMove* is an indicator variable that equals 1 if the analyst changed employers during the quarter. Following Hong and Kubik (2003) and Ljungqvist et al. (2007), we define relative forecast accuracy (*RelAccuracy*) based on the analyst's average earnings forecast accuracy across all followed stocks.[13] Finally, *AllStar* is an indicator variable that equals 1 if the analyst is a ranked as an All-Star by *Institutional Investor* magazine during year *t-1*,

[12] Ljungqvist et al. measure investment bank size as the number of registered representatives employed by the IB.
[13] For each analyst following each firm, we first estimate the absolute value of the difference between the analyst's most recent forecast of fiscal-year earnings and actual earnings, scaled by prior year price. We then rescale such that the most accurate analyst following the firm scores 1 and the least accurate analyst scores 0. Finally, each analyst's relative forecast accuracy is defined as their mean score across all stocks followed over years *t*-2 through *t*. See Appendix Table A1 for a more complete description.

and 0 otherwise. For the mean (median) observation in our sample the analyst has seniority of 5.4 (4.9) years, seasoning of 2.3 (1.4) years, and follows 11 (10) stocks. The mean and median values of relative accuracy are 41.23% and 40.96%, respectively. Finally, 18.9% of the recommendation observations in our sample are issued by All-Star analysts and 3.2% by analysts that changed employers during the quarter.

Our last set of control variables is related to firm characteristics. *ANF* is the number of analysts issuing recommendations for the firm during the previous 12 months, based on I/B/E/S recommendations. *MV* is the firm's market value of equity at the end of the prior calendar year, as defined by CRSP. *InstHoldings* is the percentage of shares held by institutional investors at the end of the quarter, based on Thomson Reuters' 13F filings. Lastly, *Proceeds* is the total value of transaction by the firm during the previous 36 months, defined for each transaction type (equity, debt, or M&A) and across all combined transactions (overall). Across all observations in our sample, mean (median) values are 11 (1) for analyst following, $9.6 ($1.9) billion for market capitalization, and 62% (70%) for institutional holdings. Three-year proceeds average $77 million, $428 million, and $1,055 million for equity, debt, and M&A, respectively. Across quarterly observations with non-zero proceeds, these averages increase to $300 million, $1,145 million, and $2,981 million.

Panel B of Table 1 provides mean values of all variables for the subsamples of observations involving sanctioned and non-sanctioned banks. As expected, sanctioned banks tend to be larger and have higher market shares than non-sanctioned banks. For example, the mean values of *IB_Size* (i.e., number of analysts) and equity market share are 116.2 and 7.2% for sanctioned banks, compared to 52.1 and 1.01% for non-sanctioned banks. Other categories of market share and measures of investment banking relationships provide similar results. Analyst and firm characteristics also differ significantly between the two groups of banks, though the differences are smaller economically than the differences in bank size and market share. Analysts employed by sanctioned banks are more likely to be ranked as All Stars, have higher seniority and seasoning, and follow more stocks than analysts employed by non-sanctioned banks.

In addition, analysts employed by sanctioned banks tend to follow larger stocks, with higher institutional ownership and more equity, debt, and M&A activity. While forecast bias and accuracy are similar across the two groups of analysts, recommendations and relative recommendations tend to be higher for analysts at non-sanctioned banks, on average. As a result, we control for differences between sanctioned and non-sanctioned banks in our analysis to follow. Despite the observed differences described above, non-sanctioned banks and the firms that hire them are involved in a significant fraction of equity, debt, and M&A activity over our sample period and account for a large fraction (43%) of the quarterly analyst observations in our data.

To highlight the relation between investment banking relationships and analyst recommendations, Figure 2 plots the frequency of various recommendations for sanctioned and non-sanctioned banks across the entire sample of quarterly observations. Frequencies are further categorized by whether or not the analyst was affiliated with the covered firm, where affiliation is defined based on the overall investment banking relationship over the previous 36 months. Results for the period prior to the Global Settlement are provided in Panel A and results for the period following Global Settlement are provided in Panel B.

The plots on the left show frequencies based on a 5-tier recommendation scale. From these graphs, it is clear that Sell and Strong Sell recommendations are rare in the period before the Global Settlement. While negative recommendations are more common in the post period, they remain relatively rare. Most importantly, the graph shows that affiliated analysts are more likely to issue Strong Buy recommendations and less likely to issue Hold or Sell recommendations than unaffiliated analysts. Although the bias is reduced in the period after the Global Settlement, it does not appear to be eliminated for either sanctioned or non-sanctioned banks, and remains particularly strong for non-sanctioned banks.

Kadan et al. (2009) note that, following the Global Settlement, many large investment banks shifted from 5-tier to 3-tier recommendation schemes. This shift is also evident in our data. For example, from 1998-2001, Deutsche Bank's investment recommendations included the five categories: Strong Buy, Buy, Hold, Underperform, and Sell. In contrast, from 2004-2009, Deutsche Alex Brown's investment recommendations included the three categories: Buy, Hold, and Sell. To ensure that our results are robust to this shift in recommendation schemes, we reassign all recommendations to a 3-tier scale. Frequencies based on this redefined scale are shown on the right side of Figure 2. The results from this redefined scale are consistent with those from the 5-tier scale, with affiliated analysts being less likely to issue Sell or Hold recommendations and more likely to issue Buy recommendations.

The results in Figure 2 suggest that analyst affiliation bias persists following the Global Settlement. However, these frequencies do not control for other factors that may affect analyst recommendations. In the next section, we therefore analyze analyst recommendations in a multivariate framework.

4. Results

In this section, we describe our main results related to analyst affiliation bias. Using the quarterly data described above, we estimate variations of the following general model specification:

$$\begin{aligned} RelRec_{ijkt} = \alpha + \beta_1 \times IB\_GS + \beta_2 \times IB\_NonGS + \beta_3 \times IBRel_{jkt} \times IB\_GS + \beta_4 \times IBRel_{jkt} \times IB\_NonGS \\ + \sum_{i=1}^{I} \delta_i \times AnalystChar_i + \sum_{j=1}^{J} \gamma_j \times IBChar_j + \sum_{k=1}^{K} \lambda_k \times StockChar_k + \varepsilon_{ijkt}, \end{aligned} \quad (1)$$

where $IBRel_{jkt}$ indicates an investment banking relationship between investment bank j and firm k during the 36 months ending in quarter t, and the remaining variables represent controls for analyst, investment bank, and stock characteristics. Our main tests are based on a comparison of the relationship interaction terms involving *IB_GS* and *IB_NonGS*, which are dummy variables that distinguish between investment banks that were and were not sanctioned in the Global Settlement, respectively. To examine the impact of the Global Settlement on analyst affiliation bias, we provide two sets of analysis. In the full period analysis, we interact the relationship variables with a dummy variable equal to one for all quarters after the Global Settlement and zero otherwise. We also provide separate analyses for the sub-periods 1998-2001 and 2003-2009. Following Kadan et al. (2009), we define the implementation date for the Global Settlement as September 2002, but because the investigations related to investment banking conflicts of interest were ongoing during 2002, we exclude 2002 from the sub-period analysis. Our general specifications also include year and firm fixed effects.

4.1 Relative Recommendations and Investment Banking Relationships

The full period regression results are presented in Table 2. *P*-values based on robust standard errors clustered by firm are reported below the coefficients. Examining the coefficients on the control variables, we see that relative recommendations are lower for large investment banks and for analysts that cover a large number of stocks, and higher for more experienced analysts and for stocks followed by a large number of analysts. Investment bank market share is positively related to relative recommendations for equity, M&A, and overall relationships, but negatively related for debt relationships. The coefficient signs for investment bank market share, for analyst All-Star ranking, seasoning, and number of firms followed, and for the firm's analyst following are generally consistent with results reported in Ljungqvist et al. (2007), but the negative coefficient on investment bank size differs from their results.[14] Consistent with expectations, the coefficient on the post-Global Settlement dummy variable indicates that relative recommendations dropped in the post period. As in Table 1, there is also evidence that non-sanctioned banks tend to have higher recommendations than sanctioned banks, especially in the post-Global Settlement period.

Turning to the results for investment banking relationships, we find strong evidence that both sanctioned and non-sanctioned banks exhibited significant affiliation bias in the pre-Global Settlement period. This result holds for each type-specific relationship (equity, debt, and M&A), as well as for the overall relationship. However, the post-GS interaction terms point to significant differences between sanctioned and non-sanctioned banks in the period following the Global Settlement. For sanctioned banks, the interaction terms suggest that analyst affiliation bias is significantly reduced in the post-Global Settlement period. In particular, the combined post-Global Settlement effects listed at the bottom of the table show that analyst affiliation bias is insignificant in the post period for equity relationships, and marginally significant for debt and M&A relationships. The results for overall relationships point to

[14] In our analysis of the sub-period from 1998-2001 (Table 3 Panel A), we obtain a positive and significant coefficient on investment bank size, consistent with Ljungqvist et al.'s (2007) results for the 1994-2000 sample period.

statistically significant affiliation bias for sanctioned banks in the period after the Global Settlement, but the magnitude of the effect is substantially reduced from the pre period. Based on the coefficients on the overall relationship variable (0.160) and the post-GS interaction term (-0.129), affiliation bias is reduced by approximately 81% in the post Global Settlement period for sanctioned banks.

The results for non-sanctioned banks provide a sharp contrast. For these investment banks, analyst affiliation bias is not reduced significantly in the period following the Global Settlement. The results provide strong evidence of a continued analyst affiliation bias in the period following the Global Settlement for non-sanctioned banks, regardless of whether relationships are measured based on equity, debt, or M&A transactions, or across all combined transactions. Based on the coefficients on the overall relationship variable (0.171) and the post-GS interaction term (-0.010), affiliation bias is reduced by only 5.9% in the post Global Settlement period for non-sanctioned banks and this reduction is statistically insignificant.

To better understand the effects of analyst affiliation bias in the periods before and after the Global Settlement, we estimate models using two sub-periods: 1998-2001 and 2003-2009. The results are presented in Panels A and B of Table 3, respectively. As in Table 2, the results for the first sub-period point to significant analyst affiliation bias for both sanctioned and non-sanctioned banks. For sanctioned banks, the coefficient on *IBRel* is positive and significant for all type-specific and overall relationships. For non-sanctioned banks, the coefficient is positive and insignificant for equity and debt relationships, positive and marginally significant for M&A, and significantly positive for the overall relationship measure. . Equality of coefficients between sanctioned and non-sanctioned banks cannot be rejected for any of the relationships measures in the pre-settlement sub-period.

The results for the second sub-period (Panel B) confirm the findings from Table 2. For sanctioned banks, the coefficient on *IBRel* is positive but insignificant for equity relationships, positive and marginally significant for debt and M&A, and significantly positive for overall relationships. However, as in Table 2, the impact of investment banking relationships on relative recommendations is substantially