Measuring luck in CEO outperformance[1]

Seoyoung Kim[2]

January 2015

Abstract

Firm performance is a crucial factor in how CEOs are evaluated. However, a CEO can be repeatedly lucky or unlucky, adding noise to performance outcomes as a measure of managerial ability. In this study, I examine how much of the observed cross-sectional dispersion in outcomes can be attributed to differences in luck rather than differences in skill. Using bootstrap simulations, I find that the best performing CEOs perform too well relative to the median to be explained by luck alone. However, the true underlying differences in skill are substantially smaller than suggested by simply looking at the raw performance differential.

JEL Classification: G3, G30, M40, M49
Keywords: luck, randomness, sustained performance, performance evaluation

[1]I thank Sanjiv Das, David Denis, Diane Denis, Mara Faccio, Byoung-Hyoun Hwang, Di Li, John McConnell, Max Moroz, Jin Xu, Deniz Yavuz, and Xiaoyan Zhang for helpful comments.

[2]Leavey School of Business, Santa Clara University; 500 El Camino Real; Santa Clara, CA 95053. Email: srkim@scu.edu.

Firm performance is a crucial factor in how CEOs are evaluated. Strong firm performance is associated with higher compensation (Murphy, 1985) and subsequent CEO appointments at other firms (Fee and Hadlock, 2003) as well as with subsequent board appointments (Brickley, Linck, and Coles, 1999). Likewise, poor performance is associated with a greater likelihood of CEO turnover (Warner, Watts, and Wruck, 1988; Weisbach, 1988; Kaplan and Minton, 2008; Jenter and Kanaan, 2008) and even litigation (Donelson, McInnis, Mergenthaler, and Yu, 2010).

Moreover, anecdotal evidence suggests that greater consideration is given to outcome than process, with extreme performances attracting the most attention. News articles praise or denounce CEOs for their firms' stock-return performances,[1] and Business Week annually publishes the best and worst chief executives based on profitability and changes in shareholder value. Similarly, Forbes ranks the best and worst CEO performances, declaring that "some [executives] are so bad they should be paying the shareholders" (Forbes, 2005).

However, a CEO can be lucky or unlucky (and repeatedly so), adding noise to performance outcomes as a measure of managerial ability and soundness of judgment. That is, even accounting for systematic economic events over which the CEO has no control, the remaining firm-specific "skill" portion of his performance may not reflect skill after all. Chance events may accumulate to produce sustained profitability in a firm (Denrell, 2004), and extreme success may actually reflect poor

[1] "Peter Cartwright of Calpine, a firm that develops and runs gas-fired power plants. This is not a profitable venture at the moment, and the average annual return to shareholders over the past six years has been -7%. For this unelectrifying performance Cartwright has pocketed an average annual $13 million." (Forbes, 2005)

skill and judgment (Denrell, 2005; Denrell and Fang, 2010; Denrell and Liu, 2012). In addition to the false positives, luck can also produce false negatives. For instance, Khanna and Poulsen (1995) argue that the actions taken by CEOs of firms filing for Chapter 11 bankruptcy are similar to those taken by CEOs of otherwise comparable, matched firms that did not experience such financial distress

More generally, the large cross-section of CEOs guarantees the occurrence of extreme differences in outcomes, even if everyone were equally skilled and put forth the same amount of effort. Thus, a natural question arises as to how much of the actual observed cross-sectional dispersion in outcomes can be attributed to differences in luck as opposed to differences in skill. That is, how much variation in performances can we expect in an economy where all CEOs are equally skilled, and in comparison, what does the actual empirical distribution of outcomes tell us about the true underlying differences in ability across CEOs?

To explore this question, I simulate a time-series of stock returns for each CEO-firm pair under the assumption that all CEOs are equally capable. That is, I impose the constraint that the true α is the same across all CEO-firm pairs. Thus, any performance differential I observe in a given simulation (i.e., any cross-sectional differences in simulated sample alphas) is entirely due to differences in luck rather than differences in skill, providing a benchmark of the extent to which CEOs are expected to substantially over- or underperform by chance alone.

I use stock returns, as opposed to alternative performance metrics, because stock returns provide more frequent data points, which is necessary for reliable bootstrapping. Moreover, insofar as the market can only estimate, but not fully

know, the value added or destroyed by an incoming CEO, stock returns provide timely, ongoing assessments of the CEO's accomplishments or failures, the average of which provides a metric of the total value created or destroyed by the CEO. Thus, changes in shareholder value reflect the extent to which the CEO has exceeded or failed ex-ante performance expectations.

Empirically (i.e., based on the actual observed data), the difference in benchmark-adjusted performances between the 90th percentile and median CEOs was *1.56% per month*,[2] or 18.72% annually. In comparison, results from bootstrap simulations indicate that by sampling variation alone, we would, on average, expect to observe a performance differential of *1.46% per month*, suggesting an annualized performance differential of 17.52% that is entirely attributable to luck. That is, although the best-performing CEOs perform too well relative to the median to be completely explained by luck alone, the simulations operating under the (extreme) assumption that all CEOs are equally skilled approximate the observed actual data quite well.

A similar observation applies when I extend my analysis to poor performance outcomes. I observe that, in actuality, the benchmark-adjusted performance differential between the median and 10th percentile CEOs was *1.18% per month*. Simulation results indicate that, by luck alone, we would expect the 10th percentile CEO to underperform the median by *1.45% per month*; in fact, in all 999 simulated

[2] Specifically, these benchmark-adjusted performances are the alphas obtained from regressing excess monthly returns on excess market and excess industry returns.

cross-sections, the difference between the 10th percentile and median performers was at least as extreme as the actual, observed difference.

Together, these findings demonstrate that the wide dispersion in observed poor performance outcomes is not statistically distinguishable from a world where all CEOs are equally skilled. Moreover, these results highlight the importance of accounting for the role of luck borne by large cross-sections when interpreting performance outcomes, particularly in the context of CEO replacement decisions, which is arguably one of the most important (and potentially costliest) decisions made by firms.

In my last set of analyses, I examine whether the distinction between luck and skill is more pronounced in some industries than in others. To measure the extent to which the empirical cross-section of performances is distinguishable from the simulated equal-skill cross-section, I use the two-sample Kolmogorov-Smirnov test, which tests whether the two samples are likely to have been drawn from the same distribution. Based on these test statistics, I observe substantial cross-industry variation in the extent to which the empirical distribution of performances differs from the simulated distribution. However, the evidence suggests that this distinction does not materially impact how performance outcomes are interpreted. Specifically, I find that CEO turnover is only modestly more sensitive to performance in industries where the difference between the empirical and simulated distributions is more pronounced, indicating that the labor market does not prioritize these cross-industry differences when evaluating CEO performances.

Overall, the results show that we can expect substantial differences in realized performance outcomes even if all CEOs in a given sample are equally skilled, suggesting that the true underlying differences in CEO skill are substantially smaller than implied by simply looking at the raw difference in performance outcomes. Thus, this study highlights the importance of considering not only the extent to which CEOs outperform their benchmarks and peers but also the extent to which they are expected to do so by sampling variation alone, since even accounting for industry- and market-wide movements, the remaining "skill" portion may still reflect a substantial amount of luck. Consequently, the measure of luck introduced here is distinct from and complementary to that considered by studies exploring whether executives are rewarded for system-wide economic events (Bertrand and Mullainathan, 2001; Garvey and Milbourn, 2006), and thereby adds to the vast literature on relative performance evaluation (e.g., Holmstrom, 1982; Gibbons and Murphy, 1990; Murphy, 1999; Jenter and Kanaan, 2008).

This paper also speaks to how we interpret extreme changes in performance surrounding executive turnovers, since a substantial number of CEO-turnover events are expected to result in significant performance changes (even in an economy where the departing CEOs are truly no better or worse than their replacements). Thus, this paper also complements the studies exploring whether industry- or firm-specific factors dominate performance outcomes, and whether corporate strategy matters (e.g., Brush, Bromiley, and Hendrickx, 1999; Bowman and Helfat, 2001; Ruefli and Wiggins, 2003). Similarly, this paper pertains to studies examining the effect of firm performance on CEO labor-market outcomes (e.g., Coughlan and Schmidt,

1985; Warner, Watts, and Wruck, 1988; Weisbach, 1988; Fee and Hadlock, 2003; Kaplan and Minton, 2008; Jenter and Kanaan, 2008), and more broadly, to any study pertaining to the theory and use of performance measurement (Bourne, 2013; Micheli and Mari, 2013).

This paper is organized as follows. In Section 1, I describe the data and provide summary statistics. In Section 2, I describe the bootstrap method that I use to evaluate CEO performances. In Section 3, I present my results, and in Section 4, I conclude and discuss.

1. Data

1.1. Sources

My sample period spans 1992 to 2009 and consists of the S&P 1500 CEO-firm pairs, which I obtain from the Standard & Poors Executive Compensation database.[3] Although the Execucomp database begins in 1992, I consider a CEO's entire tenure when evaluating his performance, even if his term begins before 1992. For example, if a CEO of firm X began his term in March of 1985, I use the returns from April of 1985 to the penultimate month of the CEO's term to calculate his alpha. I obtain monthly returns data from the CRSP database, and I require that each CEO-firm pair have at least 24 months of returns data. I exclude those CEO-firm pairs whose stock prices dip below $5.00 during that CEO's term to avoid estimation problems that

[3] Prior to 1994, the Execucomp database does not provide compensation data for the entire S&P 1500.

accompany small, illiquid stocks (such as the bid-ask bounce).[4] My final sample consists of 21,365 firm-years, with 3,320 distinct CEO-firm pairs. Of these, I have 2,127 turnover events for which I have at least 24 months of returns data preceding and following the new CEO's arrival.

1.2. Summary Statistics

In Table 1, I present summary statistics on basic CEO and firm characteristics. The average CEO in my sample is 55.4 years of age, and 28% of CEOs are 60 years of age or older. The average tenure is 7.2 years, with a CEO-turnover event occurring in 11.7% of firm-years, and the average firm has $14.0 billion in total assets, with an average market capitalization of $12.4 billion.

2. Performance-Evaluation Method

To account for general market- and industry-wide price movements, I estimate the following regression for each CEO-firm pair,

$$r_{i,t} = \alpha_i + \beta_{i,MKT} \cdot MKTRF_t + \beta_{i,IND} \cdot INDRF_{i,t} + \varepsilon_{i,t}, \quad (1)$$

where $r_{i,t}$ is the monthly excess stock return for CEO-firm pair *i* at time *t*, $MKTRF_t$ is the monthly excess market return at time *t*, and $INDRF_{i,t}$ is the monthly excess industry portfolio return for firm *i*'s industry at time *t* based on the Fama-French (1997) 30-industry classification.[5]

---

[4] Other studies using this $5.00 filter include Jegadeesh and Titman (2001) and Chan, Chan, Jegadeesh, and Lakonishok (2006).

[5] Obtained from Ken French's website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. I observe very

To be clear, skill is not a uniquely defined concept, and in removing the market and industry component of returns, I may remove aspects of CEO skill. For instance, part of a CEO's job is to determine the firm's optimal exposure to external factors (Gopalan, Milbourn, and Song, 2010). Furthermore, CEOs can collectively influence the aggregate performance measures used as benchmarks (Aggarwal and Samwick, 1999). Nonetheless, my focus in this paper is to examine how much of the observed differences in actual performance outcomes can be attributed to differences in luck versus difference in skill. To this end, I study the best and worst (out)performances, irrespective of the overall level of skill in the economy and irrespective of whether CEOs have market-timing skill.

From regression equation (1), I obtain OLS parameter estimates and a time series of residuals, $\hat{\varepsilon}_{i,1},\ldots,\hat{\varepsilon}_{i,T_i}$. I then sample with replacement from this residual vector, assigning equal probability to each $\hat{\varepsilon}_{i,t}$, and I construct a simulated time-series of residuals, $\varepsilon^*_{i,1},\ldots,\varepsilon^*_{i,T_i}$. Although the sample mean of the residuals from equation (1) is zero by construction, the sample mean of the bootstrap error terms need not be, since some of the $\hat{\varepsilon}_{i,t}$'s may be selected multiple times and others may not be selected at all.[6] My bootstrap data-generating process is then:

$$r^*_{i,t} = \alpha_{NULL} + \hat{\beta}_{i,MKT} \cdot MKTRF_t + \hat{\beta}_{i,IND} \cdot INDRF_t + \varepsilon^*_{i,t}, \qquad \varepsilon^*_{i,t} \sim EDF\left(\hat{\varepsilon}_i\right) \qquad (2)$$

---

similar results using the 49-industry classification. Similar observations apply when I use a three-factor model.

[6] For example, if the residual vector from equation (1) is [-1, 0, 1]', then different iterations could yield bootstrap error-term vectors of [-1, 0, 1]', [-1, -1, -1]', [1, 0, 0]', and so on.

in which $\hat{\beta}_{i,MKT}$ and $\hat{\beta}_{i,IND}$ are the OLS parameter estimates obtained from regression equation (1), and $EDF\left(\hat{\varepsilon}_{i,t}\right)$ is the empirical distribution function that assigns equal probability, T_i^{-1}, to each $\hat{\varepsilon}_{i,t}$ in the residual vector $\hat{\varepsilon}_i$.[7] Using this nonparametric bootstrap procedure, I simulate a time-series of returns for CEO *i* on whom I impose the constraint that his true skill is equal to some pre-specified magnitude (i.e., true $\alpha_i = \alpha_{NULL}$). Nonetheless, a bootstrap sample may yield an estimated alpha that is substantially different from α_{NULL} since, by chance, it may have drawn more of the positive (or more of the negative) $\hat{\varepsilon}_{i,t}$'s.

Finally, to evaluate CEO performances accounting for the large cross-section to which they belong, I employ a cross-sectional bootstrap procedure.[8] Specifically, I follow the above process for all CEO-firm pairs, *i* = 1, ..., *N*, and I repeat this process *B*=999 times to form *B* cross-sections of *N* bootstrapped alphas and *t*-statistics.[9] Because the bootstrap data-generating process in equation (2) imposes the constraint that all CEO-firm pairs are equally skilled (i.e., true $\alpha_i = \alpha_{NULL}$ for all *i*), any cross-sectional variation in simulated sample alphas is entirely due to differences in luck as opposed to differences in skill, providing a benchmark of the extent to which

---

[7] In my implementation, I scale the residual vector by a factor of $[T_i / (T_i - K)]^{½}$, because the empirical distribution of the residuals from equation (1) has variance $T_i^{-1}\Sigma\,\hat{\varepsilon}_{i,t}^2 = T_i^{-1}(T_i - K)\hat{\sigma}^2$, in which *K* equals the number of regressors and $\hat{\sigma}^2$ is the unbiased estimator of σ_ε^2.

[8] Other studies using this procedure include Kosowski, Timmerman, Wermers, and White (2006) and Kosowski, Naik, and Teo (2007).

[9] *B* is chosen such that $\lambda \cdot (B+1)$, in which λ is the size of the test, is an integer. Otherwise, the probability of Type I error cannot be exactly λ (though this becomes increasingly irrelevant for large *B*).

CEOs are expected to achieve highly positive or negative benchmark-adjusted performances purely by chance.

3. Results

3.1. Differences in CEO performances

I begin by examining the extent to which the actual best and worst benchmark-adjusted performance outcomes differ from the average, and I compare this difference to the performance differential that we can expect in a simulated economy where all CEOs are equally skilled. That is, in each of the B=999 simulated cross-sections of individual CEO performances, I see the differences in performance outcomes between the top and average performers, the average of which provides the expected performance differential due solely to differences in luck as opposed to differences in skill. I then calculate bootstrapped p-values as the proportion of bootstrap iterations yielding a performance differential that is at least as extreme as the actual difference observed in the actual data.

The results, presented in Table 2, show that, empirically, the difference in benchmark-adjusted performances between the 90th percentile and median CEOs was 1.56% per month, or 18.72% annually. In comparison, results from bootstrap simulations indicate that, purely by sampling variation, we would expect the 90th percentile CEO to outperform the median by 1.46% per month, indicating an annualized performance differential of 17.52% that is entirely due to greater luck and a more modest 1.20% that is attributable to greater skill; based on the average

sample market capitalization of $12 billion, this 1.20% skill-related difference translates to an average dollar amount of $144 million that is due to greater skill.

The bootstrapped *p*-value of 0.01 signifies that the simulated performance differential between the 90th percentile and median CEOs exceeds the actual observed difference of 1.56% in only 1% of all bootstrapped cross-sections. Thus, while the performance differential observed in actuality exceeds the simulated differences in performance (under the assumption that all CEOs are equally skilled), the results suggest that the performance differential due to differences in skill is not as large as implied by the raw difference in outperformance between the top and median outcomes.

The difference between the top 10% of performers and the mean cross-sectional performance tells a similar story: in actuality, the top 10% of performance outcomes exceeded the mean by 2.68% per month; by sampling variation alone, this difference in benchmark-adjusted performances is expected to be 2.54% per month, suggesting an annualized performance differential of 1.68% (or, 0.14% per month) that is attributable to greater skill (bootstrapped *p*-value = 0.04). Similarly, the 99th percentile performers exhibit not only greater luck but also greater skill relative to the 90th percentile performers, with an annualized performance differential of 4.56% (or, 0.38% per month) that is attributable to greater skill (bootstrapped *p*-value = 0.05). On the other hand, although the actual best performer outperformed the 99th percentile CEO by 6.44% per month, the bootstrap simulations indicate that even if all CEOs are equally skilled, the best performer is expected to outperform the 99th percentile CEO by 7.24% per month (and in 56% of

simulated cross-sections, the performance differential was at least as large as the 6.44% observed in actuality).

Similar observations apply when I extend my analysis to poor performance outcomes (Panel B). I observe that in actuality, the benchmark-adjusted performance differential between the median and 10th percentile CEOs was 1.18% per month. Simulation results indicate that, by poor luck alone, we would expect the 10th percentile CEO to underperform the median by 1.45% per month, and in each of the 999 bootstrapped cross-sections, the performance differential between the 10th percentile and median CEOs was at least as extreme as the difference observed in actuality (bootstrapped *p*-value = 1.00).

The results suggest that the best CEOs in actuality perform too well (relative to the average performer) to be completely explained by greater luck. However, a substantial portion of the observed differences in performance outcomes are attributable to differences in luck, indicating that the true underlying differences in skill are substantially smaller than suggested by simply looking at the raw difference in benchmark-adjusted performance outcomes. Overall, the results highlight the importance of gauging the extent to which CEOs are expected to outperform benchmarks and peers by sampling variation alone; this particularly applies to large samples, which guarantee the occurrence of seemingly abnormal performance outcomes even if all CEOs are truly, equally skilled.

3.2. Best / Worst CEO Performances, and the Frequency of Extreme Outcomes

To further illustrate the importance of accounting for large cross-sections when evaluating the role of luck in CEO performances, I consider the highest performing CEO in my sample of S&P 1500 firms, who enjoyed a market- and industry-adjusted performance of 11.4% per month, with a standard error of 4.1%. Evaluated in isolation, the probability of a CEO performing so well by luck alone is very low, since a *t*-statistic of 2.78 indicates a 0.5% chance that a true zero-alpha performer would generate the observed outcome (or better). However, across 999 bootstrapped cross-sections of CEO performances, whereby I impose the assumption that all CEOs have zero skill, 40% produced a maximum performer whose monthly outperformance was at least as high. That is, in 40% of simulations, the best performer, whose 'achievement' is solely an artifact of sampling variation, performs at least as well as the best performer observed in actuality (untabulated).

To expound this example, I examine the actual frequency of CEOs who achieve extreme performance outcomes, and I compare this to the number of simulated CEOs who are expected to experience such extreme outcomes purely by luck. The results, presented in Table 3, show that, in actuality, 132 CEO-firm pairs (out of my sample of 3,320 CEO-firm pairs) had benchmark-adjusted performances of at least 3% per month (Panel A). In comparison, bootstrap simulations indicate that under the assumption that all CEOs have zero skill (i.e., true α=0), 72 CEOs are still expected to enjoy such positive outcomes by sampling variation alone.

With regard to poor performance outcomes (Panel B), I observe that in actuality, 32 CEO-firm pairs had benchmark-adjusted performances of less than -3%

per month. Simulations indicate that, by poor luck alone, 70 zero-skill CEOs are expected to experience such negative performance outcomes. Even in a sample where all CEOs have a true skill level of α=0.57%, which is the average benchmark adjusted performance in my sample of S&P1500 CEOs, simulations indicate that 41 CEOs are expected to suffer monthly benchmark-adjusted performances of less than -3% per month, and in a sample where all CEOs have a true skill level of α=1.00%, 29 CEOs are still expected to experience such negative outcomes by poor luck alone. Similar observations apply when I extend my analysis to the frequency of extreme *t*-statistics.

Overall, the results show that in a large sample, many unskilled performers are sure to be sufficiently and repeatedly lucky (or unlucky) so as to produce outcomes that are statistically significant at conventional cutoffs. That is, a substantial number of CEOs are expected to achieve extreme performance outcomes based on extreme luck (as opposed to extreme skill), pointing to the importance of evaluating individual CEO performances within context of the large cross-section to which they belong. Consistent with the previous analyses, the simulations suggest that luck generates a substantial portion of benchmark-adjusted performance outcomes, and further highlight the importance of gauging not only the extent to which CEOs outperform industry benchmarks and peers, but also the extent to which they are expected to do so by sampling variation alone.

3.3. Cross-industry differences in luck/skill distinction

In my last set of analyses, I examine whether the distinction between luck and skill is more pronounced in some industries than in others. That is, I form simulated cross-sections as before, this time under the assumption that all CEOs within an industry are equally skilled, with each α_i set to the industry mean. Then I compare the simulated distribution of industry performances with the empirical distribution, which allows me to assess how much of the observed differences in performance outcomes, within a given industry, are due to differences in luck rather than differences in skill.

To measure the extent to which the empirical cross-section of performances is distinguishable from the simulated equal-skill cross-section, I use the two-sample Kolmogorov-Smirnov test. The resulting test statistic quantifies the extent to which the two samples differ, with higher values signifying a lower likelihood of falsely rejecting the null hypothesis that the two samples are drawn from the same distribution.

In Table 4, I present the two-sample Kolmogorov-Smirnov test statistics by industry, using the Fama-French 12-industry classification.[10] Overall, I observe substantial cross-industry variation in the extent to which the empirical distribution of performances differs from the simulated equal-skill distribution: the *Business Equipment* industry (category 6) has the lowest KS statistic, with a value of 0.100, and the *Utilities* industry (category 8) has the highest, with a value of 0.649. A natural

[10] I use the 12-industry classification because finer industry classifications result in sparser partitions, which is problematic to forming reliable cross-sectional distributions.

question that arises from these observations is whether such cross-industry differences affect how performance outcomes are interpreted, particularly in the context of CEO replacement decisions, which is arguably one of the most important decisions made by firms.

To examine turnover-performance sensitivity, I estimate the following binary response model using the logistic function:

$$CEOTurnover_{i,t} = \alpha + \beta_1 \cdot RETRF_{i,t-1} + \beta_2 \cdot RETRF_{i,t-1} * P_{KS_i} + \beta_3 \cdot P_{KS_i} + X_{i,t}\beta_{4-5} + \varepsilon_{i,t}, \qquad (5)$$

where $CEOTurnover_{i,t}$, the dependent variable, is an indicator that equals one if a CEO turnover occurs at firm *i* in year *t*, and zero otherwise; $RETRF_{i,t-1}$ is the annual excess stock return for firm *i* in year *t-1*; and P_KS_i is the *p*-value from the two-sample Kolmogorov-Smirnov test comparing the empirical and bootstrapped distributions of firm *i*'s industry (Column 2 uses the *p*-value of the KS statistic, and Column 3 uses an indicator that equals one if the *p*-value of the KS statistic is greater than or equal to 0.10, and zero otherwise).

$X_{i,t}$ is a vector consisting of year dummies, as well as a CEO-age indicator that equals one if the CEO is at least 60 years of age, and zero otherwise, which serves to account for the natural succession process. The 60-year cutoff for the age indicator was chosen following Parrino (1997) and Huson, Parrino, and Starks (2001), among others, and represents the age at which the CEO title is transferred (given a typical retirement age of 65). Reported *p*-values account for clustering by firm, which adjusts for serial correlation (Petersen, 2009).

The results, presented in Table 5, show a modest decrease in CEO turnover-performance sensitivity conditioned on KS_i. That is, CEOs are less likely to be replaced following poor stock-return performance in industries where the distinction between the empirical and simulated distributions is not as pronounced (coefficient estimate = 1.587, *p*-value=0.05). For a decrease in returns from 20% to -20%, the probability of turnover increases by 1.2% less in industries where luck is more difficult to distinguish from skill.[11] Overall, while the reason for departure is unclear, the labor market does not appear to place much weight these cross-industry differences when evaluating CEO performances.

4. Conclusion

In this paper, I provide evidence that the true underlying differences in CEO skill are substantially smaller than suggested by the dispersion in actual performance outcomes. Firm performance is a crucial factor in how CEOs are evaluated, with extreme performances attracting the most attention. However, as the simulation results demonstrate, a few random CEOs from a large cross-section are certain to be repeatedly lucky or unlucky, guaranteeing extreme differences in performance outcomes even if everyone is equally skilled and puts forth the same amount of effort. This paper also speaks to how we interpret extreme changes in performance surrounding executive turnovers, since a substantial number of CEO-turnover events

[11] To be precise, this difference-in-difference is derived by comparing the probability differential when the p-value of the industry KS statistic is 0.00 against the probability differential when the p-value is 0.20

are expected to result in significant performance changes (even in an economy where the departing CEOs are truly no better or worse than their replacements).

Evaluating how much variation in performances we can expect due solely to differences in luck, thus, has important implications for replacement decisions, for designing incentive contracts, and more broadly, for how CEO performance outcomes are interpreted in attempting to measure managerial effort or ability.

References

Aggarwal, R.K., and A.A. Samwick, 1999. Executive compensation, strategic competition, and relative performance evaluation: Theory and evidence. *Journal of Finance* 54, 1999–2043.

Bertrand, M. and S. Mullainathan, 2001. Are CEOs rewarded for luck? The ones without principals are. *Quarterly Journal of Economics* 116:3, 901–932.

Bowman, E.H., and C.E. Helfat, 2001. Does corporate strategy matter. *Strategic Management Journal* 22, 1–23.

Bourne, M., 2013. Emerging issues in performance management. *Management Accounting Research* 25, 117–118.

Brickley, J.A., J.S. Linck, and J.L. Coles, 1999. What happens to CEOs after they retire? New evidence on career concerns, horizon problems, and CEO incentives. *Journal of Financial Economics* 52, 341–377.

Brush, T.H., P. Bromiley, and M. Hendrickx, 1999. The relative influence of industry and corporation on business unit performance: An alternative estimate. *Strategic Management Journal* 20, 519–547.

Chan, K., L.K.C. Chan, N. Jegadeesh, and J. Lakonishok, 2006. Earnings quality and stock returns. *Journal of Business* 79, 1041–1082.

Coughlan, A.T. and R.M. Schmidt, 1985. Executive compensation, management turnover, and firm performance: An empirical investigation. *Journal of Accounting and Economics* 7, 43–66.

Denrell, J., 2004. Random walks and sustained competitive advantage. *Management Science* 50, 922–934.

Denrell, J., 2005. Should we be impressed with high performance. *Journal of Management Inquiry* 14, 292–298.

Denrell, J., and C. Fang, 2010. Predicting the next big thing: Success as a signal of poor judgment. *Management Science* 56, 1653–1667.

Denrell, J., and C. Liu, 2012. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences* 109, 9331–9336.

Donelson, D.C., J. McInnis, R.D. Mergenthaler, and Y. Yu, 2010. The timeliness of earnings news and litigation risk. *Working paper series*.

Fama, E. and K. French, 1997. Industry costs of equity. *Journal of Financial Economics* 43, 153–193.

Fee, C.E. and C.J. Hadlock, 2003. Raids, rewards, and reputations in the market for managerial talent. *Review of Financial Studies* 16, 1315–1357.

Forbes, 2005. The Best and Worst Bosses.

Garvey, G. and T. Milbourn, 2006. Assymetric benchmarking in compensation: executives are rewarded for good luck but not penalized for bad. *Journal of Financial Economics* 82, 197–226.

Gibbons, R. and K.J. Murphy, 1990. Relative performance evaluation for chief executive officers. *Industrial and Labor Relations Review* 43:3, 30S-51S

Gopalan, R., T. Milbourn, and F. Song, 2010. Strategic flexibility and the optimality of pay for sector performance. *Review of Financial Studies* 23, 2060–2098.

Holmstrom, B., 1982. Moral hazard in teams. *Bell Journal of Economics* 13:2, 324–340.

Huson, M.R., R. Parrino, and L.T. Starks, 2001. Internal monitoring mechanisms and CEO turnover: A long-term perspective. *Journal of Finance* 56, 2265–2297.

Jegadeesh, N. and S. Titman, 2001. Profitability of momentum strategies: An evaluation of alternative explanations. *Journal of Finance* 56, 699–720.

Jenter, D. and F. Kanaan, 2008. CEO turnover and relative performance evaluation. *Working paper series*.

Jones, M.C., J.D. Marron, and S.J. Sheather, 1996. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91, 401–407.

Kaplan, S.N. and B.A. Minton, 2008. How has CEO turnover changed? *Working paper series*.

Khanna, N. and A.B. Poulsen, 1995. Managers of financially distressed firms: Villains or scapegoats? *Journal of Finance* 50, 919–940.

Kosowski, R., A. Timmermann, R. Wermers, and H. White, 2006. Can mutual fund 'stars' really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance* 61, 2551–2595.

Kosowski, R., N.Y. Naik, and M. Teo, 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics* 84, 229–264.

Micheli, P., 2013. The theory and practice of performance measurement. *Management Accounting Research* 25, 147-156.

Murphy K.J., 1985. Corporate performance and managerial remuneration: An empirical analysis. *Journal of Accounting and Economics* 7, 11-42

Murphy K.J., 1999. Executive compensation. North Holland, Amsterdam.

Parrino, R., 1997. CEO turnover and outside succession: A cross-sectional analysis. *Journal of Financial Economics* 46, 165-197.

Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Finance Studies* 22, 435-480.

Ruefli, T.W., and R.R. Wiggins, 2003. Industry, corporate, and segment effects and business performance: A non-parametric approach. *Strategic Management Journal* 24, 861–879.

Warner, J.B., R.L. Watts, and K.H. Wruck, 1988. Stock prices and top management changes. *Journal of Financial Economics* 20, 461-492.

Weisbach, M.S., 1988. Outside directors and CEO turnover. *Journal of Financial Economics* 20, 431-460.

Table 1
CEO and firm characteristics

This table presents summary statistics on CEO and firm characteristics for my sample of S&P 1500 firms during the period of 1992 to 2009. Panel A presents CEO characteristics, where: *CEO Age* is the CEO's age in years; *CEO Age (≥60)* is an indicator variable that equals one for CEOs who are at least 60 years of age, and zero otherwise; and *CEO Tenure* is the incumbent CEO's tenure (as CEO of that firm) in years. Panel B presents firm characteristics, where: *CEO Turnover* is an indicator that equals one for firm-years in which a CEO turnover occurs, and zero otherwise; *Total Assets* is the book value of total assets in millions; and *MV(Equity)* is the market value of equity in millions.

Variable	Mean	(Std. Dev)
Panel A. CEO characteristics		
CEO Age	55.37	(7.25)
CEO Age (≥60)	0.28	(0.45)
CEO Tenure	7.17	(7.25)
Panel B. Firm characteristics		
CEO Turnover	0.117	(0.32)
Total Assets ($million)	13,980	(78,878)
MV(Equity) ($million)	12,037	(40,293)
No. of firm-years	21,365	---

Table 2
Differences in CEO performances

This table presents the empirical versus bootstrapped differences in performance across CEOs. For all CEO-firm pairs having at least 24 months of returns data (*N*=3,320), I estimate monthly industry/market-model alphas by regressing monthly excess returns on the excess market return and the relevant excess industry return (based on the Fama-French 30-industry classification). The 'empirical difference' is the actual observed difference. The 'bootstrapped expected difference' is the expected performance differential based on simulations under the assumption that all CEO-firm pairs are equally skilled (i.e., true alphas are equal). The bootstrapped *p*-value, presented below in brackets, reports the proportion of bootstrap iterations yielding a performance differential even more extreme than the actual observed difference.

Panel A. Actual versus bootstrapped differences among strong performers

	Top 10% (avg) minus mean	90th pctl minus median	99th pctl minus 90th pctl	Maximum minus 99th pctl
Empirical difference (%)	2.68	1.56	2.93	6.44
Bootstrapped expected difference (assuming no differences in skill)	2.54	1.46	2.55	7.24
Bootstrapped p-value	[0.04]	[0.01]	[0.05]	[0.56]

Panel B. Actual versus bootstrapped differences among poor performers

	Bottom 10% (avg) minus mean	10th pctl minus median	1st pctl minus 10th pctl	Minimum minus 1st pctl
Empirical difference (%)	-2.13	-1.18	-2.18	-2.06
Bootstrapped expected difference (assuming no differences in skill)	-2.49	-1.45	-2.43	-5.82
Bootstrapped p-value	[1.00]	[1.00]	[0.91]	[1.00]

Table 3
Frequency of extreme empirical versus bootstrapped performances

This table presents the cumulative number of CEO-firm performances above or below a certain threshold. For each CEO-firm pair having at least 24 months of returns data (*N*=3,320 pairs), I estimate the monthly alpha and corresponding *t*-statistic by regressing monthly excess returns on the excess market return and the relevant excess industry return (based on the Fama-French 30-industry classification). The row labeled 'Empirical outcome' reports the number of actual estimated alphas or *t*-statistics that exceed the reported threshold. The rows under 'Bootstrapped outcome' report the number of simulated alphas or *t*-statistics that are expected to exceed the reported threshold based on a bootstrapped distribution under the assumption that all performers have true alphas of -1.00%, -0.57%, 0.00%, 0.57%, and 1.00%, respectively. The corresponding 5th and 95th percentiles of each of these simulated distributions are reported below in brackets.

Panel A. Frequency of strong performances

	Frequency of alphas ≥ ...					Frequency of *t*-statistics ≥ ...		
	1.0%	1.5%	2.0%	2.5%	3.0%	1.64	1.96	2.57
Empirical outcome	946	516	329	210	132	316	153	36
Bootstrapped outcome:								
assuming true α = -1.00%	185 [171, 200]	114 [102, 127]	73 [64, 83]	48 [41, 55]	32 [28, 38]	22 [19, 24]	10 [7, 11]	2 [1, 2]
assuming true α = -0.57%	288 [268, 309]	172 [158, 186]	107 [96, 119]	69 [59, 78]	45 [38, 51]	49 [40, 57]	22 [17, 27]	4 [3, 5]
assuming true α = 0.00%	541 [511, 569]	309 [287, 330]	183 [167, 199]	113 [101, 125]	72 [62, 82]	165 [144, 182]	82 [69, 95]	18 [12, 24]

Table 3 continued.

Panel B. Frequency of poor performances

	Frequency of alphas ≤ ...					Frequency of *t*-statistics ≤ ...		
	-3.0%	-2.5%	-2.0%	-1.5%	-1.0%	-2.57	-1.96	-1.64
Empirical outcome	32	51	85	123	231	4	28	46
Bootstrapped outcome								
assuming true α = 0.00%	70 [56, 83]	115 [99, 133]	194 [171, 215]	337 [309, 364]	554 [522, 587]	28 [20, 37]	108 [91, 125]	201 [178, 224]
assuming true α = 0.57%	41 [32, 52]	63 [50, 75]	102 [87, 119]	170 [149, 189]	291 [266, 318]	7 [3, 12]	31 [23, 41]	63 [51, 77]
assuming true α = 1.00%	29 [21, 37]	45 [36, 56]	73 [60, 87]	119 [103, 137]	185 [163, 204]	3 [1, 6]	14 [9, 21]	30 [21, 40]

Table 4

Measuring differences in empirical versus bootstrapped distributions across industries

This table presents Kolmogorov-Smirnov statistics from two-sample KS tests comparing the empirical and zero-skill bootstrapped cross-sectional distributions of CEO-firm performances by industry, with higher values signifying a lower likelihood of falsely rejecting the null hypothesis that the two samples are drawn from the same distribution. Industry categories are based on the Fama-French 12-industry classification.

	Two-sample KS statistic
Category 1: Consumer nondurables (*N* = 221) - food, tobacco, textiles, apparel, leather, toys	0.442
Category 2: Consumer durables (*N* = 78) - cars, tv's, furniture, household appliances	0.283
Category 3: Manufacturing (*N* = 437) - machinery, trucks, planes, office furniture, paper	0.355
Category 4: Energy (*N* = 134) - oil, gas, coal extraction	0.466
Category 5: Chemicals (*N* = 126) - chemicals and allied products	0.506
Category 6: Business equipment (*N* = 484) - computers, software, electronic equipment	0.100
Category 7: Telecommunications (*N* = 66) - telephone and television transmission	0.402
Category 8: Utilities (*N* = 224) - electric, gas, water supply	0.649
Category 9: Shops (*N* = 378) - wholesale, retail, and some services (laundries, repair shops)	0.335
Category 10: Health (*N* = 247) - healthcare, medical equipment, drugs	0.179
Category 11: Finance (*N* = 572) - banks, insurance companies, and other financials	0.275
Category 12: Other (*N* = 353) - mines, construction, building maintenance, transportation, hotels, business services, entertainment	0.159

Table 5
Turnover-performance sensitivity

This table presents estimates from a pooled logit model of $CEOTurnover_{i,t}$ on past returns during the period of 1993 to 2009. $CEOTurnover_{i,t}$, the dependent variable, is an indicator that equals one if a CEO turnover occurs at firm *i* in year *t*, and zero otherwise; $RETRF_{i,t-1}$ is the annual stock return (in excess of the risk-free rate) for firm *i* in year *t-1*; P_KS_i is the *p*-value from the two-sample Kolmogorov-Smirnov test comparing the empirical and bootstrapped distributions of firm *i*'s industry, based on the Fama-French 12-industry classification (Column 2 uses the *p*-value of the KS statistic, and Column 3 uses an indicator that equals one if the *p*-value of the KS statistic is greater than or equal to 0.10, and zero otherwise). Also included are year dummies as well as a CEO-age indicator that equals one if the CEO is at least 60 years of age, and zero otherwise. Two-tailed *p*-values are reported in brackets below and account for clustering by firm.

	Coefficient estimate [*p*-value]		
		Interacted with...	
	unconditional	... *p*-value	... indicator
	(1)	(2)	(3)
$RETRF_{i,t-1}$	-0.521	-0.619	-0.618
	[0.00]	[0.00]	[0.00]
$RETRF_{i,t-1} * P\_KS_i$		1.587	0.263
		[0.05]	[0.05]
P_KS_i		0.938	0.156
		[0.01]	[0.01]
No. of firm-years	21,365	21,365	21,365
Likelihood ratio χ^2	475.03	490.96	491.10