Deloitte.

A Random Search for Excellence

Why "great company" research delivers fables and not facts

Michael E. Raynor, Deloitte Consulting LLP
Mumtaz Ahmed, Deloitte Consulting LLP
Andrew D. Henderson, University of Texas at Austin

Contents

Many believe that we can learn how to be great by studying greatness. But what is great performance? It turns out that we typically measure the wrong thing and set the bar far too low. Consequently, researchers who think they are studying successful companies are usually studying the winners of a random walk. What does this mean for the soundness of some of the most popular and influential management research?

The bottom line: you can't trust it.

The genre of management research that we refer to as the "success study" has a long and illustrious pedigree, and enjoys an ever-tighter grip on our collective imagination. You know the kind of book we're talking about; it has a well-known recipe. Start with a population of companies and identify the most successful among them. Examine their behaviors and look for patterns associated with that success. Distill those patterns into a general framework. Claim that if you use that framework to guide your own behaviors you can achieve those results.

In Search of Excellence in 1982 got the ball rolling, and was such a monster hit that it sucked all the oxygen out of the room for fourteen years. *Built to Last* in 1996 proved someone other than Tom Peters could do this, and that opened the floodgates: over the last twelve years there have been at least a dozen more such efforts released. In 1999 we got *The Alchemy of Growth* followed by *Peak Performance* in 2000; in 2001 we saw *Profit from the Core, Creative Destruction,* and the still-reigning successor to *Search, Good to Great*. In 2003 we were told *What Really Works*, followed in 2006 by a *Blueprint to a Billion* and *Big Winners and Big Losers;* then *The Breakthrough Company, The Granularity of Growth, Stall Points* and *The Momentum Effect* all hit the shelves in 2008. All the while there has been a steadily increasing rainfall of reports, studies and white papers from think-tanks and consulting firms taking a similar approach to specific management practices.

Many managers have found the prescriptions in one or more of these studies helpful, perhaps even enormously so. And we're sympathetic to the notion that if it works, don't knock it. But we've come to the rather disturbing conclusion that every one of the studies that we've investigated in detail is subject to a fundamental, irremediable flaw that leaves us with no good *scientific* reason to have any confidence in their findings.

It is this: success studies typically don't substantiate the claim that they are studying unexpectedly successful companies. By our measures, they are instead, by an overwhelming majority, studying a sample of firms with performance profiles that are statistically indistinguishable from fortunate random walks. In other words, they are not studying demonstrably great companies, and may very well be studying merely lucky companies. And since there are many more lucky companies than good ones, the inputs to every success study we can lay our hands on are very likely the wrong inputs. This has material consequences for the confidence we can have in the advice offered, for no matter how rigorous the data collection, no matter how Aristotelian the logic, as the saying goes, "garbage in, garbage out".

Because these studies fail as science, managers cannot hope to achieve reliably the results they are told to expect. It's only too likely that whatever benefit practitioners have realized has been distressingly haphazard, the consequence of a form of placebo effect (you expect it to help, so you perceive that it does, quite independently of any true causal connection), a Hawthorne effect (the mere act of focusing on something you were neglecting improves performance regardless of what motivated the increased attention), or luck (even a broken clock is right twice a day).

We continue to believe that an effective way to learn about greatness is to study greatness. In the service of that belief, this monograph will, we hope, begin a dialog about a critically important – but, as far as we can tell, largely ignored – question: how good do you have to be before you can claim to be great? For only when we have a measurable degree of confidence in the greatness of any given company can we have confidence in what we learn from studying it.

Outliers

We are not the first to suggest that there are some weaknesses in popular management research. In particular, two fairly recent books have had the biggest names in the business – *especially* the biggest names in the business – in their sights. *Hard facts, Dangerous Half-truths and Total Nonsense* by Pfeffer and Sutton, and *The Halo Effect and the Eight other Delusions that Deceive Managers* by Phil Rosenzweig have each explained how specific lapses in research design or defective reasoning undermine the confidence we can have in findings based on particular types of evidence.

For example, the "halo effect" that gives Rosenzweig's book its title is a consequence of insufficient skepticism on the part of researchers when evaluating various sources of insight into firm behavior. For example, the stories that show up in newspapers and magazines – and business school case studies for that matter – are often colored by the very performance one hopes to explain. As a result, rather than uncovering what *causes* great performance, one is more likely revealing how great performance is *described*.

This and other criticisms are severe and often justified. But they have not been enough to disillusion most managers. Our collective confidence in the "success study" method is perhaps tempered by these attacks, but there's been no run on the bank as yet. After all, even though the data might sometimes be suspect or a particular argument might have smuggled in a premise or two, for most, faith in the benefits of studying great companies remains strong.

But perhaps not for much longer. The bedrock assumption of every success study has, as far as we can tell, gone unquestioned in popular management writing, to wit, that the companies being studied are in fact remarkable. If we ever doubt that the "excellent" companies were excellent or that the "great" companies were great, at least when the relevant researcher said they were, everything collapses in a heap. You can forgive every halo in paradise but if the object of our admiration is in fact nothing special, any blueprints to a billion we might see are no more significant than finding the profile of Elvis in a slice of pound cake.

Prima facie this might seem a ridiculous concern. In *Good to Great* for instance, the great companies deliver shareholder returns over a fifteen year period that outstrip the broader market indices by up to 18 times. How could that possibly be anything but remarkable?

And that's precisely the problem: success studies tend to rely exclusively on intuition to justify the noteworthiness of any particular performance profile. But our intuition is easily fooled. Rebecca Henderson at MIT illustrates the problem as follows. In her words:

> I begin my course in strategic management by asking all the students in the room to stand up. I then ask each of them to toss a coin: if the toss comes up "tails" they are to sit down, but if it comes up "heads" they are to remain standing. Since there are around 70 students in the class, after six or seven rounds there is only one student left standing. With the appropriate theatrics, I approach the student and say "HOW DID YOU DO THAT??!! SEVEN HEADS IN A ROW!! Can I interview you in *Fortune*? Is it the T shirt? Is it the flick of the wrist? Can I write a case study about you? …"

What's at work here is our propensity to confuse the long run consequences of systemic variability with individual attributes such as skill. It's easy for us to make this mistake, because in any system subject to variation in outcomes – which is every system there is – streaks of high and low performance that confound our intuition are in fact to be expected, not due to any unique attributes of a given individual, but thanks to the inherent variability of the system.

Some are put off when this sort of variation is labeled "random" because for many that connotes some sort of magical, non-rational, or inexplicable animating force. So perhaps it's useful to borrow from the world of statistical process control and think of outcomes as determined by "common causes" (attributes of the system) or "special causes" (attributes of the individual). Outcomes driven by common causes can be explained post facto, but those explanations have no predictive power. For example, when someone wins the lottery they can explain why they chose the numbers they did: their anniversary, their favorite number, the month, and so on. And so some might argue that their choices weren't "random" because they can be "explained." But whether or not those numbers were the winning numbers was a function of the system (i.e., common causes), not the decision-making process (i.e., the special causes) that led that ticket to have those numbers. In short, that explanation does us no good in predicting what numbers will win next time around.

On the other hand, if we find someone who wins the lottery five times in a row, that looks like *special cause* variation – it's an outcome that is sufficiently unlikely to warrant investigation into the causes that are unique to that outcome (or, at the limit, common to a class of similar outcomes). Wanting to know more about how five-time winners pick their numbers is entirely reasonable.

The trick is to separate out which individuals have delivered sufficiently unlikely results to warrant a special cause hypothesis, and there's only one study we're aware of that even acknowledges this as an issue: *Creative Destruction* by Foster and Kaplan. They point out that unless a firm delivers performance outside the range of the system within which it functions, it has done nothing exceptional.

This is a great start, but more is required. Landing outside the normal range of expected performance is a necessary condition of remarkable performance, but it is not sufficient. Essentially, falling outside the "normal" range is equivalent to tossing seven heads out of seven tosses – it's a very unlikely event if you only get one shot of seven tosses.

To finish the job we have to correct for the number of people tossing coins. When you're actually tossing coins, this is arithmetically straight-forward. The players in Henderson's game are all independent of each other – that is, any given player's outcomes don't depend on anyone else's – and each individual's tosses are independent of all other tosses by that individual, so the odds of getting heads this time are the same regardless of what one got the last time. So, with 70 students in the room we'd expect about one person to get seven heads in a row. There's no good reason to think the lucky winner is special in any way, despite the seemingly highly unlikely nature of someone tossing seven heads in a row, simply because of the number of people in the system tossing coins. Since we *expect* one person to toss seven heads in a row and we *observe* one person with seven, claiming special cause variation for the winner, even for this seemingly unlikely event, exposes us to a very high likelihood of a "false positive" – that is, claiming that an outcome is due to special causes when in fact it can be explained entirely by common causes, that is, by the operation of chance alone.

On the other hand, if we got far more players with seven heads in a row than were expected due to common causes alone, then we'd be justified suspecting that special causes are at work for at least some of them. For example, if we observed 15 players out of 70 with seven heads in a row there's good reason to believe somebody's on to something. But how many? And which ones?

Our expectation of one out of 70 with seven heads in a row is itself merely an estimate, and is subject to a confidence interval of almost two. So, 95% of the time, we'd expect to see between zero and three players with seven heads out of seven tosses. Should we observe 15, we'd have a very high degree of confidence that 12 of the 15 people who tossed seven heads were "true" positives. Sadly, the statistics won't tell us which ones, but if we were to study, say, five of those 15 selected at random, we could expect three or four of them to be the real thing – that is, exploiting special causes to flip heads "deliberately." Studying that group of five and looking for patterns in how they do it is likely to have a high "signal to noise ratio" and yield meaningful insights.

When it comes to assessing whether a particular company's performance profile – say, beating the market by 10-fold over a 10 year period – is a consequence of common causes or special causes, and then correcting for the possibility of false positives, we have a much more challenging problem on our hands.

Competition between companies can be seen as a system, and the distribution of companies by performance (however measured) in a given year is in part a consequence of the common causes that define that system. Some companies do better or worse not because they are fundamentally any different, but simply due to the random and unpredictable perturbations endemic to the system itself. With coin tosses we can characterize the odds of getting heads or tails. But with the competitive economy, we can't easily determine what the expected distribution of outcomes due to common causes actually is.

At the risk of repetition, we can always explain after the fact why Acme Inc. delivered, say, 15% return on assets while XYZ Corp. delivered only 12%; but if the difference is due to common causes, our explanation has no predictive power: depending on the structure of the system, next year it could be just as likely that XYZ will get the better of Acme. This is an important point, for as Malcolm Gladwell makes clear in his recent book *Outliers*, people who achieve exceptional results can typically trace their success to deeply contextual and highly idiosyncratic explanations. The problem is this is as true of Joe the Plumber as it is of Bill Gates: everyone's got a story. The difficulty is determining which stories capture common versus special causes.

In addition, each player's outcome is determined by the performance of the others because business performance is entirely relative. It's no good knowing what your chances are of delivering 15% ROA. What we care about is the likelihood that you'll do better than a given percentage of the rest of the population. In other words, we want to know what the odds of a specified relative performance will be, and that clearly depends on how well everyone else does.

Finally, whereas coin tosses are independent of each other, how a company does in year two of a time series is strongly influenced by how well it did in year one. Performance is sticky, and unless we can correct for that stickiness we're very likely to attribute exceptional performance to companies evincing nothing more than system-level inertia.

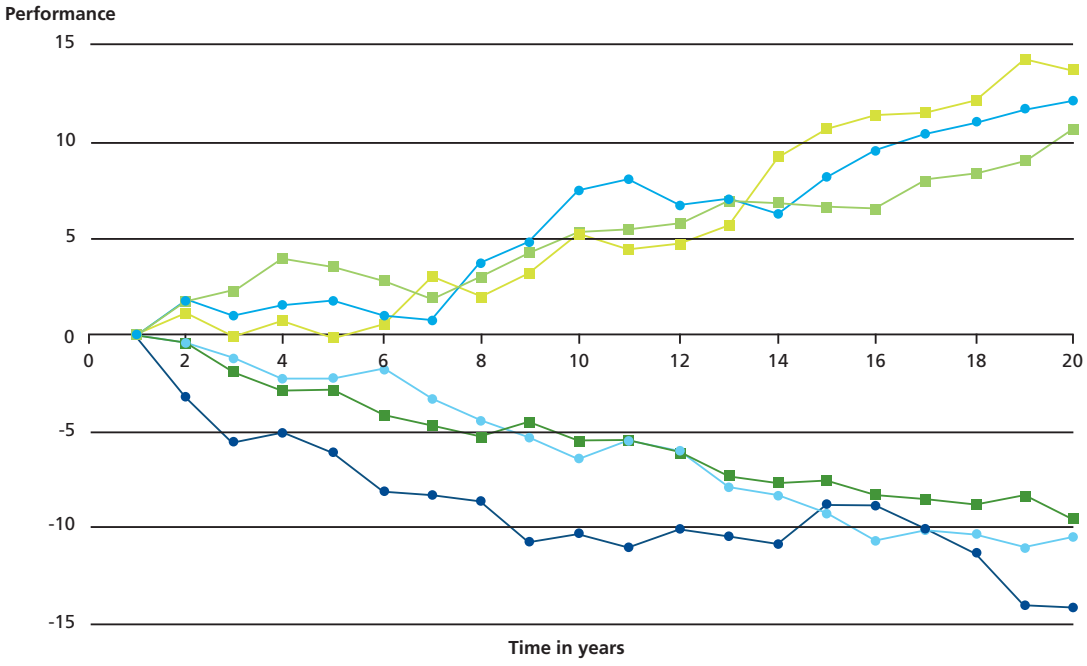
Our intuition fails us when assessing runs of coin tosses; to see how easily we can be led astray when considering a much more complex system, consider the following simulation of twenty years of performance for 100 firms. In this experiment each year's performance equals the previous year's performance (stickiness) plus a random "bump," which can be positive or negative (common causes). All 100 firms start with a performance level of 0, and the bump is normally distributed with an average of 0 and a standard deviation of 1. This simulation is known as a "random walk."

Figure 1 shows how fate treated six of the 100 firms. It's hard to resist concluding that we have a sample of three "big winners" and three "big losers." But we know in truth there's nothing to learn about the impact of individual attributes on performance by studying these six firms: we defined the simulation such that all the individuals were identical (there are no special causes), and any differences between individuals are exclusively consequences of randomness in the system – the common causes of performance.

When our spider sense is so easily fooled even when we know what is actually going on, imagine how difficult it is to differentiate between good and great when we have no idea what the underlying properties of the system in question truly are. It's only too likely that seemingly slam-dunk great performance might actually be nothing of the sort.

Making these determinations with any confidence is very difficult, but it is possible to generate some reasonable estimates of what's going on. Using a database of 41 years of corporate-level performance for approximately 85% of all companies traded on US public equity markets, we've attempted to create an algorithm for identifying the (plausibly) truly exceptional.

Figure 1. The Illusion of Greatness



Taking measure

We tackled this question, ironically enough, because we are on our own quest to uncover the secrets of sustained performance – in other words, we're doing our own "success study." And when looking for outperforming companies, the first question is always *how do you measure performance*?

A common, but hardly unanimous, practice is to look at total returns to shareholders (TSR): *Blueprint to a Billion* used sales, *Built to Last* used a composite measure of general esteem, *The Breakthrough Company* used primarily sales growth. Some studies wanted to examine a specific phenomenon (how to get to a billion in sales), while others had to cope with constraints on data availability (if you're looking at private companies, you can't use TSR).

We took the view that at first principles, the object of all of these studies, ours included, is to identify noteworthy *management* practices. This has some important implications, of which perhaps the most surprising is that TSR is arguably the worst measure to use.

Shareholder returns are a function of the capital market's estimate of future performance. A good fraction of TSR tells the story of changing hopes for the future rather than delivering on past promises. Consequently, strong returns over time are often largely the result of consistent upside surprises that serve to ratchet up expectations, which is then made manifest in a rising stock price.

This is perhaps most evident in companies that deliver great performance but lackluster TSR: a company can deliver fabulous profitability and eye-popping growth and have a share price that goes nowhere because at some point in the past the markets "figured them out" and priced that performance into the shares.

Take, for example, Wal-Mart. From 1975 to 2000 the stock chart was an upward sloping 45-degree line (on log paper, no less!) with hardly a hiccough (there was a brief plateau from 1992-1997). Every time the company delivered outstanding results, most investors simply couldn't believe that this company was going to continue to grow so quickly and so profitably for much longer. But they were wrong, and every time Wal-Mart surprised on the upside, the stock went up as investors re-set their expectations. And so Wal-Mart was often seen as a company worth studying. Since 2000, however, the stock has gone essentially sideways, more or less tracking the market (although it hasn't dropped by nearly as much as many broader indices since the Crash of '08 began). Is Wal-Mart no longer well managed? Did Wal-Mart managers stop doing all the great things that made them great for so long?

Hardly. In fact, according to our analysis, Wal-Mart has delivered excellent and exceptionally rare operating performance on a consistent basis throughout its existence. Wal-Mart appears to have gone from great to good (as measured by TSR) not because of how it has changed but because of how the market's expectations have changed: in 2000 investors as a group finally "got it" and priced Wal-Mart's consistent, profitable growth into the stock. In other words, Wal-Mart has continued to deliver outstanding performance; it just hasn't continued to deliver surprising performance. But whether investors are surprised or not is as much, or more, a function of the investors as of the company itself.

Now, if one is interested in what behaviors surprise investors, TSR might be just the ticket. But if *great management* is what you want to understand, operating measures of performance are much better. Of the operating measures available, we like return on assets as an overall measure of profitability. Other measures such as economic margin also make sense, but data availability can limit their utility.

Attempting to isolate the impact of management on firm performance has other important implications. For example, some industries routinely deliver higher levels of performance than others not because of systematically superior management but because of systematically superior structural characteristics. For example, the pharmaceutical industry has historically enjoyed very high barriers to entry and low levels of internal rivalry due to patent-protected monopolies on specific products. Consequently, pharmaceutical companies can expect higher ROA than companies in, say, the commodity chemicals industry simply because of differences between those two industries – not due to differences between the quality of the management one finds in those two industries.

Controlling for the impact of different exposure to economic cycles, company size, longevity, survivor bias, financial structure and other factors that affect ROA but don't reflect directly on strategic or operating managerial savvy each poses its own challenges, but none is insurmountable. The model we developed to strip away these influences leaves us with a pretty good estimate of the performance attributable primarily to managerial choices (see sidebar *Isolating Firm-level Effects on Performance*).

Isolating Firm-level Effects on Performance

Our basic model is a regression. (We used quantile regression rather than ordinary least squares to avoid running afoul of the stringent parametric assumptions that come along with the latter.) We control for effects we are not interested in using the standard approach of including "control" variables. These controls "soak up" the variation in performance systematically associated with the features they capture.

Each industry, as defined by four-digit SIC codes gets its own variable, as does the level of competitiveness in each industry in each year. These capture stable industry effects (e.g., the importance of patent protection as a barrier to entry in pharmaceuticals) as well as competitive intensity within an industry, which changes over time (e.g., the rush into computers in the early 1980s).

Firm size is controlled for, as is a firm's market share, since these can be seen as endowments from prior years that affect performance independently of managerial competence. In addition, although capital structure is a managerial choice variable, we controlled for it as well: our belief is that success studies are properly about strategic and operating management, not financial engineering.

Firms listed on US markets via ADRs get their own control variable.

Finally, and perhaps unique to our study, survivor bias has been controlled for as well. The longer a firm stays in the sample the more opportunities it has to string together a number of great performances due to common cause variation. To account for this, we coded three variables. First, our database begins in 1966, as firms exiting the database prior to 1966 are not captured. Simply by being in business before our observation window opened, such firms may have stumbled onto a winning combination. Second, to account for the number of chances on a rolling basis that a firm has had to post great performances, we coded for the years a company had been observed within the sample window, updated annually. Finally, some firms may have relatively stable unobserved characteristics that affect their longevity, so every firm was coded from the start with its total number of observations. That way, any impact longevity has on performance is removed.

The power of 10

It's critically important to remember that when measuring firm performance for the purposes of identifying true outliers, absolute performance measures are of very little use. What matters is the relative performance of firms – how well they do compared to each other.

With that in mind, we structured our model to generate "decile rankings" for firms based on their "corrected" ROA values. That is, we ran the entire dataset of over 230,000 firm-year observations through our regression algorithm to create an ROA value stripped of everything but firm-level, or management, effects. Then for each year each firm received a ranking of 0 through 9 based on which decile of the total population it fell into.

Recall that the objective is to get some estimate of the variability of the underlying system so that we can separate common causes from special causes. To generate that estimate we took our set of decile rankings and calculated the frequency, as a percentage of total observations, with which firms in any given decile in any given year end up in any given decile the following year. In other words, we observed how likely is it that a firm in say, the 4th decile in year 1 would end up in each of the 0th through 9th deciles the following year. We did this for every decile. The result is a 10x10 "Decile Transition Matrix" (DTM) that we use to characterize the performance arising from common causes in the overall system of competition among companies (see Table 1).

The matrix shows that common cause variation can readily mislead our intuitions about what is and is not remarkable. Two features are especially noteworthy. First, the most likely outcome for a firm in any decile is to repeat that decile the following year. For all we hear about the pervasiveness of change, your best bet with firm performance, like the weather, is that tomorrow will look like today.

Second, this stickiness in performance is especially pronounced at the high and low ends of the spectrum. Here's a specific example: should a firm find itself in the 9th decile simply due to luck, there's a 49% chance that it can expect to remain there due to luck rather than anything special about the firm itself.

To determine precisely what kind of performance profiles this system can create, we ran 1,000 simulations of the last 41 years of all 22,000+ individual companies, giving each company the same life span that it actually had and starting in the same decile in which it first showed up in our database. Subsequent

Table 1. Decile Transition Matrix Probabilities of Moving from Starting Decile (t-1) to Outcome Decile (t), along with Expected Values of Decile Outcomes for each Starting Decile (t)

		Decile Outcome(t) Probabilities										Expected Value of Decile Outcome(t)
		0	1	2	3	4	5	6	7	8	9	
Starting Decile (t-1)	0	0.5077	0.1819	0.0890	0.0518	0.0348	0.0268	0.0235	0.0208	0.0169	0.0468	1.6311
	1	0.2091	0.3144	0.1735	0.0947	0.0581	0.0392	0.0313	0.0249	0.0203	0.0345	2.2092
	2	0.1072	0.1926	0.2375	0.1679	0.0959	0.0650	0.0449	0.0321	0.0252	0.0317	2.8608
	3	0.0652	0.1138	0.1726	0.2229	0.1611	0.0972	0.0609	0.0447	0.0316	0.0299	3.4588
	4	0.0421	0.0714	0.1146	0.1750	0.2104	0.1596	0.0990	0.0612	0.0388	0.0280	4.0492
	5	0.0300	0.0463	0.0736	0.1158	0.1714	0.2142	0.1695	0.0981	0.0497	0.0314	4.6814
	6	0.0222	0.0345	0.0515	0.0746	0.1113	0.1748	0.2276	0.1734	0.0881	0.0419	5.3420
	7	0.0176	0.0265	0.0365	0.0540	0.0694	0.1098	0.1719	0.2500	0.1914	0.0729	6.0573
	8	0.0148	0.0203	0.0279	0.0334	0.0455	0.0638	0.1009	0.1878	0.3277	0.1781	6.8212
	9	0.0229	0.0257	0.0274	0.0300	0.0341	0.0389	0.0558	0.0839	0.1905	0.4909	7.3654

performance was determined exclusively by common causes – that is, next year's decile of performance was randomly drawn from a distribution of outcomes with probabilities defined by the DTM. In other words, we re-ran the last 41 years of history to see what range of outcomes is possible given the observed variability of the actual system.

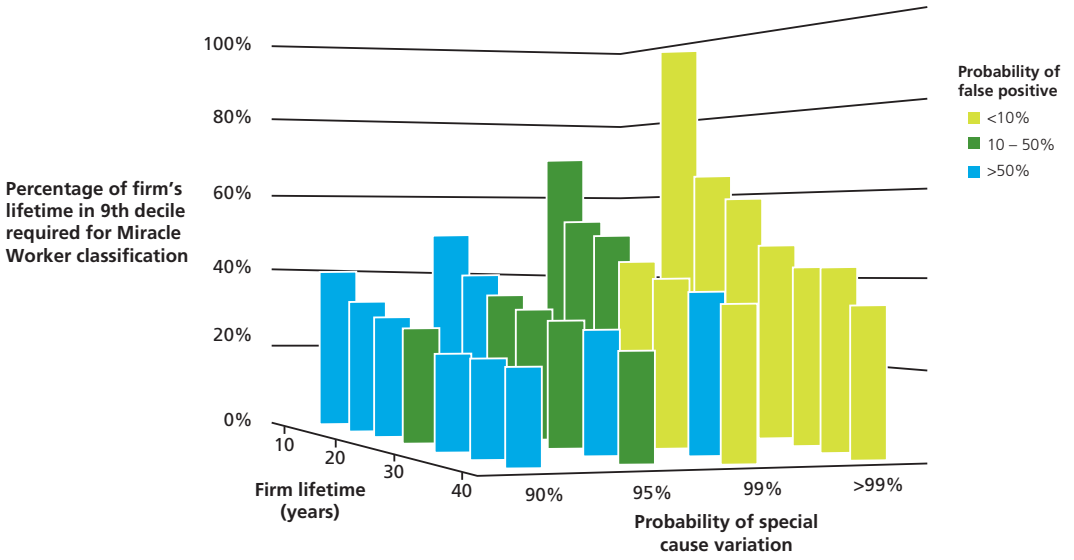
With these benchmarks, we could now begin looking for firms with performance profiles sufficiently improbable to warrant the hypothesis that special causes – that is, firm-specific attributes – are at work. For example, it turns out that for firms with exactly 11 years of data, there is a less than 4.3% chance that they would deliver five or more years in the 9th decile. That is, in 957 of our 1000 simulations, a firm with 11 years of data delivered no more than four years in the 9th decile. Now, a 5% confidence level that we have an outcome determined by special causes might seem pretty good, but we're only half-way home. So far, all we've done is identify the (im)probability of a given performance profile. Back to coin-tossing: we've estimated how unlikely it is that an individual will toss a specified number of heads; we haven't corrected for the number of coin-flippers in order to quantify the likelihood of false positives.

It turns out there are 856 companies in our sample with 11-year lifespans, so we'd *expect* to see 37 companies with 11-year lifespans and five or more years in the 9th decile (4.3% of 856). We *actually* see 45 such companies. The implication is that eight of those 45 companies have done something remarkable. Unfortunately, the statistical analysis cannot reveal which ones.

Consequently, if we pick a firm from this sub-population (11 years of data, five years or more in the 9th decile) we have an 82% chance (37 out of 45) of studying "lucky" firms (that is, performance due to common causes) rather than "good" ones (that is, performance due to special causes). To get the likelihood of false positives down to 5%, we'd have to insist on 10 years out of 11 in the 9th decile.

Figure 2 shows how many years a firm must end up in the 9th decile conditional on its lifespan to achieve specified confidence levels for special cause variation and false positive results. Close inspection of the values reveals just how demanding our benchmarks need to be if we are to place our trust in lessons learned from an examination of firm-level behavior.

Figure 2. Percentage of total lifespan required in the 9th decile of performance to meet various combinations of special cause variation and false positive benchmarks



Fooled by randomness

Our database is sufficiently extensive, and most success studies are sufficiently well-documented, that we can use our method to evaluate the performance of the firms they studied. The results of this meta-analysis will give us some indication of the degree to which previous success studies have been examining firms with plausibly remarkable performance or merely examining, at best, the right tails of a random distribution.

Before turning to the specific comparisons, it's worth making one more comment on method. It is common practice in success studies to examine specific periods of time: *What Really Works* takes a microscope to the decade 1986-1996; *Big Winners and Big Losers* focuses on 1992-2002. Companies with (allegedly) noteworthy performance within the chosen period are subjected to in-depth study.

In almost every instance the period chosen is only a slice of the full lifetime of the companies selected for clinical analysis. This creates a very real possibility of what is known as a "Texas Sharpshooter" problem, in which the target is defined only after the shots have been fired. When you set the target after you've shot, you can easily create the illusion of accuracy by placing the bull's eye over whatever random cluster of bullet holes you can find. Focusing on specific periods of time that capture only a portion of most focal firms' existences and separating out companies with suggestive performance profiles is a form of this error. Specifically, long-lived firms in a system with common causes characterized by the Decile Transition Matrix can be expected to generate periods of seemingly exceptional performance without behaving any differently than they did during periods of mediocre returns.

To get a sense of the magnitude of this problem, consider that both *What Really Works* and *Big Winners and Big Losers* see Campbell Soup as a company worthy of study. The period under examination in the two studies overlaps by five years (from the beginning of 1992 to the end of 1996). Yet *What Really Works* (1986-1996) holds up Campbell Soup as a "winner" while *Big Winners and Big Losers* (1992-2002) sees the company as a "big loser." Who's right? Both…and neither.

Campbell Soup, across the sweep of its lifetime, has had its ups and downs. If you look at a "down" period, you'll conclude the company is poorly managed; if you look at an "up" period, you'll conclude it is well managed. And if by chance you happen to begin your analysis at an inflection point, you'll conclude management is either brilliant or benighted.

Looking simply at stock prices (both studies used TSR as their measure of performance) Campbell Soup had a great run from 1986-1998: the stock went from $4.17 to $54.61, outstripping the market more than 3-fold. Then from 1998-2002 it declined to $23.47, such that for the full 16 year period covered by the two studies it was even with the DJIA, where it has stayed, with some fluctuation, ever since.

The moral is that if you want to say anything about the performance of any company you have to look at all the data available for it (i.e., how many "heads" it actually flipped) and look at the company in the context of the relevant peer group (i.e., how many other companies had the same number of tosses). For Campbell Soup, the data are suggestive, but inconclusive: the company's pattern of ROA results is strongly positive and quite rare, so the "big loser" claim seems suspect.[1] At the same time, the firm has such a large cohort that there is a disheartening likelihood that its particular profile could arise by chance, and so even if it is an exceptionally good company, it's difficult to distinguish its performance from that of a merely lucky firm. Consequently, claims of excellence or incompetence by these authors should be dismissed as artifacts of how the data were collected and analyzed. In short, the numbers say that we can't say anything. It would appear, then, that it's a lot easier to decide whether Campbell soups are mmm-mmm good than it is to determine whether Campbell Soup is mmm-mmm great.

But couldn't one argue that Campbell Soup experienced different "regimes" – sometimes well managed, sometimes not – which were reflected in periods of stronger and weaker performance? Couldn't the company have been a "winner" over one 10-year period, and so a source of insight into great

[1]Campell Soup's performance is so close to exceptional that how one defines the industry in which it competes turns out to make the difference: defined narrowly, Campbell Soup is "ambiguously good" – that is, not clearly distinguishable from extreme good luck; defined more broadly, Campbell Soup stands above even very lucky firms in the magnitude and duration of superior profitability.

management, and a "loser" over a different 10-year period, and so an example of what not to do? That's a tough claim to make in this specific instance since the two ten-year periods overlap by 50%. But even if it's conceptually possible, it turns out that the statistical hurdle is even higher when attempting to distinguish between random fluctuations and true "state changes" over the lifetime of a single company.

The DTM allows us to assess whether a company's performance profile as a whole is statistically exceptional. We employed a very different machinery to determine whether companies showed evidence of trends in performance over time. The short answer is that very few show patterns – either consistent streaks or rising or falling trajectories – that would justify claims of different performance regimes (see sidebar *Identifying Patterns in Firm Performance*).

None of this should be taken to suggest that management doesn't matter in firms with statistically unremarkable profiles. Rather, we're arguing that there is nothing demonstrably different, based purely on an examination of performance, about what management achieved in those firms. Remember, performance that is defensibly attributable to nothing other than common causes is still caused. But those causes are available, in a real sense, to all comers. The players in the drama of competition will certainly feel that they are working hard…because they are. But they are working no harder and, more to the point, no more effectively than the norm.

Identifying Patterns in Firm Performance

The canon of business mythology is replete with tales of transformation, of greatness undone by hubris only to find salvation in six sigma and customer focus. We are trained to believe that dramatic swings in performance are necessarily a consequence of dramatic and fundamental changes in behavior wrought by great leaders, either heroic or humble – depending on which success study you read.

All the same traps await those who would analyze seeming changes in the performance of a single firm as befall those who compare performance among firms. Specifically, we must separate out the signal from the noise when assessing whether a firm's performance has actually changed over time, or we risk chasing patterns in cloud formations.

Recent advances in the statistical theory have made it possible to do just this. Working with James Scott, a statistician at Duke University, we used a Bayesian modeling algorithm to infer archetypal patterns of performance from our full data set. Thirteen trajectories of performance emerged, which fell into six general categories: rising, falling, flat-high, flat-low, bouncing, and random; this last is synonymous with "no evident pattern".

Our model does not unambiguously allocate a firm to any trajectory, but provides a probability distribution for each firm across all of the archetypes. Although there are no definitive cut-offs, firms with a greater than 50% chance of belonging to any category other than random should be seen as at least potentially having changed their performance, and so a hypothesis of "special cause" variation within that firm over time is defensible.

Of the more than 21,000 firms in our database covering the period 1966-2006, fewer than 400 meet this criterion. None was used by any of the success studies we examined as an example of a performance trajectory consistent with what we found.

We do not take the paucity of firms with clear changes in performance as evidence that very few firms have ever changed their performance. We take it as evidence that very few firms have ever changed their performance enough to be distinguishable from the roar of white noise arising from the volatility endemic in a dynamic and unpredictable marketplace.

No doubt better tools than we used will be developed and better researchers than we are will tackle the problem. Until then, the lack of evidence forces us to withhold our assent, however much we believe in the plausibility of the claim. After all, even the Higgs boson is only a hypothesis until the Large Hadron Collider actually finds it.

The standard of excellence

There's one more step before we can determine what constitutes statistically remarkable performance: we have to define a benchmark. This step is ultimately somewhat arbitrary. Any combination of deciles is a potential standard of excellence. We could look for companies that had whatever number of 4th, 7th and 9th decile years that were statistically improbable, as defined by the simulations generated using the DTM. That should seem a rather arbitrary profile to want to examine, but at some level, so is any particular profile. We need a theoretical reason to study one pattern of outcomes and not another.

For the purposes of our success study, we have defined two categories of exceptional performance, tentatively labeled "Miracle Workers" (MWs) and "Long Runners" (LRs). The former deliver whatever number of 9th decile years is statistically unlikely given their lifespans, while the latter deliver whatever number of years in the 6th-8th decile band that is similarly improbable. Our definition of excellence is motivated by a desire to contrast the behaviors of firms that do exceptionally well for a long period of time with those who do merely well (rather than exceptionally well) over long periods.

A fuller description of our database of MWs and LRs is under development. For now, note only that of the 22,403 distinct companies captured in our 41 years of data, this process (the DTM-based simulations) and these standards (9th decile for MWs, 6th-8th deciles for LRs) identified 169 MWs and 184 LRs among firms that are traded in US public markets. (Applying these standards to DTM-based simulations for TSR reveals that there are essentially no companies with exceptional performance; see sidebar *No Surprises*.)

To assess the degree to which other studies have chosen truly remarkable firms upon which to base their analyses, we need only compare their lists of allegedly exceptional firms first with the firms that our method is able to categorize, and then with the list of 353 companies that are either MWs or LRs. Companies that we can categorize but that are neither MWs nor LRs are, by our lights, statistically unremarkable, with performance profiles determined by common causes. (With one important caveat; see sidebar *Mind the Gap*.)

Finally, to avoid basing our conclusions on data unavailable to researchers at the time they conducted their studies we examined each company's performance profile from the time it first appears in our database to the end point of each individual study.

With all this in place, we can finally report our results, summarized in Table 2.

Table 2.

	# High Performers	# We Categorize	% (#) @ >90% of SCV	% (#) @ >90% SCV and <10% FP
Alchemy of Growth	29	11	73% (8)	55% (6)
Big Winners/Big Losers	9	8	38% (3)	25% (2)
Blueprint to a Billion	26	24	63% (15)	13% (3)
Breakthrough Company	9	6	83% (5)	17% (1)
Built to Last	18	14	50% (7)	14% (2)
Creative Destruction	11	9	44% (4)	33% (3)
Good to Great	11	8	63% (5)	0% (0)
Granularity of Growth	42	24	63% (15)	29% (7)
In Search of Excellence	14	13	62% (8)	23% (3)
Peak Performance	21	11	45% (5)	9% (1)
Profit from the Core	32	18	72% (13)	39% (7)
Stall Points	67	30	33% (10)	13% (4)
What Really Works	14	13	69% (9)	36% (6)
Total (No Duplicates)	**288**	**184**	**57% (104)**	**23% (42)**

Note: Totals are not the sum of the columns because some companies are used by several studies (e.g., General Electric and 3M). The somewhat cryptic headings of the two right-most columns should be read as follows. For each study, we identify those firms that it features as high performing companies that we can categorize using our method that have a 90% chance of having a performance profile that is the result of special cause variation (SCV) at a 90% confidence level. That number is given in parentheses, and is expressed as a percentage of the total number of high performing companies mentioned in a study that we can categorize. We then "correct" this number for the likelihood of a given statistically remarkable high performer being a "false positive" (FP) – that is, that the firm owes its seemingly improbable outcome to a large sample size rather than truly remarkable performance. Our cutoff for false positives is 10%; that is, we must be 90% sure that a given result is a true positive before we are willing to state that a firm has a statistically remarkable performance profile given the subpopulation of firms from which it was drawn.

We examined 13 popular success studies. These studies identified 288 distinct company-time period combinations as high performers, of which we were able to categorize 184 using our approach. Of these, 104, or 57%, had a performance profile that had a 90% chance of resulting from special cause variation. The range, however, is quite large: from a low of 33% in *Stall Points* (10 of 30 companies) to a high of 83% in *The Breakthrough Company* (5 out of 6). Consequently, by this initial cut, some studies would appear to be on shaky ground, while others seem to have a solid foundation.

But we were shocked to find that, when we insisted on no more than a 10% likelihood of false positives, the 13 success studies we examined had, collectively,

No Surprises

Total Shareholder Return (TSR) is so pervasive a measure of firm performance in success studies that it is incumbent upon us to at least examine how it fares under our method, whatever our philosophical objections to its use.

Constructing a Decile Transition Matrix (DTM) for TSR figures is instantly revealing. Unlike the DTM for ROA or other performance measures such as Economic Margin (EM), TSR shows little evidence of stickiness: the likelihood of moving from any given decile to any other is very nearly uniform (see table). This means that every year every firm starts with a blank slate, and has to impress shareholders all over again. This is consistent with our argument that share prices reflect expectations for the future, and that as those expectations rise, share price quickly rises to reflect that. In order to achieve high-decile returns again, the firm must surprise on the upside again – and the market would appear to be sufficiently efficient to make it impossible to know whether or not that would happen.

(As a thought experiment, if there were a systematic bias in the DTM, and firms had a higher or lower likelihood of performing in a particular way conditional on how they performed this year, one could go long or short on those firms and reap surplus returns.)

Using the TSR-DTM and the same method detailed in the main paper to look for Miracle Workers and Long Runners based on TSR, we find that there are none to be had. Literally. In other words, markets rapidly bid up any firm that is delivering exceptional returns so that it very quickly is no longer delivering exceptional returns. The seemingly astonishing runs of high TSR that some companies deliver is simply the right tail of a stable distribution of returns.

		Ending TSR decile (t+1)									
		0	1	2	3	4	5	6	7	8	9
Starting TSR Decile (t)	0	0.1630	0.1126	0.1000	0.0879	0.0823	0.0766	0.0771	0.0842	0.0912	0.1250
	1	0.1193	0.1006	0.1094	0.1019	0.0969	0.0946	0.0879	0.0918	0.0936	0.1040
	2	0.0972	0.0984	0.1016	0.1057	0.1043	0.1001	0.1045	0.0992	0.0965	0.0925
	3	0.0878	0.0981	0.0992	0.1014	0.1071	0.1029	0.1056	0.1056	0.0992	0.0930
	4	0.0805	0.0941	0.1047	0.0996	0.1000	0.1146	0.1047	0.1070	0.1045	0.0902
	5	0.0786	0.0892	0.0975	0.1045	0.1039	0.1103	0.1090	0.1071	0.1068	0.0930
	6	0.0769	0.0955	0.1011	0.1009	0.1048	0.1072	0.1073	0.1070	0.1046	0.0946
	7	0.0794	0.0937	0.0961	0.1047	0.1030	0.1087	0.1087	0.1026	0.1093	0.0939
	8	0.0896	0.0971	0.0957	0.1003	0.1009	0.1017	0.1029	0.1091	0.1033	0.0994
	9	0.1131	0.1115	0.0971	0.0963	0.0913	0.0927	0.0880	0.0973	0.1004	0.1123

only 42 companies, or 23% of those studied, with defensibly remarkable performance.[2] Some variation remained, of course: *Alchemy of Growth* now comes out on top with 55% of its high performers clearing our bar for demonstrable excellence, while *Good to Great*, for which we were able to categorize nine of its 11 companies, had precisely none that were statistically remarkable. (See sidebar *The Benefit of the Doubt*.)

If one finds this analysis convincing, the implications are inescapable and confounding to the success studies we have examined, and very likely just about all others besides. When the best of these efforts can claim little more than half of their exceptional firms are defensibly exceptional – and when the most famous and influential have no bona fide excellent firms at all on which to base their conclusions – the alleged determinants of success are no more than patterns imposed on randomness. That is not science. It is astrology.

There is little to be gained from further rhetorical flourish, but we do strongly encourage the reader to reflect on the potential significance of this observation.

Mind the Gap

You might have noticed that by defining Miracle Workers as "enough 9th decile years to be improbable" and Long Runners as "enough 6th-8th years to be improbable" we left open the possibility that some firms might have a good many 9th decile years without enough to qualify as MW and because of that also have too few 6th-8th years to be improbable. In short, a firm could be too good to be a Long Runner but not good enough to be a Miracle Worker. If previous success studies were looking at firms with a performance profile that falls in the gap of the two categories we have defined for our research purposes, they would be studying defensibly exceptional firms that our method overlooked.

We choose not to study firms with this particular profile because we feel we won't be able to distinguish defining management behaviors. However, for the purposes of assessing whether other studies have been looking at randomness, we do need to take this wrinkle into account. To do so, we re-ran our analysis looking for companies with performance in the 6th-9th deciles frequently enough to be improbable. Call them "Super Long Runners" (SLRs) – a category that, on its face, is more inclusive than the conjunction of our Miracle Worker and Long Runner categories.

As it turns out, we ended up with even fewer SLRs than we had MWs and LRs combined. The reason is straightforward: as we expand the performance band within which a firm can fall and still be considered exceptional, the likelihood of landing within that band by chance alone increases. Consequently, the number of observations within that band required to rise above the statistical noise goes up significantly, and few firms stick around long enough to have any chance at all of differentiating themselves.

Nevertheless, there are twenty-two SLR firms that were neither MWs nor LRs in our original analysis. Our assessment of the randomness infecting other studies counts these firms as statistically exceptional in addition to those companies that meet our definition of MWs and LRs. In other words, we have used the most generous definition of "exceptional", and mixed two conceptually different (even if closely related) benchmarking approaches.

[2] Note that this assessment is made using a 90% confidence interval, since our estimate of how many firms with a given performance profile one would expect is an estimate, itself subject to a distribution. If we were to make these assessments using the expected value as a point estimate, the percentage of true high performers rises to 29%. This doesn't seem to us to be a material impact.

The Benefit of the Doubt

In this sort of large-scale statistical modeling there are many assumptions and simplifications one must make. Usually there is no best answer for any one of them, and so it is important to test the sensitivity of one's final results to the impact of specific choices.

Perhaps the most theoretically vexing is how best to control for industry effects. When using a decades-long time series from Compustat, the most convenient industry categorization scheme available is the Standard Industrial Classification (SIC) system. Companies are classified first into one of ten divisions (A through J), then into two-digit major groups, three-digit industry groups, and finally into four-digit industries.

When controlling for industry effect, one has to choose which level of the SIC system to use. In most academic literature, the two-digit level is considered the least specific that is still defensible. However, much of the work on this topic sees the two-digit similarities capturing industry *relatedness*, rather than similarity in the sorts of industry-level characteristics that can affect corporate results. Defining industry at the four-digit level tends to be the more common approach when controlling for industry-level effects on firm-level performance.

But this choice is not without its own drawbacks. In particular, it raises the spectre of over-specifying the regression models by leaving us with industries that have too few participants to allow for accurate estimates of the underlying phenomenon of interest. Defining industry at the two-digit level would leave more "variability" to be "explained" by firm-level performance, and so likely increase the number of firms classified as "remarkable."

Categorization Sensitivity Analysis

			Percentage of companies categorized by our method as exceptional		
	# High Performers	# We Can Categorize	Original Analysis	Two Digit SIC	Miracle Workers on ROA Only
Alchemy of Growth	29	11	55%	36%	0%
Big Winners/Big Losers	9	8	25%	25%	13%
Blueprint to a Billion	26	24	13%	21%	8%
Breakthrough Company	9	6	17%	33%	17%
Built to Last	18	14	14%	57%	7%
Creative Destruction	11	9	33%	33%	0%
Good to Great	11	8	0%	25%	0%
Granularity of Growth	42	24	29%	50%	0%
In Search of Excellence	14	13	23%	38%	0%
Peak Performance	21	11	9%	36%	0%
Profit from the Core	32	18	39%	33%	17%
Stall Points	67	30	13%	13%	0%
What Really Works	14	13	46%	46%	0%
Total (No Duplicates)	**288**	**184**	23%	33%	4%

A second choice worth testing is the types of performance profiles worthy of being called "exceptional." In the main text, we have chosen to include companies that clear our "Miracle Worker" and "Long Runner" benchmarks on either Return on Assets (ROA) or Economic Margin (EM). We have also included so-called "Super Long Runners" on either measure, for although this is a category of no theoretical interest to us, it is worth including in the interests of completeness. This is the most inclusive definition of "remarkable" our method will permit.

More exacting standards would presumably reveal fewer companies with statistically exceptional performance. For example, the calculation of EM requires estimating a number of parameters, whereas ROA is built exclusively on publicly available data. In the clinical research informed by the method described here, companies that are Miracle Workers on ROA will be the "focal companies," and their behaviors will be contrasted with Long Runners.

The table on the previous page shows the impact of defining industry at the two-digit level and of comparing the incidence of exceptional performance in success studies when only Miracle Workers are accepted as the benchmark of excellence. We have reproduced the relevant results from Table 2 in the main text to facilitate comparisons.

As expected, the two-digit SIC industry definition increases the number of companies identified by other success studies that our method characterizes as exceptional, but not appreciably: the best of the lot is still running under 60%, and overall the population of studies is still studying randomness fully two-thirds of the time. When we restrict exceptional behavior to the Miracle Worker standard, eight of the 13 success studies we looked at have no clearly exceptional companies in their studies, and overall only 4%, or eight out of 184, cleared the Miracle Worker benchmark.

The fable of the fables

We have taken pains to be transparent in explaining our method because the weight one gives to our findings is determined by the defensibility of our underlying assumptions. But if you find our premises true and our reasoning valid, then our argument is sound: the explicit claims of these studies – to reveal the principles of success, or an "organizational physics" to quote one author – must be rejected. These authors cannot be seen to have achieved what they set out to achieve because they were not studying what they said they were studying. Rather, just as patterns perceived in ink blots are seen by some to reveal underlying character traits, the secrets of success identified in what is in the end, at best, a randomly chosen sample from the right tail of a distribution almost certainly says more about the researcher than it does about the evidence. This doesn't mean, however, that you should necessarily dismiss the advice offered in existing success studies. The authors are savvy observers of the business world. Their recommendations can be useful, but more in the manner of fables than evidence-based advice. And we use fables very differently from science.

For example, no one reads "The Tortoise and the Hare" and, faced with a chance to bet on such a race, chooses the tortoise. Rather, people take from this tale the idea that there is merit in perseverance while arrogance can lead to a downfall. Similarly, because the prescriptions of most success studies lack an empirical foundation, they should not be treated as how-to manuals, but as a source of inspiration and fuel for introspection. In short, their value is not what you *read* in them, but what you read *into* them.

The truth is a harsh mistress

We have been uncharacteristically forthright in our criticisms of popular and influential management books authored by accomplished and respected academics and business professionals. We have been deliberately, but we hope not unnecessarily, provocative. In so doing, it is only fair to recognize the limitations of the work we present here.

First, and most importantly, the method we describe here is only one way to measure the remarkableness of any given firm's performance. There is more than one way to skin this cat, and the choice of method often matters. But as far as we know, no success study we've examined has engaged this fundamental question in a substantive way.

Second, as one of the first large scale attempts to explore the problem of distinguishing between luck and skill in corporate-level performance, it is all but certain that material improvements lie ahead. The proposition that we got it exactly right on our first try is unlikely. Consequently, it is our hope that this monograph is not the end of our attempt to contribute to the advancement of management theory and practice, but rather the beginning.

To that end, we have taken our method for identifying exceptional performers and identified "triplets" in over a dozen industries consisting of a Miracle Worker, a Long Runner, and an "Average Joe" (AJ) – a company with statistically unremarkable life span, performance level, and performance variability. Our intent, in the time honored tradition of the success study, is to compare and contrast the behaviors of these firms and thereby tease out the necessary and sufficient conditions for exceptional performance.

In so doing, we hope eventually to expose our own work to the kind of careful and, no doubt, critical analysis from a broad community of scholars and practitioners. Allow us to state here for the record that if we should in any way succeed in this endeavor, it is only because we have had the opportunity to learn so much from others' works – learning that was possible only by placing their efforts under the microscope. Prior work may well be subject to the shortcomings that we have identified, or fall short in ways specified by others. But by pointing out the mote in another's eye we in no way suggest that we have, or shall not have, any occlusions of our own. Nevertheless, progress requires that we both identify the faults in earlier attempts to find truth and risk making new mistakes as we continue the quest. We have here attempted only the first, and by far the easier, half of that obligation.