Good to Great, or Just Good?

by Bruce Niendorf and Kristine Beck

Executive Overview

Good to Great has been on *BusinessWeek's* best-seller list since its October 2001 release. In *Good to Great*, author Jim Collins identified a set of 11 firms as great, then used them to derive five management principles he believed led to "sustained great results." We contend that due to two fatal errors, *Good to Great* provides no evidence that applying the five principles to other firms or time periods will lead to anything other than average results. We explain the two errors and empirically test our contention. When ranked with the 2006 Fortune 500, the 11 *Good to Great* firms have an average ranking of 202nd. In addition, in terms of long-term stock return performance, the *Good to Great* firms do not differ significantly from the average company on the S&P 500. Our evidence is consistent with the conclusion that although the *Good to Great* firms may be good, they aren't great.

The book *Good to Great* (hereafter GTG) has attained long-running positions on the *New York Times*, *Wall Street Journal*, and *Business Week* best-seller lists, has sold 3 million hardcover copies since publication, and has been translated into 35 languages.[1] Since its publication, GTG has become popular worldwide not only in the boardroom, such as at Barnes-Jewish St. Peters Hospital in St. Louis, but also in classrooms across the country, such as at the Wharton School at the University of Pennsylvania.[2] GTG has also had great influence. For their 2007 Business Book of the Year Award, the *Financial Times* asked readers to choose the "best book of all time" from a short list of five books. The list was created by soliciting suggestions from a variety of business executives, including GE's Jeff Immelt and eBay's Meg Whitman. GTG was listed as one of the top five books. The fact that GTG shares a position on this list with Adam Smith's *Wealth of Nations* (1776) and Peter Drucker's *The Effective Executive* (1966) places it among the most influential business books ever written.

We thank Peter Cappelli, an anonymous reviewer, and Barbara Rau for their helpful comments.

[1] http://www.jimcollins.com/bio/index.html.

[2] See "Good to Great, a B-School Staple," Businessweek Online, January 19, 2006.

What Collins Did

In GTG, author Jim Collins (2001) and his 21-person research team used a set of screens to sift through 1,435 companies and identify a list of 11 elite firms that "made the leap" to business greatness. In addition to screening for high stock returns, they screened for firms they felt outperformed their industries, were well established, and were still on an upward trend 15 years after they made their transition to greatness. The 11 companies were:

(a) Abbott Laboratories
(b) Circuit City
(c) Fannie Mae
(d) Gillette
(e) Kimberly-Clark
(f) Kroger
(g) Nucor
(h) Philip Morris (now Altria)
(i) Pitney Bowes
(j) Walgreens
(k) Wells Fargo

After identifying these 11 firms, Collins' next step was to contrast them with a set of comparison firms he felt had failed to make the leap from good to great. By identifying characteristics common to the great firms, but lacking in the failed firms,

***Bruce Niendorf** (niendorf@uwosh.edu) is Associate Professor of Finance, University of Wisconsin Oshkosh.*
***Kristine Beck** (beck@uwosh.edu) is Associate Professor of Finance, University of Wisconsin Oshkosh.*

Table 1
Commonalities Found by Collins in the 11 *Good to Great* Firms

● Level Five Leadership
○ Leaders focus on the firm rather than themselves
● First Who, Then What
○ "Get the right people on the bus before you decide where to drive it"
● Confront the Brutal Facts
○ Cannot make good decisions without knowing the facts
● Hedgehog Concept
○ Know one thing rather than trying to know many things
● Build Your Company's Vision
○ Preserve core values but be willing to change operating practices and business strategies

Collins derived the five GTG principles. Table 1 presents the GTG principles that Collins claimed would lead to "sustained great results" (p. 190).

What Collins Did Wrong

We contend that due to two fundamental research design errors, GTG does not show that the five GTG principles lead to "sustained great results," as Collins claims. Rather, it shows only that the 11 GTG firms had these principles in common during the specific time period studied by Collins. The difference in these two conclusions is enormous, and the principal implication is this: GTG provides absolutely no evidence that applying the GTG principles to other firms during other time periods will lead to anything other than average business performance. Thus, Collins has not identified "timeless, universal answers that can be applied by any organization" (p. 5).

The two fatal errors in GTG are data mining and mistaking association for causation. In this article we discuss these two errors, and then test the performance of the 11 GTG companies after the time period used in the GTG study. By doing so, we find evidence that addresses the generalizability of the GTG principles. Our test results provide no empirical evidence for Collins' claims that applying the GTG concepts leads to "sustained great results." A discussion of data mining and correlation and causation, as well as a description of our empirical tests, are below. These are followed by our results.

Data Mining

The use of data mining is a fundamental research flaw in GTG. Data mining is the process of collecting and searching for patterns in data and then once patterns are identified, formulating explanations that are treated as underlying causes or principles. The problem is that since the patterns may depend completely on the specific time period and dataset gathered, data mining provides no legitimate evidence of applicability outside the sample firms or time period. This problem with data mining (also called data dredging, data grubbing, or fishing) is aptly described by Hand (1998, p. 112):

> *By definition, data that are not simply uniform have differences which can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure. The object of data analysis is not to model the fleeting random patterns of the moment, but to model the underlying structures which give rise to consistent and replicable patterns. To statisticians, then, the term data mining conveys the sense of naïve hope vainly struggling against the cold realities of chance.*

Thus, according to Hand, researchers using data mining may well draw conclusions that are based either on purely random patterns or on patterns that exist only in the sample firms or time period studied.

Collins' method of identifying his five principles in GTG is data mining. By Collins' description, he started with a list of companies that appeared in the Fortune 500 rankings and screened them four times to arrive at the 11 GTG companies. He then studied the 11 GTG companies relative to 11 comparison companies to "discover the essential and distinguishing factors at work" (p. 3). This is data mining. As Collins put it: "It is important to understand that we developed all of the concepts in this book by making empirical deductions directly from the data. We did not begin this project with a theory to test or prove. We sought to build a theory from the ground up, derived directly from the evidence" (p. 10). Given the data mining methodology in GTG, Collins provided no evidence that his five principles are anything other than five "fleeting random patterns." In other words, if Collins is correct, his statistics suggest that it's just luck. As

Walker (2006) noted, this "creates a neat little circle of inference: Great firms do these things, so do these things and your firm will be great" (p. 120).

If used appropriately, data mining can be an extremely powerful tool. To use data mining appropriately, a researcher would search data for patterns, formulate explanations for those patterns, and then statistically test those explanations to identify their applicability outside the sample firms and time period used to identify the patterns. This method is used extensively in medical research, for example. The critical and necessary ingredient to its success is extensive and painstaking testing of any patterns identified in the data. Collins provided no such tests in GTG. Rather, Collins presented his potentially "fleeting random patterns" as five principles that will lead a firm to greatness.

Association vs. Causation

A second problem in GTG involves interpreting association (correlation) as causation. In GTG, Collins chose firms he defined as great, investigated the qualities of those firms relative to a set of comparison companies, and identified five characteristics common to the "great" firms. In other words, Collins identified an *association* between the great firms and the five characteristics; he found that greatness and the characteristics appeared together in the 11 GTG firms during the time period he studied. He did not, however, provide any evidence of the *causation* he claimed: Established Company or Startup + *Good to Great* Concepts = Sustained Great Results (p. 190).

The GTG methodology did not show that these characteristics made the firms great. It showed only that the 11 firms had these characteristics in common. This error may be much more damaging in practice than data mining since it leads business managers to believe that applying the five principles will move them toward greatness. In reality, there are at least two alternative explanations. First, being great allows firms to live by the principles; for example, being profitable allows a firm to spend more to hire the right people. Second, there could be an outside factor, such as an unidentified environmental factor, that simultaneously creates sustained results and the principles. Again, GTG provided no evidence of causation, only claims.

In the epilogue to GTG, Collins provided answers to frequently asked questions. The third question regarded statistical significance given that the total study size was only 28 companies. Collins replied that he "engaged two leading professors to help us resolve this question" and that they "concluded that we do not have a statistics problem" (pp. 211–212). In a further explanation, Collins quoted one of the professors, William P. Briggs from the University of Colorado:

> *What is the probability of finding by chance a group of 11 companies, all of whose members display the primary traits you discovered while the direct comparisons do not possess those traits (p. 212)?*

Collins then stated:

> *He (Briggs) concluded that the probability is less than 1 in 17 million. There is virtually no chance that we simply found 11 random events that just happened to show the good-to-great pattern we were looking for. We can conclude with confidence that the traits we found are strongly associated with transformations from good to great (p. 212).*

This rationale does not apply to the methodology actually used in GTG. Collins didn't start by identifying the "primary traits" and then evaluate the characteristics of any companies that displayed them, as the explanation implied. Collins started with the companies and then looked for common traits in those companies. Thus, the rationale provided by the professors does not apply to Collins' methodology. In addition, given Collins' actual methodology, the "11 random events" were not random events—they were events/companies that were intentionally and painstakingly selected, and then mined for their commonalities. The question the professors should have been asking was:

> *What is the probability that, in a sample of firms selected for like characteristics with respect to performance, you could find five points of commonality between them?*[3]

[3] This statement reflects the fact that the probability the professors were attempting to quantify is actually a conditional probability. The question could also be stated as "What are the odds that 11 selected firms share five attributes *given* that they have outstanding stock returns?"

We propose that this probability is far closer to 100% than to 1 in 17 million. Although Collins used 11 handpicked comparison firms, his methodology doesn't tell us if the practices in the 11 GTG firms are different from those of the other firms in the Fortune 500. Poor performers may have Level 5 leadership as well, but we'd never know this from his study.

The Right Way

We appreciate Collins' efforts to think outside the box and "not begin this project with a theory to test or prove" and to "build a theory from the ground up" (p. 10). However, identifying a theory, setting up hypotheses, and then sampling, testing, and drawing conclusions regarding the validity of the theory is the method necessary to justify the claims that Collins made. To clarify our position, we do not argue that Collins' conclusions and claims are incorrect, but that he provided no evidence that they are anything other than random patterns.

To provide bases for his claims, Collins (or others) needs to pick up where he left off in GTG and test the validity of his claims using classical statistical analysis. The first step in this process would be to identify hypotheses regarding the relationship between the five GTG principles and firm success (measured, for example, by stock returns). These hypotheses would be grounded not only in Collins' inductive work but also in prior management theory research addressing the same relationships. For example, there exists a vast body of literature on the importance of valid selection procedures to get "the right people on the bus." After formulating specific hypotheses, the next step would be to collect a sample of publicly traded firms that is representative of all firms. Finally, one of a wide variety of statistical methods, such as regression analysis, could be used to test for support of the relationships specified in the hypotheses.

What We Did

Methodology

In addition to our explanations of the statistical errors in GTG, we provide tests of the performance of the GTG companies. We do this to provide empirical evidence in addition to our statistical arguments that there is no basis for Collins' claims. We investigated the generalizability of the GTG results in two ways. First, we checked to see how the performance of the 11 GTG firms compared to the Fortune 500 firms in the 10 years following the time period used in GTG. If the 11 firms were truly great, they should have outperformed their peers in time periods other than the period used to select them. We used 10-year investment returns to measure performance.[4] Specifically, we ranked the 11 GTG firms with the 2005 Fortune 500 firms by 10-year return to investors to see how the 11 firms performed during 1995 through 2005. One of the screening criteria used to select the 11 GTG firms was that the firms still be on "an upward trend" at the end of the time period considered in GTG. To the extent that this criterion was measurable and effective, this should have created an upward bias in the ranking of the 11 GTG firms, thereby making them look better in our rankings than they otherwise would.

As a second test, we compared the return performance of the 11 GTG stocks to the return performance of the S&P 500 after the sample period used in GTG.[5] If Collins' results are generalizable, i.e., they apply to other time periods, these firms should outperform their peers in other time periods. We hypothesized that there was no significant difference between the returns of each of the 11 GTG firms and the returns of the S&P 500 in the time period after the one Collins used to select the firms. This hypothesis can be stated as:

> *H_0: There is no difference between the mean return of each GTG firm and the mean return of the S&P 500 index.*
>
> *H_1: The mean return of each GTG firm is greater than the mean return of the S&P 500 index.*

[4] Collins and his team also used long-term stock returns as a measure of performance in choosing the 11 firms.

[5] We used the S&P 500 index because there is not a consistently quoted return index calculated for the Fortune 500, which makes sense given the relatively frequent changes in the list. Many of the firms in the Fortune 500 in any given year are also in the S&P 500. Also, as an anonymous reviewer pointed out, Collins did try to benchmark his firms against a set of companies in a matched pair analysis; unfortunately, inspection of the matched pairs shows that the matches were highly flawed.

Data

We tested this hypothesis in two time periods: during the post-GTG sample period (1995–2005) and with as long a time period as possible without overlapping GTG's selection period. One of the GTG screening criteria was that to be great, a firm had to have a 15-year period demonstrating a great cumulative stock return (more than three times the general market) preceded by a 15-year period demonstrating a good cumulative stock return (less than one third its great cumulative return). The point in time between the 15-year periods was deemed the company's "transition point"—the point at which it made the transition from good to great. The year of this transition point for the GTG firms ranged from 1964 to 1984. To get as long a time period as possible, we tested each firm's returns against the returns on the S&P 500 during the time period beginning with each firm's transition point + 15 years through 2005.

Results

We ranked the 11 GTG firms with the 2006 Fortune 500 firms based on return to investors in the period 1995 through 2005. According to Collins' theory, if the 11 GTG firms were run by the most skilled management teams, they should have demonstrated "great" performance relative to their peers in time periods other than those used in GTG. Table 2 shows the top 10 Fortune 500 firms, ranked by return to investors. This list, topped by NVR and Dell, produced 10-year average returns to investors ranging from 53% (NVR) to 31.6% (Kinder Morgan Energy), with a top 10 group average 10-year return of 37.8%. In contrast, the 11 GTG firms provided a 10-year group average return to investors of 10.9%, less than one third the average return of the top 10 group. Although several of the GTG firms produced returns that were good (for example, Walgreens at 20.2%), the returns ranged down to a 6.2% return produced by Kimberly-Clark. In terms of the overall Fortune 500 rankings, the GTG firms ranged from 62nd (Walgreens) down to 302nd (Kimberly-Clark), with an average ranking of 202. Thus, considering both the rankings and the returns themselves, we find no evidence of greatness in the 10-year time period following the sample period used in GTG.

The results in Table 2 are especially strong when considered in light of the upward bias in the rankings of the 11 GTG firms created by Collins' requirement that to be considered great, "the company should still show an upward trend" at the end of the period in which it was selected. The GTG companies are clearly strong companies managed by competent managers, but ranking an average of 202 even with the help of the bias and having average returns that are one third of the top 10 do not support GTG's label of greatness.

The results in Table 3 also shed light on the question of the generalizability of Collins' findings to other time periods. Panel 1 of Table 3 provides the results of paired t-tests of the null hypothesis that there is no difference between the mean return of the GTG firms and the mean return of the S&P 500 index during the time period 1995 through 2005. As the results show, in 10 of the 11 cases, we failed to reject the null hypothesis. We must therefore conclude that in those 10 cases, no significant difference exists between the returns of the GTG firms and the average firm in the S&P 500. Only one of the GTG firms, Walgreens, outperformed the S&P 500 during that period. Thus we found little evidence of the "greatness" described in GTG.[6]

We also tested the return performance of each firm against the S&P 500 during the time period beginning 15 years after its transition point through the end of 2005. This results in a longer time period being tested than in Panel 1, with an average of 14 years per firm. The results of Panel 2 in Table 3 also fail to support the generalizability of Collins' findings. In nine of the 11 cases, we failed to reject the null hypothesis. Thus we concluded with even greater certainty that there was little difference in the performance of the GTG firms and the performance of the average firm in the S&P 500. In other words, we found no evidence that the GTG firms were associated with greatness in the time period following that used by Collins to select the firms.

[6] We also tested the 20-year time period from 1986 through 2005. Despite strong bias toward rejecting the null due to a 10-year overlap with GTG's time period, we failed to reject the null in six of the 11 cases.

Table 2
2006 Fortune 500 Ranked by 10-Year Return to Investors 1995–2005[1]

Rank by Return	2006 Fortune 500 Ranking	Top 10 Fortune 500 Firms	Shareholder Return % 1995–2005
1	132	NVR	53.0
2	225	Dell	39.4
3	10	Jabil Circuit	38.7
4	448	Best Buy	37.7
5	335	Frontier Oil	37.6
6	131	Ryland Group	36.8
7	123	Whole Foods Market	36.6
8	96	MDC Holdings	33.8
9	181	Qualcomm	32.5
10	330	Kinder Morgan Energy	31.6
		Average 10-Year Return	**37.8%**
		***Good to Great* Firms**	
62	45	Walgreens	20.2
98	46	Wells Fargo	17.1
122	20	Philip Morris (now Altria)	15.2
223	177	Nucor	10.3
235	226	Circuit City	9.8
240	93	Abbott Laboratories	9.4
247	394	Pitney Bowes	9.1
280[2]	not ranked	Fannie Mae	7.6
283[3]	not ranked	Gillette	7.5
289	21	Kroger	7.3
302	140	Kimberly-Clark	6.2
202		**Average**	**10.9%**

[1] Return period begins January 2, 1996, and ends December 30, 2005. This is the same time period used by *Forbes* for the firms on the 2006 Fortune 500 list.
[2] Although Fannie Mae did not make the Fortune 500 list in 2006, its return and position on the list, had it made the Fortune 500, is provided.
[3] Gillette returns only through 2004, when it was purchased by Procter & Gamble.

Weaknesses of This Study

As with any statistical study, there are several reasons the results of our study could be inadequate. For example, it could be that the 11 GTG firms abandoned the GTG principles during the time period we examined. If this were somehow true, the average performance of these firms could be consistent with Collins' claims. Given the power of inertia in a firm, the probability of this occurring for 10 of the 11 GTG firms (excepting Walgreens) seems low. It should be pointed out, however, that any failure on our part to provide adequate results does not suggest that Collins' methodology is correct. Our objective is to test his claims, and while we may have failed to adequately test those claims, his method is still definitely flawed.

Conclusion

We have provided both statistical arguments and empirical evidence regarding Collins' claim that adopting the GTG concepts leads to "sustained great results." Empirical evidence provides no support for this claim. When ranked with the 2006 Fortune 500 companies based on

Table 3
Tests of *Good to Great* Firm Returns Against the S&P 500

Panel 1: 1996 through 2005		
Firm	Paired Test Statistic	p-Value
Abbott Laboratories	0.427	0.335
Altria (formerly Philip Morris)	1.261	0.105
Circuit City Stores	1.098	0.137
Federal National Mortgage Association	0.672	0.251
Gillette*	0.436	0.332
Kimberly-Clark	0.532	0.298
Kroger	0.457	0.324
Nucor	0.567	0.286
Pitney Bowes	0.651	0.270
Walgreens	1.703	**0.045**
Wells Fargo & Co.	1.554	0.061

Panel 2: Transition Year + 15 years through 2005				
Firm	Transition Year	Years Tested	Paired Test Statistic	p-Value
Abbott Laboratories	1974	16	1.055	0.146
Altria (formerly Philip Morris)	1964	26	2.605	**0.005**
Circuit City Stores	1982	8	1.220	0.113
Federal National Mortgage Assn.	1984	6	0.273	0.393
Gillette*	1980	9	0.324	0.373
Kimberly-Clark	1972	18	1.269	0.103
Kroger	1973	17	1.066	0.144
Nucor	1975	15	1.449	0.074
Pitney Bowes	1973	17	0.729	0.233
Walgreens	1975	15	1.930	**0.028**
Wells Fargo & Co.	1983	7	1.109	0.135

Table 3 shows returns for the 11 GTG firms compared to the S&P 500. The null hypothesis of equal returns is rejected if the individual stock returns are significantly higher than the S&P 500 returns over the time period examined. If the null hypothesis is rejected, we can conclude the average stock return is higher than the average S&P 500 return. If the null hypothesis is not rejected, there is not a significant difference in the returns. Significant p-values are bold.

* Gillette returns only through 2004, when it was purchased by Procter & Gamble.

return to investors from 1995 through 2005, the 11 GTG firms ranked an average of 202nd, and produced returns less than one third of those produced by the top 10 ranking companies. Paired t-tests of the return performance of the 11 GTG firms to the S&P 500 index also suggest that there is little difference between the 11 GTG firms and the average firm in the S&P 500.

Our explanation for this is that the method used to select the great firms in GTG is a classic example of data mining. This error was further compounded by mistaking association for causation. Thus the management principles "discovered" in GTG do not appear to be "timeless and universal answers that can be applied by any organization" (p. 5). Rather, there is no evidence that they can be generalized to other firms or to other time periods, and the methodology and evidence in GTG do not justify claims that the five principles cause greatness. Although the 11 GTG firms are certainly good, we find no evidence that they are great.

References

Collins, J. (2001). *Good to great*. New York: HarperCollins.

Collins J., & Porras, J. (1994). *Built to last*. New York: Harper Business.

Drucker, P. (1966). *The effective executive*. New York: Harper Collins.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician*, *52*, 112–119.

Smith, A. (1776). *The wealth of nations*. London: Methuen and Co., Ltd.

Walker, G. (2006). Good to great: Why some companies make the leap—and others don't. [Review of the book *Good to great*]. *Academy of Management Perspectives*, *20*(1), 120–122.