Research

Financial Analysts Journal | A Publication of CFA Institute
https://doi.org/10.1080/0015198X.2019.1567194

What Is Quality?

Jason Hsu, Vitali Kalesnik, and Engin Kose

Jason Hsu is chairman and CIO at Rayliant Global Advisors, Hong Kong. Vitali Kalesnik is a partner and the director of research for Europe at Research Affiliates Global Advisors (Europe), Limited, in London. Engin Kose is vice president and senior analyst at Allianz Global Investors, San Diego.

Unlike standard factors, such as value, momentum, and size, "quality" lacks a commonly accepted definition. Practitioners, however, are increasingly gravitating to this style factor. They define quality to be various signals or combinations of signals—some that have been thoroughly explored in the academic literature and others that have received limited attention. Among a comprehensive group of the quality categories used by practitioners, we find that profitability, accounting quality, payout/dilution, and investment tend to be associated with a return premium whereas capital structure, earnings stability, and growth in profitability show little evidence of a premium. Profitability and investment-related characteristics tend to capture most of the quality return premium.

Disclosure: The authors report no conflicts of interest.

CE Credits: 1

"Quality" as a factor in equity investing is a collection of metrics designed to capture the indicators of higher-quality financials in companies. Quality metrics are popular in the practitioner investment community, but no standard definition for the quality factor has been agreed on. In contrast, factors such as value and size have clear and accepted definitions. Although an extensive literature is dedicated to a few specific facets of quality, certain facets used in practitioner definitions have been only minimally explored in the academic literature.

As with the conventional factors, such as value and size, quality has been widely adopted as a target for factor indexes. In the 2010s, MSCI, FTSE Russell, Standard & Poor's, Research Affiliates, EDHEC, and Deutsche Bank, among others, have created smart beta indexes based on some quality factor. Moreover, they typically include quality as an element of their multifactor offerings. In conversations with investors, the quality factor is pitched by index providers as an independent source of return and as a source of diversification because of its supposedly low correlation with the value factor.

The challenge for researchers is that the quality factor is constructed differently from other factors. The value and low-beta factors, for example, are created from a particular stock characteristic (or a set of highly related stock characteristics) to capture a risk premium associated with an undiversifiable economic risk or to capture an anomalous return associated with a persistent investor behavioral bias. For example, the value factor is generally ascribed to stocks that have a high book-to-price ratio, high earnings-to-price ratio, high dividend-to-price ratio, or some combination of these three valuation measures. The portfolio resulting from construction based on one or more of these definitions owns low-valuation stocks.

In contrast, quality factor portfolios are constructed differently by the various providers. One provider might tag a stock as high quality if it has a high score on some combination of the following attributes: earnings growth, earnings growth stability, low return volatility, high profitability, high return on assets, low debt ratio, and accruals-related accounting quality. Because the quality label is vague, we assess in this article each of the quality definitions proposed by practitioners to determine which, if any, is a reliable source of return.

We begin by examining the definitions of quality implemented in various product offerings. We then examine various quality portfolios available in the marketplace to assess the risk of data mining and biases.[1] Based on the criteria, we consider what are the reliable sources of return premiums.

Survey of Quality Metrics in Product Offerings

Several major index providers offer quality indexes for passive investing. In **Table 1**, we list the company characteristics used to construct six quality factor indexes offered by six providers. We consider these characteristics to be a means to compare the indexes. For example, the quality indexes that use gross profitability, ROE (return on equity), or ROA (return on assets) are seeking to proxy company profitability, whereas indexes that use debt-to-equity and debt-to-cash-flow ratios are seeking to proxy a corporation's financial conservatism in its capital structure.

We can group the characteristics into seven categories used by product providers to define quality:

- Profitability
- Earnings stability
- Capital structure
- Growth
- Accounting quality
- Payout/dilution
- Investment

The six quality product providers listed in Table 1 use substantially different characteristics in their portfolio construction. For example, a highly profitable company does not necessarily have stable earnings or low leverage or exhibit fast growth. An examination of the existing literature does not find any research exploring how growth and accounting quality, combined with low debt, would capture a risk exposure or a persistent irrational unwillingness on

Table 1. Popular Quality Factor Index Definitions

Index	Measures Defining Quality	Corresponding Broader Quality Category
Index 1	Return on equity (ROE)	Profitability
	Debt to equity (D/E)	Capital structure
	Growth variability in earnings per share (EPS)	Earnings stability
Index 2	EPS growth	Growth
	Growth in dividends per share (DPS)	Growth
	EPS stability	Earnings stability
	DPS stability	Earnings stability
Index 3	Return on assets (ROA)	Profitability
	Change in asset turnover	Growth
	Debt to cash flow	Capital structure
	Accruals	Accounting quality
Index 4	Return on invested capital (ROIC)	Profitability
	Accruals	Accounting quality
Index 5	Gross profitability (GP)	Profitability
	Growth in total assets	Investment
Index 6	Multiple variables	Profitability
		Growth
		Safety
		Payout

the part of investors to own this desirable combination of company attributes. Nor could we find any work in the academic literature that claims these groups of variables might proxy for a common source of covariation.

To empirically study whether quality variables are homogeneous or heterogeneous, we examined the pairwise correlation of the excess returns produced by quality portfolios. The correlations reported in **Table 2** reveal a lack of similarity, indicating that the variables are not proxies for a common hidden factor.[2] The suggestion is that these leading quality index products provide a collection of heterogeneous attributes linked by the theme of financial and accounting quality. No evidence exists that these variables proxy for a unique homogeneous source of risk or a single anomaly. Therefore, quality indexes are more appropriately interpreted as multifactor portfolios whose primary commonality is that they are constructed mostly from the less well-known and less vetted company characteristics.

The other common thread—one that seems driven more by marketing than by theory or data—is that all of the selected quality characteristics are viewed as being attractive company attributes, those characteristics investors would generally be willing to "pay up" for. Implicit in the product design is an assumption that the high-growth and high-profitability companies with low debt and conservative accounting practices are underpriced and thus will generate high returns. This assumption should raise alarms for the economists among us. It's not just a free lunch; it's a free feast!

Taking to heart the caution against data-mining bias in multisignal research offered by Harvey, Liu, and Zhu (2016; hereafter, HLZ) and Novy-Marx (2016), we explored each of the seven categories of quality characteristics to determine whether a portfolio of stocks based on the desirable characteristics could indeed generate meaningful excess returns.[3] The framework we used for validating factor robustness is the three-step procedure of Hsu, Kalesnik, and Viswanathan (2015). Instead of calculating a single hard number, such as an upward-adjusted *t*-statistic, which can often feel blunt and is still potentially gameable, this method offers a suite of qualitative and quantitative diagnostics to help inform investors as to the validity of a particular factor strategy. Specifically, a return premium is more likely to be "real" if

1. it has been sufficiently explored in peer-reviewed publications,
2. its statistical significance remains robust to variations in time period and geography, and
3. its statistical significance remains robust to reasonable perturbations in definitions.

For the categories that passed these three tests, we further checked to see whether they satisfied the criteria suggested by HLZ and McLean and Pontiff (2016). The goal was to determine which, if any, of the popular quality attributes are true sources of long-term return. This knowledge would provide guidance to investors on which blend of the popular quality attributes should produce the best outcome. The three-step procedure was applied by Beck, Hsu, Kalesnik, and Kostka (2016) in their exploration of the robustness of various factors—specifically, the robustness of the gross profitability characteristic proposed by Novy-Marx (2013). Our study subsumes that result because we explored the broader category of profitability as well as six additional characteristics.

Three-Step Factor Validation Procedure

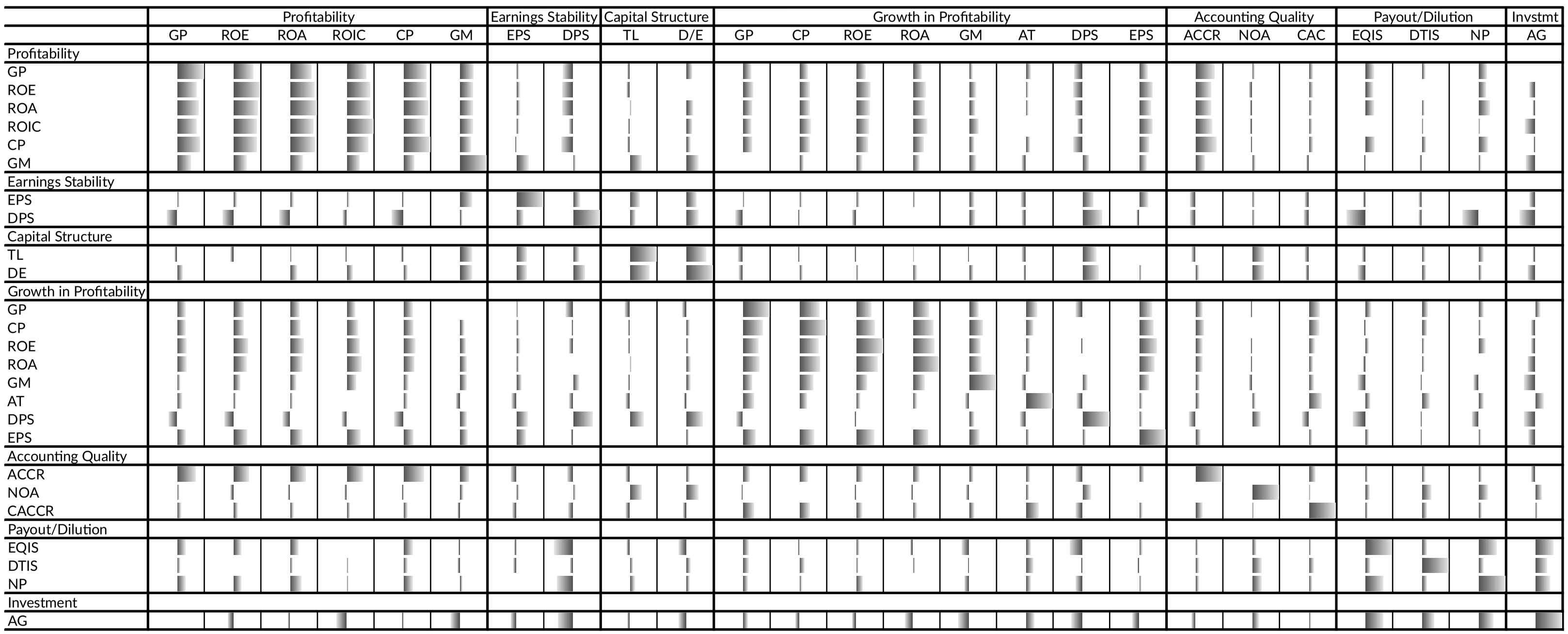
In this section, we describe the results of applying the three criteria to study the validity of a signal used in a quality factor, as proposed by Hsu et al. (2015). The first criterion is the degree to which a factor is explored in the finance literature.

Literature Review.

A thorough exploration in the literature of a source of excess return ensures that many highly trained economists have examined its merits. Furthermore, thorough coverage helps rule out the possibility that the published findings were driven by coding or data error.[4] In this section, we summarize the literature on each of the seven categories of characteristics commonly used in constructing products designed to exploit the quality factor.

Profitability. Profitability might be the most commonly used characteristic in the construction of quality portfolios. It is included in five of the six quality indexes we examined. As of the writing of this article, at least seven top-tier academic articles have studied profitability. Fama and French (2006, 2008, 2015, 2016); Novy-Marx (2013); Hou, Xue, and Zhang (2015); and Ball, Gerakos, Linnainmaa, and Nikolaev (2015) all found a positive premium associated with the profitability characteristic. Specifically, they found that the more profitable companies earn

Table 2. Correlation of Variables Used by Index Providers in Various Quality Categories



Notes: ACCR = accruals; AG = asset growth; AT = asset turnover; CACCR = change in accruals; CP = cash flow profitability; DPS = dividends per share; DTIS = debt issuance; EQIS = equity issuance; GM = gross margins; NOA = net operating assets; NP = net payout; and TL = total leverage. Table A3 in Appendix A of the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194) provides a detailed view of the pairwise correlations of a larger list of variables.

Sources: US data from CRSP and Compustat; other data from Worldscope and Datastream.

an excess return vis-à-vis less profitable companies and that portfolios based on company profitability have negative correlations with value portfolios.[5] The million-dollar question is, Why do investors fail to recognize this characteristic and consequently bid up the price of the more profitable company, thus increasing its valuation and decreasing its return?

Q theory argues that, in equilibrium, higher profitability must imply greater risk and thus a higher cost of capital.[6] Novy-Marx (2013), among others, argued from a mispricing perspective that investors underreact to high profitability because of its complexity relative to other ratios. His argument is substantiated by the observation that the less manipulated proxies for corporate profitability—excluding highly managed accounting variables such as depreciation, amortization, and other noncash variables—tend to perform better in terms of future company profitability and that investors tend to underreact to "good" attributes, which require the meticulous removal of the "polluting" items.[7]

Earnings stability. Dichev and Tang (2009) found that earnings growth volatility contains information about both short-term and long-term earnings growth. Donelson and Resutek (2015) found that earnings uncertainty is correlated with an overly optimistic expectation about earnings growth. Connecting these observations to asset pricing, Hsu, Kudoh, and Yamada (2013) found that low earnings growth volatility and the associated analyst and investor optimism are related to the low-beta effect. This finding suggests that earnings growth stability might be more appropriately categorized as a variant of the low-beta characteristic than considered as a distinct factor characteristic.

Capital structure. Empirical findings on the relationship between corporate leverage and expected equity returns are, unfortunately, mixed. Bhandari (1988) and Fama and French (1992) documented a strong and positive relationship between leverage, when computed from market prices for corporate bonds, and returns. Fama and French (1992); Penman, Richardson, and Tuna (2007); George and Hwang (2010); and Gomes and Schmid (2010) showed that after market leverage is controlled for, book leverage is negatively related to stock returns.[8]

The negative relationship between book leverage and return in the cross-section is likely a result of the cross-sectional relationship between volatility and book leverage. Companies with high book leverage also tend to have high volatility and high betas. The documented low-beta anomaly, then, suggests low returns for companies with high book leverage. Again, in this case, book leverage might be more appropriately classified as a variant of the low-beta characteristic.

Growth in earnings. We were unable to identify any papers that explored the relationship between return and past earnings growth.

Accounting quality. Managers can make decisions that affect financial reporting and temporarily boost earnings. One way to boost current earnings is to aggressively book sales that may never translate into actual cash flows. Implicit in a corporate manager's choice to incur meaningful costs and risks to manipulate earnings is the assumption that investors can be fooled, even if only temporarily. Sloan (1996); Hirshleifer, Hou, Teoh, and Zhang (2004); Dechow and Ge (2006); and Chan, Jegadeesh, and Lakonishok (2006) documented that companies with high accruals tend to have low subsequent returns. Hirshleifer et al. attributed this relationship to market participants' tendency to focus on headline earnings while ignoring indications of manipulation of those earnings.

Payout/dilution. Extensive academic research has investigated payout and issuance anomalies. We grouped these measures into a single channel because a company's payout and issuance policies are inherently tied together. Some forms of payout, such as repurchases, can be viewed as negative issuance.

Boudoukh, Michaely, Richardson, and Roberts (2007) showed that various measures of payout contain information on future stock returns. The companies that pay out more have higher subsequent returns. Both payout (dividends plus repurchases) and net payout (dividends plus repurchases minus equity issuance) predict higher stock returns in the cross-section of equities. Moreover, these return premiums cannot be explained by standard risk factors.

Loughran and Ritter (1995) documented that most forms of share issuance lead to underperformance. This finding held true for both initial public offerings and secondary issuances. Debt issuance creates a similar effect on subsequent returns. Lee and Loughran (1998) documented poor stock and operating performance in the years following convertible bond offerings. Spiess and Affleck-Graves (1999) showed that share prices of debt issuers significantly underperform those of nonissuers. This effect tended to be strongest for small, young, and NASDAQ-listed companies. Pontiff and Woodgate (2008) found that the share issuance premium is stronger than the size, book-to-market, and momentum premiums. Finally, in dissecting a number of financial anomalies, Fama and French (2008) found that the anomalous (negative) returns associated with net stock issues are robust. All these studies found a robust negative relationship between issuance and stock returns.

Investment. Titman, Wei, and Xie (2004) and Cooper, Gulen, and Schill (2008) found that companies with a conservative level of investment (asset growth) tend to achieve superior returns. Fama and French (2008, 2016) confirmed their finding.

Using *q* theory, Hou et al. (2015) argued that companies that can finance a high level of investment must be deploying capital into safer projects, which tend to have less upside. These companies and their projects are simply less risky and thus produce lower returns. Alternatively, Roll (1986) asserted that companies that invest aggressively and produce lackluster outcomes are overinvesting out of CEO hubris or are engaging in empire building because of misalignment of managerial incentives.

In summary, our literature research indicates that profitability, investment (asset growth), accounting quality, and payout/dilution are all strongly related to future return. The supporting analysis validates historical cross-sectional patterns and provides credible models that explain the phenomenon.[9] Characteristics such as low book leverage and low volatility in earnings growth appear to be too closely related to the low-volatility (low-beta) characteristic to warrant independent consideration. The past earnings growth characteristic has failed to find empirical or theoretical support in the mainstream finance literature.

We turn now to the second and third steps in our methodology—testing factor robustness by using data from various non-US regions to estimate the factor premium (the second step) and by perturbing the definitions for constructing the factor portfolio (the third step).

Robustness across Geographies and Definitions. Most of the empirical research on the cross-section of equities uses US data, which extend back to the early 1960s.[10] The long history is often necessary to establish the significance of a candidate return factor. If a candidate factor earned a premium in the United States but no other markets—worse yet, if it earned a negative premium in other markets—that finding casts doubt on the validity of the factor as a reliable source of excess return. Indeed, it suggests that the US results are probably spurious and data-mined. The reason is that a behavioral bias or source of risk is unlikely to be unique to the US market and not present in other markets that have less informed retail trading or less complete systems for risk sharing. In our study, we examined factor performance in five regions: the United States, global developed markets, Japan, Europe, and Asia Pacific excluding Japan.[11]

For companies in our US tests, we used CRSP for stock returns and market capitalization and used Compustat for company accounting information. For companies in the international tests, we used Datastream and Worldscope for, respectively, stock returns and accounting information. We excluded companies with negative book values but did not restrict our sample to observations for which all necessary data items for each quality measure were available. Instead, we used all available data for each measure.

Construction methodologies that do not show a strong in-sample *t*-statistic are never published or proposed for product launches. Thus, a natural upward bias exists in the *t*-statistic of the published and commercialized factors. To combat this bias, the three-step procedure perturbs the candidate methodology to examine the impact on the resulting *t*-statistics. A sign of data mining could be a large change in the performance of a factor as a result of a small change in the methodology. To explain why the book-to-market ratio, say, should perform very differently from trailing earnings or from the cash-flow-to-market ratio would be difficult. We certainly would not expect the sign of the estimated premium to change from one definition of value to a nearly equivalent definition of value.

To perform the robustness tests for each category, we selected three to eight definitions. We followed two guidelines in our selections. First, we included measures that were used in the index product we had selected for the examination or that were popular in the practitioner literature. For instance, because ROE and ROA are both popular definitions of profitability and are used in many quality index products in the global marketplace, we included them as perturbations of the profitability construct. Second, we included diverse measures that were, nonetheless, highly correlated with the portfolio constituents they generated.

The perturbed variable definitions for the seven categories are listed in **Table 3**. The list includes all variables we found in the index definitions as well as the more common variables used in the literature to capture these seven categories.[12]

Measurement of each quality-related variable depended on specific accounting rules. A company's assets and equities may be overstated or understated depending on transactions, such as investing in intangibles or inadequate asset write-offs. Some companies operate in highly profitable industries, and others in low-margin/high-sales environments. Financial leverage for a bank is a completely different animal from financial leverage for a utility company. In other words, quality definitions are likely to contain large industry-specific components. These industry-specific features may have obscured our robustness checks. Therefore, we used industry-neutral factors in our main analyses. For completeness, we provide standard (no industry neutrality) versions of our tests in Appendix A of the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194).

All of our portfolios were rebalanced at the beginning of July of each year. The market capitalization was measured at the end of June of the same year. Financials were lagged to have at least six months

Table 3. Categories of Quality Factor Definitions and Alternative Definitions for Each Category

Growth in Profitability	*Payout/Dilution*
Long-term change in gross profitability	Equity issuance
Long-term change in cash flow profitability	Debt issuance
Long-term change in ROE	Total payout
Long-term change in ROA	Net payout
Long-term change in gross margins	
Short-term change in asset turnover	*Investment*
Year-over-year change in DPS	Low asset growth
Year-over-year change in EPS	Low book growth
	Low capital expenditure growth
Accounting Quality	Low fixed assets growth
Accruals	
Accruals2	*Capital Structure*
Net operating assets	Total leverage
Short-term change in accruals	D/E
	Financial leverage
Profitability	
Operating profitability	*Earnings Stability*
Gross profitability	Stability of EPS growth
ROE	Stability of DPS growth
ROA	Stability of gross profitability
ROIC	Stability of cash flow profitability
Cash flow profitability	Stability of gross margins
Gross margins	

Note: Accruals2 is the difference between net income and cash flows from operations scaled by total assets. Accruals follows Sloan (1996).

between the fiscal year-end and the portfolio formation date. We augmented the Fama–French methodology with straightforward sector neutralization to form portfolios. Industries were defined according to the Fama–French 12-industry specification.[13] For each industry, we first broke the universe of stocks into large-size and small-size groups. In the US market, "large" was defined as larger than the median stock by market capitalization in the NYSE sample for each industry. For the international companies, "large" was defined as the largest 90% by market capitalization within an industry; all other stocks were put in the small-size portfolio.

Based on each variable in Table 3, we selected a high-quality and a low-quality portfolio. The high-quality portfolio contained stocks within each industry having the characteristic aligned with high quality, as defined, and the low-quality portfolio consisted of stocks identified as low quality based on that same characteristic. For example, companies with high ROA are high-quality companies, so having a high ROA is a marker for high quality. Companies with high total leverage are low-quality companies, which implies that low leverage is a marker for high quality.

For both the large and small stock groups, we selected the top (high) 30% and bottom (low) 30% of stocks with respect to the definition of the particular variable (or quality measure). This process produced four groups of stocks (high and low within the large and small universes), which we weighted proportionally by market capitalization to form four portfolios. Finally, we equally weighted the two portfolios with the high- (low-) quality characteristic from the large and small groups to form the high- (low-) quality portfolio.

For the standard (no industry neutrality) version of our test, portfolios were sorted across the sample. Size and factor characteristic breakpoints were determined on the basis of all stocks—not within industries, as was the case for the industry-neutral portfolios.

In our robustness tests for the high- and low-quality portfolios, we examined three measures of performance:

1. *Average portfolio return difference.* We tested whether the high-quality portfolio outperformed, with statistical significance, the low-quality portfolio. The practical importance of this test for investors is that statistical significance indicates that a portfolio based on the definition is likely to outperform the benchmark on a stand-alone basis.
2. *Average Fama–French plus momentum four-factor model alphas.* A candidate factor may not lead directly to excess return. If it is sufficiently negatively correlated with other common factors, however, it could deliver strong diversification benefits in a multifactor portfolio.[14] The effect would be improved information and Sharpe ratios for the portfolio.
3. *Sharpe ratio.* Using bootstrapping, we tested whether the Sharpe ratio of the high-quality portfolio was significantly higher than that of the low-quality portfolio. We sampled monthly returns (with replacement) of high- and low-characteristic portfolios to create a bootstrapped distribution of Sharpe ratios for hypothesis testing. We then used significance test statistics to determine whether each factor provided improved risk and/or return characteristics. The practical importance of this test is that it captures not only the performance but also the risk of the factor. Of course, the risk of a portfolio can easily be adjusted by varying leverage. Many investors, however, are severely limited by their governance structure in their ability to use shorting, leverage, or derivatives. For these investors, built-in portfolio risk reduction may be as valuable a feature of the factor as improved performance.

The purpose of our robustness tests was to reduce potential data-mining bias and noise in each individual measure. We present the summary information for each category in **Table 4**. The full results of the robustness tests for the variables within the seven quality factor categories across the five regions are reported in Table S1 in the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194).

In Table 4, we report the percentage of statistically significant measures—using differences in return, multifactor alpha, and the Sharpe ratio—in each category for each region. We used these summary results to determine which categories are robust in each region. The frequency of the statistically significant outcomes for each measure is a function of

- the true Sharpe ratio, information ratio, or Sharpe ratio improvement of the underlying factor (unobservable),

Table 4. Summary of Robustness of Quality Categories: Percentage of Tests Statistically Significant at the 5% Level

	Average Return	4-Factor Alpha	Sharpe Ratio	Average
Profitability				
United States	43%	100%	86%	76%
Europe	86	86	100	90
Asia Pacific ex Japan	0	29	29	19
Japan	0	29	29	19
Global developed markets	100	100	100	100
Average of significance	*46%*	*69%*	*69%*	*61%*
Earnings stability				
United States	0%	40%	60%	33%
Europe	20	40	20	27
Asia Pacific ex Japan	20	0	0	7
Japan	20	20	40	27
Global developed markets	20	40	20	27
Average of significance	*16%*	*28%*	*28%*	*24%*
Capital structure				
United States	0%	33%	0%	11%
Europe	0	33	33	22
Asia Pacific ex Japan	0	0	0	0
Japan	0	0	0	0
Global developed markets	0	33	33	22
Average of significance	*0%*	*20%*	*13%*	*11%*
Growth in profitability				
United States	38%	50%	63%	50%
Europe	38	38	50	42
Asia Pacific ex Japan	13	13	13	13
Japan	0	0	0	0
Global developed markets	25	38	38	33
Average of significance	*23%*	*28%*	*33%*	*28%*

	Average Return	4-Factor Alpha	Sharpe Ratio	Average
Accounting quality				
United States	100%	75%	100%	92%
Europe	50	50	50	50
Asia Pacific ex Japan	0	50	50	33
Japan	0	0	25	8
Global developed markets	100	75	100	92
Average of significance	*50%*	*50%*	*65%*	*55%*
Payout/dilution				
United States	75%	100%	100%	92%
Europe	75	75	100	83
Asia Pacific ex Japan	50	50	50	50
Japan	0	0	0	0
Global developed markets	50	75	100	75
Average of significance	*50%*	*60%*	*70%*	*60%*
Investment				
United States	100%	75%	100%	92%
Europe	75	0	75	50
Asia Pacific ex Japan	0	50	75	42
Japan	0	0	0	0
Global developed markets	75	25	100	67
Average of significance	*50%*	*30%*	*70%*	*50%*

Sources: CRSP/Compustat; Worldscope; Datastream.

- the length of the sample period,
- the number of multiple tests, both in the selection of quality categories and within the category variable selection, leading to selection bias (unobservable), and
- the correlation of the returns for various variables within the categories and the correlation of the test outcomes (unobservable).

Given the noisy nature of the tests and the limited time period of the sample, we should expect less than 100% even for true factors because of false negative findings. Conversely, even in the case of no underlying driver of returns and no selection bias, we should expect 2.5% of the sample to register statistical significance as a result of false positive findings (i.e., incorrectly indicating that a particular attribute is present). The multiple tests and the potential positive correlation across variables both increase the likelihood of false positive discoveries.

For the categories in Table 4, the table shows average frequencies far exceeding 2.5%, but concluding that all the categories are robust would be premature. A selection bias probably exists whereby only variables with strong in-sample performance are chosen by product providers in creating their products. This selection bias contaminates both the creation of a "factor category" and the "factor construction" within a category. Perturbing factor definitions and verifying factor performance in multiple regions reduces the within-category selection bias. It does not address the issue of bias on the category selection level. Applying a haircut to in-sample factor performance would be prudent.

To adjust for the multiple-testing bias at the category level, we used an adjustment based on the Holm statistic proposed by Harvey and Liu (2015).[15] They provided estimates of the appropriate haircuts for various levels of "realized" Sharpe ratios in portfolio strategies to compensate for multiple testing. Specifically, with 10 multiple tests, Sharpe ratios below ~0.35 should be fully adjusted to zero. Applying their Sharpe ratio adjustment to our frequency statistics reported in Table 4 results in a cut-off of 70% for the 54-year sample and a cut-off of 40% for the 27-year sample. (We provide the details for obtaining these estimates in Appendix B of the online supplemental material, available at www.tandfonline.com/doi/suppl/10.1080/00151 98X.2019.1567194.) Frequencies below these levels are probably driven by false positive outcomes.

Statistical Significance of the Quality Categories.

Next, we examine each of the individual quality categories.

Profitability. The first measure, return difference, shows that profitability produces, as shown in Table 4, a statistically significant return advantage in the global developed markets and Europe. In the other regions, the return difference is mostly in the correct direction but lacks statistical significance. When measured by differences in multifactor alpha, profitability is significant in the US market, the global developed markets, and Europe. When measured by differences in Sharpe ratio, profitability offers statistically significant improvements in three of the five regions; the two showing no improvement are Japan and Asia Pacific ex Japan. Overall, 46% of the profitability factors are significant for differences in return, 69% are significant for differences in multifactor alpha, and 69% are significant for differences in Sharpe ratio.

Earnings stability, capital structure, and growth in profitability. The factors earnings stability, capital structure, and growth in profitability almost universally show frequency levels below the cut-off suggested by the Holm statistic of Harvey and Liu (2015). Little evidence supports these factors as delivering outperformance, whether they are considered alone, in a multifactor setting, or on a risk-adjusted basis.[16]

Accounting quality. The accounting quality factors, overall, are significant in the US and other global developed markets, Europe, and Asia Pacific ex Japan but, based on return differences, are not significant in Asia Pacific ex Japan. Overall, 50% of the accounting quality factors are significant as measured by differences in return and in multifactor alpha, and 65% are significant as measured by differences in the Sharpe ratio.

Payout/dilution. The payout/dilution factors are significant in all regions except Japan. Overall, 50% of these factors are significant by return spread, 60% are significant by multifactor alpha, and 70% are significant by Sharpe ratio.

Investment. Investment factors are significant in all regions except Japan by most measures. They are not significant in Asia Pacific ex Japan by the return difference measure. Investment as measured by differences in multifactor alpha is also insignificant except in the United States. That result arises mainly because of the high correlation between the low investment factor and the value factor in our multifactor pricing model.[17] Overall, 50% of investment variables are significant for differences in return, 30% are significant for differences in multifactor alpha, and 70% are significant for differences in the Sharpe ratio.

Summary of robustness tests. Profitability seems to bring robust benefits, but mainly on a risk-adjusted basis or a multifactor basis. This outcome also holds true for accounting quality, payout/dilution, and investment, except that investment has a weaker multifactor alpha because it is correlated with the value factor. Earnings stability, capital structure, and growth in profitability have no empirical support as factors that produce a benefit for investors.[18]

The empirical evidence matches quite well with our findings from a survey of the literature, which indicated that the heaviest research has been carried out on profitability. We observed an adequate amount of research on accounting quality, payout/dilution, and investment, but little noncontradictory research on the other three categories—an indication of their lack of robustness. Based on our literature search, we would label profitability, accounting quality, payout/dilution, and investment as robust and would label earnings stability, capital structure, and growth in profitability as nonrobust.

Table 5 contains the index definitions from Table 1 and an additional column showing the degree of robustness associated with each category. Most of the indexes use at least a few nonrobust measures in their definitions. Index 4 appears to combine a measure of profitability with a measure of accounting quality, and Index 5 combines profitability and investment. These two definitions of quality are probably the most robust in this roster of indexes.

Table 5. Popular Quality Factor Indexes, Definitions, and Degrees of Robustness

Index Provider	Measures Defining Quality	Corresponding Broad Quality Category	Robustness of the Broad Category
Index 1	ROE	Profitability	Robust
	D/E	Capital structure	Nonrobust
	EPS growth variability	Earnings stability	Nonrobust
Index 2	EPS growth	Growth in profitability	Nonrobust
	DPS growth	Growth in profitability	Nonrobust
	EPS stability	Earnings stability	Nonrobust
	DPS stability	Earnings stability	Nonrobust
Index 3	ROA	Profitability	Robust
	Change in asset turnover	Growth in profitability	Nonrobust
	Debt to cash flow	Capital structure	Nonrobust
	Accruals	Accounting quality	Robust
Index 4	ROIC	Profitability	Robust
	Accruals	Accounting quality	Robust
Index 5	Gross profitability	Profitability	Robust
	Growth in total assets	Investment	Robust
Index 6	Multiple variables	Profitability	Robust
		Growth	Nonrobust
		Safety	Nonrobust (potentially overlaps low-beta effect)
		Payout	Robust

Our results for the factors constructed without sector neutrality are reported in Appendix A, Tables A1–A4 of the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194) and are similar to our main results.

Additional Test and Parsimonious Factor Models

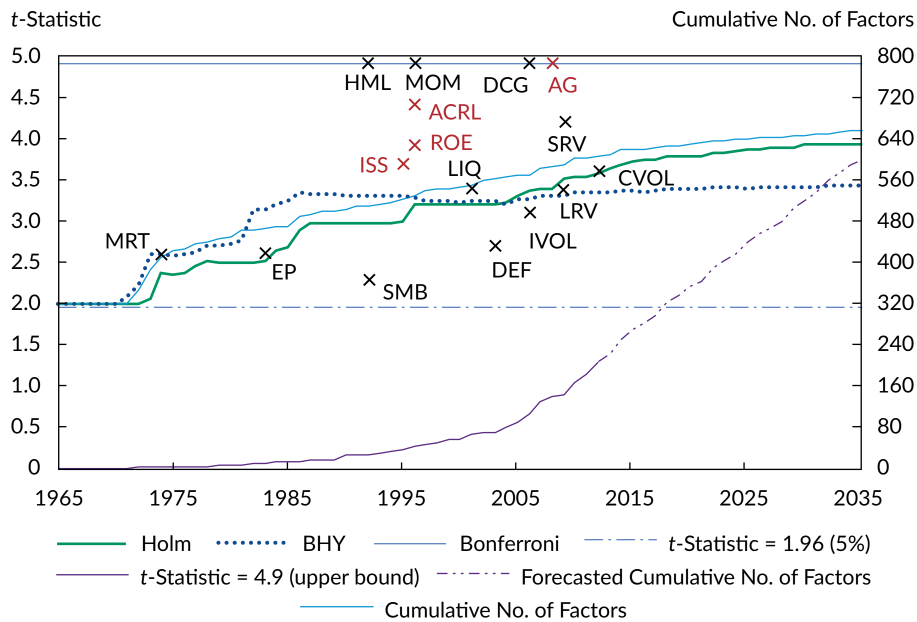
The three-step procedure we used is one way to test factor robustness. HLZ and McLean and Pontiff (2016) studied factor robustness from alternative perspectives.

HLZ argued that cut-off levels for *t*-statistics of prospective factor returns should be higher than 1.96, the usually adopted level. Given that only factors with statistical significance are published in the finance literature, we may merely be observing the positive outliers from millions of random candidate factors. In the presence of multiple tests, the *t*-statistic of 1.96 no longer corresponds to the *p*-value of 5%. A higher *t*-statistic hurdle for factor robustness tests is required; the correct hurdle would depend on the estimated total number of research backtests ever attempted. HLZ estimated this hurdle over time.

Following this procedure, we examined whether the *t*-statistic at the time of the first factor publication satisfies the more stringent criteria. We conducted this exercise for the four factors we identified as being robust—profitability, accounting quality, payout/dilution, and investment.

Figure 1 provides the corresponding factor *t*-statistics at the time of publication (as indicated with a cross), as well as the *t*-statistic cut-off level that corresponds to a 5% confidence level, *given the multiple tests* conducted in the search for factors. The *t*-statistics were computed by three different methods: BHY, Bonferroni, and Holm.[19] Note that all four robust factors (marked in red) are well above the suggested benchmarks, which indicates that they are unlikely to have been discovered purely because of a multiple search bias.[20]

Figure 1. Profitability, Investment, Accruals, and Issuance Factors



Note: ACRL = accruals; AG = asset growth; CVOL = consumption volatility; DCG = durable consumption goods; DEF = default likelihood; EP = earnings-to-price ratio; ISS = net issuance; IVOL = idiosyncratic volatility; LIQ = liquidity; LRV = long-run volatility; MRT = market beta; and SRV = short-run volatility.

Sources: Based on information in HLZ. Issuance is from Loughran and Ritter (1995), accruals is from Sloan (1996), profitability (ROE) is from Haugen and Baker (1996), and asset growth is from Cooper et al. (2008). Issuance and profitability are based on reported *t*-statistics for Fama–Macbeth (1973) regression slopes. Accruals are based on reported *t*-statistics for Jensen's alpha. Asset growth is based on reported *t*-statistics for the long–short decile portfolio return spread.

McLean and Pontiff (2016) proposed an alternative procedure. The authors investigated the postpublication performance of an extensive list of cross-sectional stock return factors and found factor performance that was significantly weaker out of the original sample. Overall, they interpreted these results from the perspective of an investor learning about these anomalies from publications and then investing in them. Higher demand for these mispricing opportunities led to price corrections and subsequently lowered the profitability of the corresponding investment strategies.

We report in **Table 6** average returns and *t*-statistics for the four robust factors before and after publication. Factor definitions were selected to reflect the variables used in the original publications that identified the factor. Table 6 statistics demonstrate that the ROE (profitability), accruals (accounting quality), and net issuance (payout/dillution) factors have similar average monthly returns in the pre- and postpublication subsamples. Asset growth has lower returns in the postpublication subsample; note, however, that the publication date for the asset growth factor is 2008, so the postpublication time period is short. Consequently, the postpublication averages and *t*-statistics for asset growth are potentially noisy estimates.

Parsimonious Quality Definitions.

Four quality categories exhibit robustness. Such a small group is appealing because academics and practitioners alike prefer parsimonious lists of factors. For example, in the 1980s, a number of articles documented that earnings to price, book to price, and similar ratios of fundamentals to price are associated with better performance. Fama and French (1992) showed that many of these drivers of return can be succinctly summarized in their three-factor model, which subsequently became ubiquitous.

The usual procedure for selecting a parsimonious list of factors follows the Gibbons, Ross, and Shanken (1989) methodology, which allows for the comparison of various models on the basis of how well they explain both the cross-section of average stock returns and stock return covariation. Fama and French (2008) showed that their three-factor model could explain many of the nonvalue and nonsize anomalies discovered by that date. In 2008, the anomalies corresponding to the four groups that we consider robust were largely unexplained by the Fama–French three-factor model (and the four-factor model that includes momentum).

Motivated by *q* theory, Hou et al. (2015) built a factor model using the profitability and investment factors together with market and size. They showed that this factor model explains many of the other anomalies better than the Fama–French three-factor model.

Stambaugh and Yuan (2017) chose a somewhat different approach. Starting with the broad list of anomalies and applying a clustering method, they reduced the list to two factors, which they intuitively

Table 6. Average Returns and *t*-Statistics for Factors, Pre- and Postpublication in US Markets

Sample	1996 High ROE	1996 Low Accruals	1995 Low Net Issuance	2008 Low Asset Growth
Full sample				
Return	0.14%	0.17%	0.30%	0.27%
t-Statistic	1.60	3.24	4.55	3.92
Prepublication				
Return	0.10%	0.18%	0.29%	0.30%
t-Statistic	1.09	2.70	4.10	3.97
Postpublication				
Return	0.21%	0.15%	0.32%	0.07%
t-Statistic	1.17	1.79	2.49	0.48

Source: CRSP/Compustat.

identified as management and performance related. The management-related factor includes accounting quality, issuance, and investment measures. The performance-related factor includes profitability, defined similarly to the category we have examined here.

Finally, Fama and French (2015) added operating profitability and asset growth factors to their three-factor model (Fama and French 1992). Fama and French (2016) showed that this five-factor model explains several anomalies, such as low beta, low volatility, and large share issues, but does not explain momentum or an accruals-related anomaly.

A survey of the literature finds that among the robust anomalies, most models considered in academia include profitability and investment as separate factors. Although both Stambaugh and Yuan (2017) and Fama and French (2016) found that issuance is related to the investment factor, they came to different conclusions about the accounting quality factor by following two different approaches: Stambaugh and Yuan found that the accounting quality factor belongs with the investment and issuance factors, whereas Fama and French demonstrated accounting quality to be a stand-alone anomaly probably related to mispricing because it does not help explain the covariation of returns as well as the other factors do.

Conclusion

So, what is quality? Quality is an industry term that refers to various company characteristics perceived to be associated with financial indicators of a company's business success. Exposure to quality is viewed as a means of generating superior return.

Quality as a category, although popular in practitioner circles, lacks a widely accepted definition. Categories of variables used in quality product offerings represent heterogeneous groups of signals and do not proxy for a unique source of risk or a single anomaly. The multisignal nature of quality products creates the potential for data mining; thus, the statistical significance of the resulting portfolios' outperformance may be overstated.

Based on the second and third steps of the three-step procedure we used to test the robustness of the various categories applied by "quality" indexes to define quality, we made the following observations:

1. Profitability as a factor delivers superior performance on a risk-adjusted or multifactor basis.
2. Accounting quality delivers overall superior performance.
3. Earnings stability, capital structure, and growth in profitability exhibit no robust evidence of being factors that lead to superior performance.
4. Payout/dilution delivers overall superior performance.
5. Investment delivers superior performance when measured by return spread and Sharpe ratio but is weak when measured by multifactor alpha.

Among these factors, profitability, accounting quality, payout/dilution, and investment are examined in many academic studies whereas earnings stability, capital structure, and growth in profitability are not well researched. Little evidence exists that earnings stability, capital structure, and growth in profitability are associated with superior performance.

The combination of the profitability and investment signals is the most parsimonious quality definition and is associated with the strongest academic evidence. Adding accounting quality and payout/dilution to those signals produces a definition that also has solid evidence of providing a historical equity premium.

All the metrics we found to be robust have a governance angle, which could be of particular interest to ESG-minded investors.[21] Specifically, the combination of strong profitability with a conservative level of investment can be interpreted as a sign of strong positive governance, which would protect against overexpansion driven by managerial hubris and excess risk taking caused by a misalignment of incentives. High accounting quality can be viewed through the ESG lens as indicative of a culture of compliance, transparency, and integrity in business reporting. Low dilution can be interpreted as good stewardship on behalf of equity shareholders.

Editor's Note:

Submitted 11 July 2018

Accepted 25 October 2018 by Stephen J. Brown

Notes

1. Previous literature on factor robustness aimed to counteract potential data-snooping and reporting biases; see Leamer (1978); Ross (1989); Lo and MacKinlay (1990); Fama (1991); Schwert (2003); and McLean and Pontiff (2016). Conrad, Cooper, and Kaul (2003) argued that most of the return spreads based on company characteristics originate from data snooping. Lewellen, Nagel, and Shanken (2010) showed the spurious nature of factors and presented biases in cross-sectional regression studies. Other studies that have contributed to the data-snooping concerns in predictive regressions are Foster, Smith, and Whaley (1997); Cooper and Gulen (2006); and Lynch and Vital-Ahuja (2012). Another line of research focuses on multiple-testing biases and their applications. It includes work by Shanken (1990); Sullivan, Timmermann, and White (1999, 2001); Ferson and Harvey (1999); White (2000); Boudoukh, Michaely, Richardson, and Roberts (2007); and Patton and Timmermann (2010). More recently, Harvey and Liu (2014, 2015); Bailey and López de Prado (2014); Hsu, Kalesnik, and Viswanathan (2015); McLean and Pontiff (2016); and Harvey, Liu, and Zhu (2016) investigated various facets of selection bias, which arises mainly because researchers publish only factors that have strong backtests.

2. Table A3 in the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194) contains a detailed view of the pairwise correlations of a larger list of variables. Additional variables are specified later in the article.

3. Commercially available quality index simulations invariably show outperformance, at least over the relatively short time periods for which they are provided. As we pointed out earlier, the quality factor is not uniformly defined. The multiple definitions of quality create the opportunity (and perhaps the temptation) to data-mine in search of the best possible outcome, thus making a simple comparison with index performance an unreliable estimate of what an investor could expect on a forward-looking basis.

4. A surprising number of published results cannot be replicated; see Bailey, Borwein, López de Prado, and Zhu (2014) for details.

5. Because growth companies tend to have more profitable investment opportunities, a high-profitability portfolio also tends to be growth oriented and thus negatively correlated with value portfolios, which tend to focus on mature, cash-cow-like companies. Additionally, a high-profitability portfolio does not depend on valuation ratios; thus, the resulting portfolio does not tautologically concentrate in high-price companies, which are known to underperform. Put another way, a high-profitability portfolio displays strong earnings growth and has a Fama–French (1992) three-factor alpha.

6. In Tobin's *q* theory of investment behavior, *q* represents the ratio of the market value of a company's existing shares (share capital) to the replacement cost of the company's physical assets (thus, replacement cost of the share capital).

7. This insight has motivated the research into variants of profitability, including gross versus net versus operating profitability.

8. This finding suggests, therefore, that the market price of corporate debt contains information on equity prices: Potentially positive cash flow information contained in corporate debt creditworthiness is not fully reflected in equity share prices.

9. Stambaugh and Yuan (2017) built a four-factor model to explain cross-sectional patterns in stock returns. Together with the market and size factors, they identified management and performance factors. Performance included profitability, and management included accounting quality, payout/dilution, and investment, thereby capturing corporate decisions such as accounting disclosure and financing.

10. Most tests in the literature are limited to the 1963 starting date because the Compustat data containing company characteristics are unavailable (on a survivorship-free basis) for previous times. Linnainmaa and Roberts (2017) is a notable exception.

11. We followed Fama and French (2016) in the choice of regions.

12. Note that the definitions used in existing product design probably suffer from upward bias for the reasons stated in the text.

13. See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_12_ind_port.html.

14. Beck et al. (2016) identified five non-quality-related factors that are well explored in the literature (three-step procedure, step 1): value, momentum, size, illiquidity, and low beta. We found value and momentum to be robust across definitions and geographies (three-step procedure, steps 2 and 3). Size was, interestingly, not robust across definitions or geographies. Beck et al. presented mixed evidence for illiquidity, which we found to be robust across definitions but not across geographies. The low-beta factor was found to be robust overall when only the Sharpe ratio of the factor return was considered, but the low-beta factor premium was statistically insignificant, even though it was economically large. The reasons for this outcome are the large negative correlation of the low-beta factor with the equity market (market factor) and its association with low absolute risk.

15. The Holm statistic method was proposed in Holm (1979) and further extended among others in Hochberg (1988).

16. Kose (2017) documented that the relationship between corporate leverage and returns is highly dependent on the maturity structure of company liabilities.

17. Goto, Goyal, Hsu, Kalesnik, and Kose (2017) showed that the correlation between value and investment is driven mostly by the factors' similar sector bets. Gerakos and Linnainmaa (2018) showed that changing the profitability

factor characteristic from operating profitability to cash profitability also keeps the high-minus-low factor (the value premium) from being redundant.

18. The tests we present in this article were conducted in the sample of nonmissing values for each individual variable. Given the number of variables used, the samples for individual tests may be quite different and possibly raise concerns about the results' sensitivity to potential outliers, as pointed out by Adams, Hayunga, and Mansi (forthcoming) and Guthrie, Sokolowsky, and Wan (2012), among others. We repeated our tests using the US universe with the best coverage for all the variables (outside the United States, limiting the coverage to the intersection of all variables available may significantly shrink the sample) and report results in Table A4 in Appendix A of the online supplemental material (available at www.tandfonline.com/doi/suppl/10.1080/0015198X.2019.1567194). As expected, the results were marginally weaker in the reduced sample because of the smaller dispersion of each variable but were still largely consistent with our main results: Profitability, investment, payout/dilution, and accounting quality largely exceed the Holm-adjusted statistics cut-off values, while the other three fail to pass the threshold.

19. For literature on the Bonferroni method, see Schweder and Spjøtvoll (1982) and Hochberg and Benjamini (1990). For the BHY method, see Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001); Sarkar (2002); and Storey (2003).

20. Figure 1 also shows *t*-statistics for value (HML), momentum (MOM), and other commonly studied factors for equity premiums.

21. ESG is environmental, social, and governance.

References

Adams, John, Darren Hayunga, and Sattar Mansi. Forthcoming. "Diseconomies of Scale in the Actively-Managed Mutual Fund Industry: What Do the Outliers in the Data Tell Us?" *Critical Finance Review.*

Bailey, David H., Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2014. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society* 61 (5): 458–71.

Bailey, David H., and Marcos López de Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management* 40 (5): 94–107.

Ball, Ray, Joseph Gerakos, Juhani Linnainmaa, and Valeri Nikolaev. 2015. "Deflating Profitability." *Journal of Financial Economics* 117 (2): 225–48.

Beck, Noah, Jason Hsu, Vitali Kalesnik, and Helge Kostka. 2016. "Will Your Factor Deliver? An Examination of Factor Robustness and Implementation Costs." *Financial Analysts Journal* 72 (5): 58–82.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57 (1): 289–300.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29 (4): 1165–88.

Bhandari, Laxmi. 1988. "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence." *Journal of Finance* 43 (2): 507–28.

Boudoukh, Jacob, Roni Michaely, Matthew Richardson, and Michael Roberts. 2007. "On the Importance of Measuring Payout Yield: Implications for Empirical Asset Pricing." *Journal of Finance* 62 (2): 877–915.

Chan, K., L.K.C. Chan, N. Jegadeesh, and J. Lakonishok. 2006. "Earnings Quality and Stock Returns." *Journal of Business* 79 (3): 1041–82.

Conrad, Jennifer, Michael Cooper, and Gautam Kaul. 2003. "Value versus Glamour." *Journal of Finance* 58 (5): 1969–95.

Cooper, Michael, and Huseyin Gulen. 2006. "Is Time-Series-Based Predictability Evident in Real Time?" *Journal of Business* 79 (3): 1263–92.

Cooper, Michael, Huseyin Gulen, and Michael Schill. 2008. "Asset Growth and the Cross-Section of Stock Returns." *Journal of Finance* 63 (4): 1609–51.

Dechow, Patricia, and Weili Ge. 2006. "The Persistence of Earnings and Cash Flows and the Role of Special Items: Implications for the Accrual Anomaly." *Review of Accounting Studies* 11 (2–3): 253–96.

Dichev, Ilia, and Wei Tang. 2009. "Earnings Volatility and Earnings Predictability." *Journal of Accounting and Economics* 47 (1–2): 160–81.

Donelson, Dain, and Robert Resutek. 2015. "The Predictive Qualities of Earnings Volatility and Earnings Uncertainty." *Review of Accounting Studies* 20 (1): 470–500.

Fama, Eugene. 1991. "Efficient Capital Markets: II." *Journal of Finance* 46 (5): 1575–617.

Fama, Eugene, and Kenneth French. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47 (2): 427–65.

——. 2006. "Profitability, Investment, and Average Returns." *Journal of Financial Economics* 82 (3): 491–518.

——. 2008. "Dissecting Anomalies." *Journal of Finance* 63 (4): 1653–78.

——. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

——. 2016. "Dissecting Anomalies with a Five-Factor Model." *Review of Financial Studies* 29 (1): 69–103.

Fama, Eugene, and James MacBeth. 1973. "Risk, Return and Equibrilium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–36.

Ferson, Wayne, and Campbell Harvey. 1999. "Conditioning Variables and the Cross-Section of Stock Returns." *Journal of Finance* 54 (4): 1325–60.

Foster, Douglas, Tom Smith, and Robert Whaley. 1997. "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2." *Journal of Finance* 52 (2): 591–607.

George, Thomas, and Chuan-Yang Hwang. 2010. "A Resolution of the Distress Risk and Leverage Puzzles in the Cross-Section of Stock Returns." *Journal of Financial Economics* 96 (1): 56–79.

Gerakos, Joseph, and Juhani T. Linnainmaa. 2018. "Decomposing Value." *Review of Financial Studies* 31 (5): 1825–54.

Gibbons, Michael, Stephen Ross, and Jay Shanken. 1989. "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57 (5): 1121–52.

Gomes, Joao, and Lukas Schmid. 2010. "Levered Returns." *Journal of Finance* 65 (2): 467–94.

Goto, Shingo, Amit Goyal, Jason Hsu, Vitali Kalesnik, and Engin Kose. 2017. "Is HML Redundant?" Research Affiliates working paper.

Guthrie, Katherine, Jan Sokolowsky, and Kam-Ming Wan. 2012. "CEO Compensation and Board Structure Revisited." *Journal of Finance* 67 (3): 1149–68.

Harvey, Campbell R., and Yan Liu. 2014. "Evaluating Trading Strategies." Special 40th Anniversary Issue, *Journal of Portfolio Management* 40 (5): 108–18.

——. 2015. "Backtesting." *Journal of Portfolio Management* 42 (1): 13–28.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "… and the Cross-Section of Expected Returns." *Review of Financial Studies* 29 (1): 5–68.

Haugen, Robert, and Nardin Baker. 1996. "Commonality in the Determinants of Expected Stock Returns." *Journal of Financial Economics* 41 (3): 401–39.

Hirshleifer, David, Kewei Hou, Siew Hong Teoh, and Yinglei Zhang. 2004. "Do Investors Overvalue Firms with Bloated Balance Sheets?" *Journal of Accounting and Economics* 38 (1): 297–331.

Hochberg, Yosef. 1988. "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 75 (4): 800–02.

Hochberg, Yosef, and Yoav Benjamini. 1990. "More Powerful Procedures for Multiple Significance Testing." *Statistics in Medicine* 9 (7): 811–18.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.

Hou, Kewei, Chen Xue, and Lu Zhang. 2015. "Digesting Anomalies: An Investment Approach." *Review of Financial Studies* 28 (3): 650–705.

Hsu, Jason, Vitali Kalesnik, and Vivek Viswanathan. 2015. "A Framework for Assessing Factors and Implementing Smart Beta Strategies." *Journal of Index Investing* 6 (1): 89–97.

Hsu, Jason, Hideaki Kudoh, and Toru Yamada. 2013. "When Sell-Side Analysts Meet High-Volatility Stocks: An Alternative Explanation for the Low-Volatility Puzzle." *Journal of Investment Management* 11 (2): 28–46.

Kose, Engin. 2017. "Dissecting Leverage on Stock Returns." Working paper.

Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Hoboken, NJ: John Wiley & Sons.

Lee, Inmoo, and Tim Loughran. 1998. "Performance Following Convertible Bond Issuance." *Journal of Corporate Finance* 4 (2): 185–207.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken. 2010. "A Skeptical Appraisal of Asset Pricing Tests." *Journal of Financial Economics* 96 (2): 175–94.

Linnainmaa, Juhani, and Michael Roberts. 2017. "The History of the Cross-Section of Stock Returns." Marshall School of Business Working Paper No. 17-17.

Lo, Andrew, and Craig MacKinlay. 1990. "Data-Snooping Biases in Tests of Asset Pricing Models." *Review of Financial Studies* 3 (3): 431–67.

Loughran, Tim, and Jay Ritter. 1995. "The New Issues Puzzle." *Journal of Finance* 50 (1): 23–51.

Lynch, Anthony, and Tania Vital-Ahuja. 2012. "Can Subsample Evidence Alleviate the Data-Snooping Problem? A Comparison to the Maximal R^2 Cutoff Test." New York University working paper (30 January).

McLean, David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance* 71 (1): 5–32.

Novy-Marx, Robert. 2013. "The Other Side of Value: The Gross Profitability Premium." *Journal of Financial Economics* 108 (1): 1–28.

——. 2016. "Testing Strategies Based on Multiple Signals." Working paper (March).

Patton, Andrew, and Allan Timmermann. 2010. "Why Do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Dispersion." *Journal of Monetary Economics* 57 (7): 803–20.

Penman, Stephen, Scott Richardson, and Irem Tuna. 2007. "The Book-to-Price Effect in Stock Returns: Accounting for Leverage." *Journal of Accounting Research* 45 (2): 427–67.

Pontiff, Jeffrey, and Artemiza Woodgate. 2008. "Share Issuance and Cross-Sectional Returns." *Journal of Finance* 63 (2): 921–45.

Roll, Richard. 1986. "The Hubris Hypothesis of Corporate Takeovers." *Journal of Business* 59 (2): 197–216.

Ross, Stephen. 1989. "Regression to the Max." Yale University working paper.

Sarkar, Sanat. 2002. "Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures." *Annals of Statistics* 30 (1): 239–57.

Schweder, T., and E. Spjøtvoll. 1982. "Plots of *P*-Values to Evaluate Many Tests Simultaneously." *Biometrika* 69 (3): 439–502.

Schwert, William. 2003. "Anomalies and Market Efficiency." In *Handbook of the Economics of Finance*, vol. 1B, edited by George Constantinides, Milton Harris, and René Stulz, 939–74. Amsterdam, Netherlands: Elsevier.

Shanken, Jay. 1990. "Intertemporal Asset Pricing: An Empirical Investigation." *Journal of Econometrics* 45 (1–2): 99–120.

Sloan, Richard. 1996. "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?" *Accounting Review* 71 (3): 289–315.

Spiess, Katherine, and John Affleck-Graves. 1999. "The Long-Run Performance of Stock Returns Following Debt Offerings." *Journal of Financial Economics* 54 (1): 45–73.

Stambaugh, Robert, and Yu Yuan. 2017. "Mispricing Factors." *Review of Financial Studies* 30 (4): 1270–315.

Storey, John D. 2003. "The Positive False Discovery Rate: A Bayesian Interpretation and the Q-Value." *Annals of Statistics* 31 (6): 2013–35.

Sullivan, Ryan, Allan Timmermann, and Halbert White. 1999. "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap." *Journal of Finance* 54 (5): 1647–91.

——. 2001. "Dangers of Data Mining: The Case of Calendar Effects in Stock Returns." *Journal of Econometrics* 105 (1): 249–86.

Titman, Sheridan, John Wei, and Feixue Xie. 2004. "Capital Investments and Stock Returns." *Journal of Financial and Quantitative Analysis* 39 (4): 677–700.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5): 1097–126.