We provide evidence that associations between realized returns and common risk proxies (factor betas and implied cost of equity estimates) vary with properties of the sample returns distribution used to estimate those associations. We consider two forms of non-random sampling from a reference distribution of realized returns for all CRSP firms with at least 12 consecutive monthly returns during 1976-2009. First, using non-random sampling based on the sign of returns, we show that risk measures are negatively (positively) associated with realized returns in bad (good) states. Second, and more generally, we focus on the effects of sample selection criteria, in the form of data requirements, that result in non-random selection from the reference distribution, using the sample for which five implied cost of equity estimates can be estimated as an example. We show that the returns in the cost of equity sample are a non-random sample from the reference distribution, and that results of association tests with realized returns are sensitive to this form of non-random sampling. We illustrate a bootstrapping approach to adjust the cost of equity sample returns distribution to match the reference returns distribution. Reversing inferences from the unadjusted sample, associations of implied cost of equity estimates with realized returns are positive in the distribution-matched samples.
1. Introduction

This paper provides evidence on how non-randomness in samples of realized returns affects the estimated associations between realized returns and common measures of risk, and proposes and illustrates a resampling technique to correct for the effects of such non-randomness.\(^1\) We define non-randomness by comparison to a definable and observable reference sample, described below, that might, but need not be the entire population. The risk measures we consider are factor betas (market, size, book-to-market and accruals quality factor betas) as well as several implied cost of equity estimates \((CofE)\) used in the accounting literature.

Our study contributes both substantively and methodologically. Substantively, we shed light on prior research showing weak or no associations between realized returns and implied cost of equity estimates.\(^2\) In contrast to this research, we find evidence of statistically reliable positive associations, once we adjust for non-randomness of a sample that is typically used in these tests. Our methodological contribution is to illustrate, using implied cost of equity measures as an example, an approach for dealing with non-random samples in empirical archival settings that arise, for example, from data availability requirements.

We start from the premise that we can define a reference sample for which the test of interest is meaningful from a theoretical perspective, and to which results from actual research samples could ideally be generalized. As our focus is on associations with realized returns, we use all CRSP firms with at least 12 consecutive monthly returns during February 1976 to July 2009 (the Full Returns sample) as the reference sample for the \(CofE\) tests. On a conceptual level, all firms in the reference sample have a cost of equity. However, we cannot observe the cost of equity, as measured by implied cost of equity metrics \((CofE)\), for observations that lack the

\(^1\) Returns are the raw holding-period returns from CRSP, and excess returns are the raw returns in Month \(t\) less the risk-free rate in Month \(t\). In cross-sectional association tests, results will be unaffected by the use of returns or excess returns; we use the terms interchangeably.

\(^2\) See Easton and Monahan, 2005; Botosan and Plumlee, 2005; Guay, Kothari and Shu, 2011. An exception is Botosan, Plumlee and Wen (2011) who find that while realized returns and implied \(CofE\) estimates are not positively associated, when they control for variables meant to proxy for the non-expected portion of realized returns, the remaining portion of realized returns is associated with several \(CofE\) estimates. However, Botosan et al. express concerns about the construct validity of this remaining portion as a proxy for expected return, since the remaining portion shows no or opposite-to-expected associations with the risk-free rate, beta and other risk proxies.
necessary data, for example, analyst forecasts, to calculate the metrics.\(^3\) Put another way, the reference sample contains complete data on one variable (returns), but not on the other variable (CofE). We hypothesize that, because of often-stringent data requirements, CofE samples are expected to differ from the reference sample; more specifically, we show that the returns of the CofE sample are not a random sample from the reference sample distribution, and that this non-randomness in returns affects results of association tests between returns and CofE measures, beyond the expected efficiency loss due to sample size reductions. We propose and illustrate a bootstrapping approach on simulated and actual data to alleviate the effects of non-randomness. We argue that our approach will increase the external validity of the CofE sample results, that is, allow for generalizations to the reference sample.

To demonstrate the influence of non-random samples and to shed light on the broader question of the influence of properties of the returns distribution, we begin by presenting results of association tests between realized returns and both factor betas and CofE measures based on an extreme but direct and intuitive selection criterion. This selection criterion partitions both the Full Returns sample and the CofE sample by the sign of the excess return, based on the premise that a positive association between realized returns and risk factors, such as factor betas or cost of equity measures, is expected in a good state, while a negative association is expected in a bad state. We use the sign of realized excess returns as a heuristic to proxy for good versus bad state.

Intuitively, when economic conditions are favorable (adverse), the realized return of a high risk firm is expected to be highly positive (negative), whereas the realized return of a low-risk firm is likely to be more moderately positive (negative). Unconditional samples contain similar, but not equal proportions of positive and negative returns observations. When these mixed-sign returns are regressed on, for example, cost of equity estimates (which are positive by both definition and construction\(^4\)), a positive association between these estimates and positive realized returns will be dampened by the influence of a negative association between

---

\(^3\) The problem is equivalent to an “item non-response” in a questionnaire that is otherwise completed. The item is known to exist, but the actual data are not available to the researcher.

\(^4\) We follow Easton and Monahan (2005) and Botosan et al. (2011) and include only those observations with positive values for all five CofE metrics in our CofE sample.
these estimates and negative realized returns.⁵ We also simulate several subsamples that give varying weight to observations with positive excess returns to document the sensitivity of association test results to non-random sampling that affects the proportions of positive versus negative returns.⁶

Consistent with this intuition, association tests using the Full Returns sample show reliably positive (at the 0.05 level or better) associations between realized returns and the market factor, size factor and accruals quality factor betas, which are interpretable as monthly factor premia. The book-to-market factor premium is negative and not significant at the 0.05 level.⁷ When we re-estimate the associations separately for positive and negative returns, we find highly significant (at the .01 level or better) associations of opposing signs. For example, regressions that condition on the sign of returns show a coefficient estimate on the market factor beta of -0.0100 (t-statistic = -9.58) for negative returns and 0.0137 (t-statistic = 5.06) for positive returns, as compared to 0.0052 (t-statistic = 2.03) in the unconditional regression.

Similarly, we analyze the relation between returns and implied cost of equity ($CofE$) estimates for, on average, 955 firms with five $CofE$ measures over 1976-2009. The unconditional correlations⁸ between returns and $CofE$ measures are reliably negative for all but the Value Line $CofE$ measure, which shows no significant relation. When we select two subsamples based on the sign of returns, we find significantly positive correlations for the positive returns sample and significantly negative correlations for the negative returns sample.

---

⁵ The same reasoning applies to tests of association between realized returns and risk factor betas. To the extent a stock with a high factor beta is risky, that stock is likely to have large absolute returns, both positive or negative. The influence of negative returns in a sample of realized returns may offset or even reverse the positive association expected for positive returns, even if the true association in the reference distribution is positive.

⁶ For example, a particular sample might contain a time period characterized by a relatively higher incidence of bad state outcomes, i.e., the 1970s. In later tests, we show that it is not the sign of the returns per se that is crucial to determining bias in the non-random $CofE$ sample, but the general shape of the returns distribution in that sample. In the more general case, a particular sample could be biased because data availability constraints limit the sample to observations that have different distributional properties from the reference distribution.

⁷ As explained in Section 4, the opposite-to-expected sign of the association between the book-to-market factor beta and realized returns is consistent with prior research that uses firm-specific data (as opposed to portfolio data).

⁸ We focus on correlation coefficients rather than univariate regression coefficients in the $CofE$ tests because correlation coefficients are unaffected by changes in standard deviations, while regression coefficients are also dependent on the relative magnitudes of the returns standard deviation and the $CofE$ standard deviation. To the extent the empirical standard deviation in $CofE$ varies across metrics, regression coefficients are not necessarily comparable, even holding the standard deviation of returns constant across our tests. Furthermore, our resampling techniques will affect standard deviations mechanically, and we aim to isolate the effects of non-randomness on correlations, rather than on standard deviations and correlations combined (as in regression coefficients).
sample. For example, the correlation for the Value Line CofE estimate is -0.1130 (t-statistic = -20.03) when returns are negative, and 0.1018 (t-statistic = 17.80) when returns are positive.9

Taken together, result of these two tests support two conclusions. First, significant positive associations between CofE metrics and returns exist for some subsamples; if we could not identify any such subsamples, we would conclude that CofE metric are pure noise or otherwise not related to returns. Second, factor premia from asset pricing tests on the Full Returns sample are also sensitive to sampling biases in returns. More generally, the sensitivity of both types of association tests to properties of the sample returns distribution suggests that results of studies that use samples with differing sample selection criteria and sample periods might not be directly comparable.

We next assess a practical and general form of non-randomness that arises because of data requirements, using both simulated data and the CofE sample to illustrate. In our simulation analyses, we simulate three types of non-random samples, compute the bias in estimates of the correlation in those samples, and show that the correlation bias is much reduced when we apply a bootstrapping (resampling) technique to produce distribution-matched samples (i.e., samples whose distributional properties match those of the reference sample).

We then apply the bootstrapping approach to the actual CofE sample. We define the Full Returns sample as the reference sample, and hypothesize that because of the data required to calculate CofE measures (e.g., analyst earnings forecasts), the CofE sample is a non-random subsample that contains larger and more stable firms with, for example, less dispersed returns than the reference sample. As a result, test results using the CofE sample would not generalize to the reference sample. In these tests, we focus on general differences in the shape of the reference distribution of returns versus the shape of the CofE sample distribution of returns.

To distinguish the effects of changes in sample sizes from the effects of non-random sampling, we first examine the implied factor premia from regressions of returns on factor betas for a random subsample of returns

---

9 We extend the sign partitioning tests to show that when we decrease or increase the likelihood of sampling from positive returns, we find that the more positive bias we induce the stronger is the estimated positive association between realized returns and risk measures, and vice versa for a negative bias.
observations from the reference sample, of the same size as the actual CofE sample; we compare these factor premia to those obtained for the actual CofE sample. The factor premia for the randomly-selected sample are similar in magnitude and statistical significance to those for the reference sample; for the actual CofE sample, however, there is no reliable association between realized returns and any factor beta. These results suggest that: (1) efficiency losses due to sample size reduction alone have little effect on qualitative results of tests of returns-risk associations; and (2) the CofE sample should not be assumed to be a random subsample of the reference sample for purposes of association tests.

To address the issue of non-random sampling in our setting, we resample the observations in the CofE sample so that the returns distribution in the adjusted (“distribution-matched”) CofE sample mimics more closely the returns distribution in the reference sample. We do this in two ways. Our first approach uses the non-parametric Kolmogorov-Smirnov (KS) test of general sample differences. The KS statistic rejects, at the 0.10 level (0.05 level) [0.01 level], the hypothesis of distribution equality in the unadjusted CofE sample compared to the reference sample in 401 (401) [393] of the 402 sample months. To distribution-match, we construct monthly resampled subsamples of CofE firms that minimize the KS statistic, and therefore minimize the deviation between the returns distributions of the CofE sample and the reference distribution, possibly reducing the size of the sample in the process. In contrast to results obtained for the actual, unadjusted CofE sample, correlations between returns and CofE measures for the distribution-matched subsamples are generally positive and highly significant. The exception is one specification with the CofE metric based on Claus and Thomas (2001), which shows an insignificantly positive association.

While the KS-based resampling procedure is effective and requires few assumptions, it imposes substantial computing burdens, especially for large samples. We therefore propose and illustrate a more practical, semi-parametric approach for distribution matching of large samples, which sorts the returns distributions of both the reference sample and the CofE sample into intervals (“bins”) first. Similar to the results found for the KS-based approach, we find that correlations between realized returns and all five CofE metrics are reliably positive in bin-based distribution-matched samples.
A possible alternative approach is to treat the CofE sample as a choice-based subsample of the reference sample, and to apply a Heckman-type selection model. That approach assumes bivariate normality of the residuals, and that the selection model itself can be estimated on a random sample of the population, which might not be possible once data on determinants are required. We illustrate and discuss the Heckman approach in Section 6.1.

We believe our findings support two inferences. First, sample selection criteria that yield samples whose returns distributions differ from those of a reference sample of returns have the potential to affect the qualitative results of association tests, including producing results that would not generalize to the reference sample. We believe many actual samples used in accounting research are likely to be characterized by non-randomness induced by sample selection criteria. Specifically, restrictive data requirements for explanatory variables are likely to affect the distribution of the outcome variable (realized returns in our setting) relative to an unrestricted, or relatively unrestricted, reference sample. To illustrate, we examine returns-risk associations after applying several selection criteria to the reference sample, including membership in the S&P 500, NYSE listing, availability of a dispersion measure of analyst earnings forecasts and share price at least $5. We find that most returns-risk associations differ depending on whether they are estimated using samples that do, or do not, meet the specified criterion. We conclude that results based on (unadjusted) non-random samples may not support generalizations to the reference sample. Our analysis of the CofE sample suggests that maximizing the size of the non-random sample, after imposing data requirements, is not a solution, and adjusting the non-random sample does improve generalizability.

The second inference speaks to the generally weak or non-existent associations between realized returns and risk proxies documented in prior research.10 These results have led some researchers to conclude that CofE

---

10 Regressions of realized returns on factor betas from asset pricing regressions also show insignificant, sometimes negative, associations (e.g., Petkova, 2006; Core, Guay and Verdi, 2008). Researchers’ interpretations of these results vary. For example, in tests of the Fama-French model, Petkova (2006, Table 5) documents a negative association between returns and the market beta, an insignificant association with the size beta, and a positive association with the book-to-market beta. Petkova does not conclude that these inconsistent associations with realized returns invalidate the model. (In subsequent tests, she finds that the Fama-French variables are correlated with innovations in variables that describe
estimates and risk factor betas are poor proxies for expected returns (or poor determinants of expected returns), that realized returns are a poor proxy for expected returns, or some combination of the two. Our inference, however, is that returns-risk associations would be expected to differ across studies that use different samples, if those samples are not randomly drawn from the population of realized returns, but instead are the direct consequence of varying data requirements. Our findings suggest that analyst-based CoFe measures have greater construct validity than previous research has shown.

The rest of the paper proceeds as follows. Section 2 describes places our work in the context of prior research. Section 3 describes our data and tests, and section 4 describes preliminary results. Section 5 applies the bootstrapping (resampling) approach we use to adjust non-random samples to “match” them with the reference sample distribution to both simulated and actual data. Section 6 discusses extensions of the analysis to other research settings and compares the bootstrapping approach to selection models. Section 7 concludes.

2. Background and Prior Research

To provide a perspective on our analysis, Section 2.1 summarizes previous research on the association between realized returns and common risk factors, specifically, market- and accounting-based factor betas and implied cost of equity metrics. Section 2.2 describes how our analyses extend, and in some ways contrast with, approaches taken in prior research.

2.1 Prior research testing for an association between realized returns and measures of risk

Accounting and finance researchers use realized returns as the dependent variable in a variety of association tests, including two-stage cross-sectional asset pricing tests (associations of realized returns with risk factor betas, Fama and MacBeth 1973) and other cost of equity tests (associations of realized returns with implied cost of equity capital metrics). Tests of these latter associations are predicated on the assumption that investment, and concludes that these findings are consistent with a discrete time version of Merton’s (1973) intertemporal capital asset pricing model.) Core et al. (2008, Table 4) also find negative, insignificant and positive associations between returns and betas on the market, size, book-to-market, and accruals quality factors. They argue that an insignificant return association with a particular factor beta invalidates that factor in an asset pricing model.
both realized returns and CofE metrics are (potentially noisy or confounded) proxies for (unobservable) expected returns. Intuitively, a firm’s expected return should be commensurate with its riskiness, with changes in the expected return caused by changes in fundamentals. Realized returns are ex-post outcome measures that might also be affected by the arrival of information during the return measurement period. In other words, realized returns consist of an expected return component and a potentially non-zero unexpected return component that is caused by news about cash flows and news about the expected return itself (Campbell and Shiller, 1988; Campbell, 1991).

Researchers typically make one of two assumptions about the relation between expected returns and realized returns. The first assumption is that realized returns are a reasonable proxy for expected returns; that is, the non-expected return component is small and/or cancels out through aggregation across firms, across time, or both. The alternative assumption is that the non-expected return component is a key, non-cancelling component of realized returns (e.g., Elton 1999 and Vuolteenaho 2002) that must be taken into account. In fact, Vuolteenaho (2002) concludes that cash flow news is the main driver of firm-specific realized returns. Elton (1999) concludes that realized returns are “a very poor measure of the expected return”, although they continue to be used in asset pricing tests without so much as a “qualifying statement” and suggests using ex ante cost of capital measures rather than realized returns.

Adopting the perspective that realized returns are a poor proxy for expected returns, researchers in accounting and finance have taken two approaches. First, they have developed and analyzed, as alternatives to realized returns, several implied cost of equity metrics (e.g., Gebhardt, Lee and Swaminathan, 2001; Claus and Thomas, 2001; Botosan and Plumlee, 2002; Easton, 2004; Brav, Lehavy and Michaely, 2005; Ohlson and Jüttner-Nauroth, 2005). The implied (or ex ante) CofE metrics are inferred from equilibrium valuation models that relate expectations about future cash flows, dividends or earnings to current price. By construction, ex-ante CofE metrics are derived from “static” equilibrium models, and, therefore, cannot be affected by “news” over a

---

11 More specifically, Lakonishok (1993) notes that 70 years of monthly returns data are needed to obtain a statistically significant market risk premium and Elton (1999) notes that there are periods exceeding 10 years during which realized stock returns are, on average, less than the risk-free rate, which makes him question the use of realized returns as a reasonable proxy for expected returns.
measurement period in the same way that realized returns might be affected. Second, some researchers have empirically purged the realized return of its non-expected (or “news”) component. For example, Campbell (1991) and Vuolteenaho (2002) propose a variance decomposition method that pre-specifies the expected returns model as a linear combination of firm characteristics; Botosan, Plumlee and Wen (2011) and Ogneva (2012) control for a specific kind of fundamental (earnings) news in realized returns in order to directly identify the cash-flow news component of realized returns. Broadly speaking, both approaches seek to increase the construct validity of the expected returns proxy by drawing inferences from the firm’s fundamentals, either directly by using a measure of expected return that is derived independently of realized returns, or indirectly by purging realized returns of non-expected returns components.

Empirical evidence of associations between asset pricing factor betas and realized returns and between implied CoFE measures and realized returns is mixed and largely disappointing. With regard to the former, Fama and French (1997) call cost of equity estimates based on either the CAPM or the Fama-French 3-factor model “woefully imprecise.” They demonstrate that, even on the industry level, it is hard to establish that risk premia are significantly positive, and they comment that the results are sure to be even worse on the firm level.

A large literature uses the statistical significance of implied factor risk premia as the criterion for the validity of asset pricing factors as proxies for priced risk. Results vary, however, across test specifications and test assets considered. In Core, Guay and Verdi (2008), for example, neither the factors in the Fama-French 3-factor model nor an accruals quality factor exhibit consistent statistically reliable implied risk premia. Results vary by specification, but in no specification is more than one of the factors significant in the expected positive direction. Using a similar design and controlling for low-priced stocks, Kim and Qi (2010) find that accruals

---

12 Related work tries to increase the association between realized returns and the respective variable of interest by filtering out an expected (as opposed to non-expected) return component. For example, Easton and Monahan (2005) and Hecht and Vuolteenaho (2006) use a variance decomposition approach to separate realized returns into expected return, cash flow news and discount rate news components. Easton and Monahan use these components to explore the weak correlation between realized returns and implied cost of capital metrics. Hecht and Vuolteenaho use the components to explore the low correlation between realized returns and contemporaneous earnings.
quality exhibits a reliably non-zero implied factor risk premium, but results on the market, size and book-to-
market factors in the Fama-French 3-factor model remain inconsistent across specifications.

Controlling for the non-expected return component of realized returns using cash flow news sometimes
improves the associations between asset pricing factor betas and realized returns. On the whole, however,
research suggests the effectiveness of the approach seems limited and inconsistent across risk factors and test
specifications. For example, Ogneva (2012) concludes that the accruals quality factor is a priced risk factor in
most specifications, but results are sensitive to test design choices. Furthermore, filtering out cash flow news
from realized returns does not qualitatively affect the remaining factor premia (in her case, the Fama-French
factors) to the same extent or even in the same direction. The market factor premium, in particular, remains
significantly negative in several specifications.

Another stream of research views realized returns and \textit{CofE} metrics as alternative, albeit imperfect proxies
for the expected return, and aims to validate \textit{CofE} metrics in association tests. Easton and Monahan (2005) and
Guay, Kothari and Shu (2011) find that the association between realized returns and several commonly used
\textit{CofE} estimates is often insignificant or significantly negative. Botosan, Plumlee and Wen (2011) find that the
association varies between positive and negative (depending on year) and is, on average, weak. Lee, So and
Wang (2011) show that the associations between implied cost of capital estimates and future returns are
statistically significant for three to four of the seven implied cost of equity estimates they consider.\footnote{Lee et al. show that the number of implied cost of capital models that yield a significant association with realized return differs depending on the test design, the horizon over which returns are measured and the source of earnings forecasts, namely, mechanical models versus analysts’ forecasts (using a mechanical model for earnings forecasting produces a larger sample and more models with significant returns associations than does using analysts’ forecasts).} Guay et al. find that modifying analyst-based \textit{CofE} metrics to account for “analyst sluggishness” improves the
associations between some \textit{CofE} proxies and realized returns. They propose two methods; in firm-specific tests
one method yields t-statistics between -0.52 and 1.93 for five implied cost of capital proxies and the other
yields t-statistics between -0.50 and 1.58.
Using variance decomposition to control for non-expected return components in realized returns, Easton and Monahan find no (a significantly negative) association between “news-purged” realized returns and three (four) of the seven implied CofE estimates they consider. In contrast, and similar to Ogneva (2012), Botosan et al. (2011) use different empirical proxies for non-expected return components and find that their CofE estimates have significantly positive associations with “news-purged” returns. However, they also document that this returns construct has either no associations or counter-intuitive associations with the risk-free rate, beta, book-to-market and other proxies for risk. These results lead them to question the validity of their “news-purged” realized returns as a proxy for expected return, and to express caution against the approach.

Based on previous research, the overall picture is that the association between returns and CofE metrics is weak or, in some studies, non-existent, and results vary across studies (that is, across samples). Moreover, adjustments to the CofE metrics (Botosan et al.) or realized returns (Easton and Monahan) do not necessarily improve this association, or not uniformly so across CofE metrics. Finally, there seems to be a tradeoff in that some adjustments to CofE metrics (notably Botosan et al. 2011) worsen the relation between CofE metrics and other proxies for risk such as beta.

2.2 Our approach to tests of association between realized returns and measures of risk.

Building on previous research, we propose a different and possibly co-existing explanation for weak and inconsistent results in tests of association between realized returns and expected return (risk) proxies (factor betas or CofE metrics). Specifically, we explore the effects of differences in the empirical distribution of realized returns in specific research samples compared to their distribution in the population (or a random sample thereof).14 Our exploration builds on the premise that findings from a sample that is non-random with respect to the construct being investigated are difficult to generalize to the population; i.e., the external validity

14 Relative to a variance decomposition approach or a news-purging approach, we require no assumption about the determinants of the expected return component, or about the functional form of the relation between news and returns. Relatedly, the measurement intervals of variables in an expected returns model does not dictate the data frequency, and disaggregated (e.g., monthly) data can be used here. In ongoing work, we evaluate more broadly the effects of aggregation of returns over time, to (compound) annual returns, for example.
of the results is questionable. We hypothesize that actual $CofE$ samples used in accounting research are such non-random samples. More specifically, we question the representativeness of the realized returns distribution in the $CofE$ sample for the overall realized returns distribution and test for the consequences of the posited unrepresentativeness for results of association tests. The $CofE$ setting provides an example in which the full distribution of returns is available for the reference sample, but data on the $CofE$ metric are missing for many of the reference sample firms.15

We conjecture that the effects of non-randomness of $CofE$ samples are likely to be both acute and variable across studies, because of the way risk manifests itself differently in a good state, when a firm performs well, and in a bad state, when a firm performs poorly. To illustrate, in a CAPM world, a high-beta firm should have larger positive returns than a low-beta firm in a good state, and worse negative returns in a bad state.16 However, an unconditional sample of realized returns will contain both positive and negative returns. In a regression of these “mixed-sign” returns on estimated CAPM betas, the positive association between positive returns and beta will be dampened by the influence of a negative association between negative returns and beta. The same reasoning applies to tests that examine the association between realized returns and implied cost of equity ($CofE$) estimates, which are positive by construction. When $CofE$ estimates are correlated with realized returns, the influence of negative returns may offset the positive coefficient that links $CofE$ measures with positive returns. Thus, even relatively small changes in the sample composition can erase or reverse the association within a specific sample.

This reasoning supports two conjectures. The first is that associations between asset pricing factor betas and realized returns (i.e., implied factor premia) and associations between implied $CofE$ measures and realized returns should be stronger when those associations are conditioned on good versus bad states, as compared to

---

15 Of course, all firms with realized returns in the reference sample have a cost of equity, but data requirements mean that the cost of equity may not be observed by the researcher. Similarly, firms not listed on CRSP, for example, private firms, also have a cost of equity capital and the fair values of their ownership interests change, but neither of these is typically observable by a researcher.

16 This reasoning is consistent with concurrent research by Bilinski and Lyssimachou (2013) who provide evidence that stocks with high CAPM betas are more likely to have very high, or very low, returns.
unconditional associations. The corresponding null hypothesis is that factor betas and CofE metrics are not associated with realized returns in good or bad states. The second conjecture is that distributional properties of the specific samples used in association tests will affect the results of those tests. For example, using a sample biased towards returns observations to the left of the mean population return can lead to the appearance of a non-existent or even opposite-to-predicted association between returns and a risk factor proxy.

To provide evidence on our first conjecture, we estimate the relation between realized returns and both risk factor betas and implied cost of equity estimates for a broad sample of returns covering a long time period (1976-2009). Mimicking an extreme and directly inducible form of sampling bias, we condition on a firm being in a good state (when the association is expected to be positive) vs. in a bad state (when the association is expected to be negative). We use the sign of excess returns as a heuristic for distinguishing between good (positive returns) and bad (negative returns) states, and as a reasonable indicator of the midpoint of realized excess returns. The approach is partially motivated by the literature on asymmetric covariances in stock returns, particularly Ang, Chen and Xing (2006) who predict and find that positive and negative returns have different implications for market betas. More broadly, their intuition is that positive and negative returns have different implications for risk, and that aggregation of returns (in their case, across time in the market beta estimation) may obscure those different implications. They calculate and examine betas separately for positive and negative returns. Our study builds on their intuition concerning the partitioning of realized excess returns by their sign; however, we do not define two different betas in the first stage of the asset pricing test, but allow for two states in the second (cross-sectional) stage of the test.

The results of our first test form the basis for our second conjecture in that these forced sample partitions, on the sign of the excess return, are a form of extreme sampling bias in the marginal distribution of excess returns. To provide evidence on the second conjecture, we first show how relatively small induced biases in

---

17 As reported in Table 1, the mean excess return in our sample is 0.0083 and the median is -0.0027.
18 In their main specification, Ang et al. condition on the sign of the difference between the current market return and the mean market return over their sample period. In sensitivity tests, they repeat their analyses using the sign of the market return itself (against zero) and using the sign of the excess market return over the risk-free rate as the state indicator, and find very similar results.
samples of returns with respect to the firm’s state of the world as compared to the population can induce, eliminate or reverse population-wide associations between risk measures and returns.19

After examining these extreme, sign-based non-random samples, we explore more generally how differences in the shape of the marginal distribution (between the sample being analyzed in a specific study and the reference sample defined by the researcher) can matter for association tests, and we provide bootstrapping techniques to remedy such non-randomness. To illustrate whether and how data requirements may result in non-random samples, we let the data requirements for five CofE metrics dictate the extent of the sampling bias in excess returns. Intuition suggests that standard data requirements for CofE metrics, such as positive earnings forecasts, positive earnings growth and analyst following, are likely to bias CofE samples towards more stable firms with less dispersion in returns than the overall returns sample; intuition does not, however, necessarily suggest an effect on associations of CofE metrics with these returns. Using the risk factor betas, we show that the CofE returns sample is indeed a nonrandom sample from the Full Returns (reference) sample.

To address the effects of using non-random samples to estimate associations between realized returns and risk factors, we develop and illustrate resampling procedures (“distribution matching”) that allow us to mimic a reference distribution of realized returns with observations from a non-random sample of that reference distribution. We validate the technique on simulated data with known correlations and across three distinct forms of non-randomness before we apply it to the (non-random) CofE sample. Our results illustrate that tests of the risk-returns association carried out using the CofE sample, adjusted by our resampling procedures, can lead to opposite conclusions compared to tests on the unadjusted CofE sample. As additional illustrations of non-random samples arising from the imposition of data availability requirements, we show that the application of several sample selection criteria used in accounting research can result in samples whose returns distributions

19 We use the word “population” to refer to broad samples of returns over long time periods. Our Full Returns sample (of the CRSP population) contains all firms with at least 12 consecutive monthly returns during the period February 1976 to July 2009. In some of our tests, we assume that this Full Returns sample can serve as the population from which other subsamples, including the CofE sample, are drawn.
are biased relative to the reference distribution. Consequently, these samples show a weaker or stronger result, or sometimes an opposite-to-expected result, with respect to the association between risk proxies and returns.

To summarize, we propose a bootstrapping technique to incorporate knowable information about a reference distribution. We illustrate our approach by revisiting previous research that uses data-restricted samples and reports weak or no associations between realized returns and both risk factor betas and CofE estimates. Our research is based on the view that changes in the distributional properties of the outcome variable (returns) may affect the ability to detect the predicted associations in a given sample. We show that the predicted associations between risk factors and realized returns are present in the Full Returns (reference) sample, except for a book-to-market factor premium, and that even small sampling biases towards more positive or more negative returns can substantively affect conclusions from these asset pricing tests. We also show that results of CofE association tests are sensitive to sample biases that arise from (non-parametric) differences between the in-sample distribution of returns as compared to the reference distribution of returns. Finally, we demonstrate the effects on associations when the CofE sample returns are resampled to mimic the reference distribution of returns.

3. Sample Data and Methods

3.1. Sample and Descriptive Data

Table 1 describes the empirical data we use in our tests. We begin by identifying all firms with monthly returns data on CRSP over the period February 1976 to July 2009 (402 months). These data are used for our asset pricing tests. The reference sample, which we sometimes refer to as the Full Returns sample, includes all firms with a return in the current month as well as in the preceding 11 months; the preceding 11 months are used to estimate univariate firm-specific risk factor betas. Effectively, a firm is required to have 12 consecutive monthly returns observations to enter the asset-pricing sample/Full Returns sample in Month \( t \). Overall, this returns sample consists of 2,460,998 firm-month observations for a total of 24,657 unique firms. Table 1,
Panel A shows that the average monthly cross section consists of 6,122 firms, and the average monthly realized raw return is 1.30%. The average median return is much smaller: 0.20% per month. The mean and median excess returns, defined as the realized return less the month-specific risk-free rate, are 0.83% (mean) and -0.27% (median). The average cross-sectional standard deviation of both raw and excess returns is 16.15%, and the interquartile ranges are about 12-13%, indicating that realized returns are quite dispersed, in particular relative to their mean. We perform all tests using excess returns, so we do not further discuss descriptive statistics of raw returns.

Panel B of Table 1 reports average cross-sectional statistics for the sample of firms for which we are able to estimate five measures of implied costs of equity (CofE). On average, those cross sections consist of 955 firms (383,955 monthly observations for 3,989 unique firms), a perfect and potentially non-random subsample from the Full Returns sample. We use the CofE sample for our cost of equity tests. Value Line cost of equity (VL CofE) estimates are derived from Value Line target prices and dividend forecasts, are re-calculated each month, and are de-annualized to the month level. We calculate four other implied CofE estimates based on the models in Claus and Thomas (2001, CT), Gebhardt, Lee and Swaminathan (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN). The firms in the CofE sample are expected to differ from those in the Full Returns/reference sample because of the firm characteristics associated with data requirements to calculate CofE measures. Specifically, all five CofE metrics require analyst following in general, and Value Line coverage in particular, as well as positive and increasing earnings forecasts. Prior literature (e.g., Francis et al. 2004) emphasizes that Value Line firms are large, and comprise about 50% of total market capitalization.

As reported in Panel B of Table 1, the mean (median) values of the CofE estimates range from 0.0071 to 0.0121 (0.0067 to 0.0118). The mean and median realized excess returns are 0.74% and 0.41% per month. With regard to higher moments of the returns distribution, the Full Returns sample is more dispersed, more positively skewed and more leptokurtic than the CofE sample. With regard to dispersion, the standard deviation

\[ \text{standard deviation} \]

\[ = \left( \frac{1}{12} \right) \text{annual CofE} - 1. \]

---

20 Because all tests are performed on the cross section, using realized returns would yield equal regression and correlation coefficients.

21 We calculate the monthly CofE as \( (1 + \text{annual CofE})^{(1/12)} - 1. \)
of excess returns for the *CofE* sample is 8.96%, a 44.5% reduction compared to the standard deviation of the Full Returns sample, and the interquartile range of excess returns of the *CofE* sample is 9.87%, a reduction of about 22% relative to the Full Returns sample. With regard to skewness, the Full Sample returns are positively skewed, with skewness coefficient of 3.74 whereas the skewness coefficient of the *CofE* sample is only .6034 (a perfectly symmetric distribution has zero skewness). Finally, with regard to kurtosis, the Full Sample returns are leptokurtic (thin-tailed), with kurtosis coefficient of 82.7 on average, while the average *CofE* sample kurtosis is 6.16 (tabulated results are expressed as excess kurtosis, over the expected level of 3 for a normal distribution).

3.2. Asset Pricing Tests

The asset-pricing tests follow a two-stage regression approach. In the first stage, we estimate slope coefficients (factor betas) in a firm-specific time-series regression of excess returns on each risk factor:

\[
R_{i,t} - R_{f,t} = a_{i,t} + b_{i,t}^F F_t + \varepsilon_{i,t} \tag{1a}
\]

where \(R_{i,t} - R_{f,t}\) is the excess return for firm \(i\) for Month \(t\);

\(F_t\) = a risk factor, specifically, the market excess return (market factor), the size factor or book-to-market factor \((SMB_t, HML_t)\) from Fama-French (1993), or the accruals quality factor \((AQfactor_t)\) from Francis et al. (2005);

\(b_{i,t}^F\) = the factor beta for risk factor \(F\);

\(t\) = subscripts the sample month.

Equation (1a) is estimated over a rolling 12-month window ending in Month \(t\). In the second stage, we estimate cross-sectional regressions of the firm-specific excess returns in Month \(t\) on the univariate first-stage factor loadings \(\hat{b}_{i,t}^F\) (that is, the risk factor betas) obtained from estimating Equation (1a):

---

22 For the time-series regressions (1a), we use the more common specification with excess returns to estimate factor betas. As all our association test results are (averages) from cross-sectional estimations, using returns or excess returns is equivalent, and we use the terms in our discussion interchangeably.
\[ R_{i,t} - R_{f,t} = \gamma_{0,t} + \gamma_{F,t} \beta_{i,t} + \theta_{i,t} \]  

(1b)

Equation (1b) is estimated for each Month \( t \). For the full sample tests, we use all firms with the necessary observations to estimate betas in the first stage. The resulting second-stage coefficient estimates (\( \gamma_{F,t} \)) are interpretable as implied risk factors in Month \( t \) (implied by the first-stage loadings). Following Fama and MacBeth (1973), the test statistic for the significance of these implied risk factors is the average monthly coefficient estimate, relative to the time-series standard error of the monthly estimates over the sample period. Theory predicts the sign (positive) but not the magnitudes of the second-stage coefficient estimates (i.e., the magnitudes of the implied factor premia). Following previous research, we test whether the \( \gamma_{F,t} \) estimates are reliably different from zero. Our unconditional or baseline tests (described in section 4.1) estimate Equation (1b) using all firm-month returns observations. In tests that condition on the sign of returns we estimate Equation (1b) separately for all positive excess returns in Month \( t \) and for all negative excess returns in Month \( t \).

3.3. Implied Cost of Equity Tests

For the implied cost of equity association tests, we estimate cross-sectional Pearson correlations between realized excess returns and each of the five CofE metrics we consider. The correlations are estimated for each Month \( t \) using all returns-CofE observations available for that month. The average of the monthly correlations over the sample period is our measure of the association between realized excess returns and a specific CofE metric. The test statistic for the significance of the correlation is the average of the monthly correlation estimates, relative to the time-series standard error of the monthly estimates over the sample period.\(^23\) We first obtain unconditional, baseline correlations using all sample returns observations. The conditional (on sign) correlations are based on separately analyzing positive versus negative excess returns in Month \( t \).

\(^{23}\) The slope coefficient from a regression of realized excess returns on CofE equals the correlation coefficient times the ratio of the standard deviation of the realized excess returns to the standard deviation of the CofE estimate. Using the average results in Panel B of Table 1, this multiplier ratio ranges from 12.9 (VL CofE) to over 25 (GLS CofE). We use the correlation coefficient instead of the regression coefficient to capture the strength of association for two reasons: First, we wish to abstract from the effects of differing standard deviations across CofE metrics. Second, our distribution matching approach might affect the standard deviations of returns and CofE metrics differently, inducing a change in the regression coefficient that is unrelated to the strength of their association. The correlation coefficient is, by definition, not affected by changes in the standard deviations.
4. Empirical Results

4.1. Unconditional results using the Full Returns sample and the CofE sample

We use the samples described in Table 1 to establish unconditional results for regression tests of association between excess returns and risk factor betas (asset pricing tests) and for correlation tests of association between excess returns and CofE measures. Those baseline results are reported in Panels A and B of Table 2, in the column labeled “Full Returns Sample” (Panel A) and “Full CofE Sample” (Panel B). Panel A shows the results of asset pricing tests; we report the second stage coefficient estimates from Equation (1b) as well as t-statistics based on the time-series standard error of the monthly estimates. These coefficient estimates capture the association between risk factor betas and realized returns, and have been interpreted as the risk premia implied by the factor betas. We find that the association between returns and the market beta is positive (coefficient estimate = 0.0052, t-statistic = 2.03), corresponding to a market risk premium of 0.52% per month. We find a similar result for the AQfactor beta (coefficient estimate = 0.0077, t-statistic = 2.44). The coefficient estimate for the SMB beta (coefficient estimate = 0.0025, t-statistic = 1.69) is significant at the .05 level, one-tailed. The association between returns and the HML beta (coefficient estimate = -0.0020, t-statistic = -1.33) is not reliably different from zero at the 0.05 level. Panel B reports unconditional correlations between realized returns and CofE estimates; as discussed more fully later, the CofE sample is a non-random subsample of the Full Returns sample. The results show either no statistically reliable relation in the case of VL CofE, or a reliably negative relation between returns and CofE estimates, with t-statistics ranging between -0.52 and -6.11. Our tests based on non-random samples of the Full Returns sample and the CofE sample, described in the next two subsections, demonstrate the sensitivity of these findings to properties of realized returns, specifically, including larger or smaller proportions of positive and negative returns.24

---

24 While prior research has mostly used annual data, we use monthly versions of the CofE estimates because the asset pricing tests reported in Table 3 are commonly performed on monthly returns. The asset pricing tests are an input to our demonstration that the CofE sample is a non-random sample from the reference sample.
4.2. Association tests between realized returns and factor betas using non-random (sign-based) samples

One premise of our prediction about the effects of non-random sampling is that the association between realized returns and risk factor betas should differ in non-random subsamples of the Full Returns sample. If associations were statistically indistinguishable from zero across all identifiable subsamples, non-random sampling would not matter. To illustrate this point, we first explore an extreme form of non-random sampling from the Full Returns sample (the reference distribution) in which we select subsamples based on the sign of returns. The columns labeled “Positive Excess Returns” and “Negative Excess Returns” in Table 2, Panel A show the results of estimating Equation (1b) separately for firms with positive excess returns in Month $t$ (“Positive”) and for firms with negative excess returns in Month $t$ (“Negative”). In the average of our 402 sample months, 51.85% of monthly excess returns are negative and 48.15% are positive.

As reported in Table 2, Panel A, with the exception of the $HML$ factor beta, the associations are positive for positive excess returns and negative for negative excess returns, and the statistical significance of all the associations is more pronounced for both the negative and positive returns sub-samples than for the estimation that uses the Full Returns sample. That is, in estimations that use a specific extreme form of non-random sampling, we find the predicted associations between realized returns and risk factor betas. For example, the market factor association is 0.0052 (t-statistic = 2.03) in the Full Returns sample, -0.0100 (t-statistic = -9.58) in the negative returns subsample and 0.0137 (t-statistic = 5.06) in the positive returns subsample. Results for the size and accruals quality factor are similar, in that the associations between the risk factor betas and realized returns are positive and significant at the .05 level or better in the Full Returns sample, negative and highly significant in the negative returns subsample and positive and highly significant in the positive returns subsample. The exception is the $HML$ factor beta, where results differ from the results for other risk factor betas in two ways. First, in the estimations that condition on the sign of returns, the t-statistics for the $HML$ risk factor are consistently smaller in magnitude than the t-statistics for the other risk factor betas. Second, the sign
of the association between realized returns and the HML factor beta is opposite the sign of the associations for the other risk factor betas.25

Our findings imply that analyses showing weak or no associations between realized returns and risk factor betas for specific samples selected after applying sample selection criteria may underestimate the actual strength of these associations, depending how the criteria affect the returns distribution as a whole.26 This is a statement about the strength of the associations, not about the equilibrium pricing consequences (i.e., the magnitude of the equilibrium risk premium). Phrased differently, we believe the results in Table 2 Panel A suggest that characteristics of the returns distribution that is used to estimate an association influence the outcome. These returns characteristics, such as whether the sample contains fewer or more positive returns, or larger or smaller returns, may in turn be influenced by researcher-chosen sample selection criteria or sample partitions that shift the distribution of returns. To examine this implication, we analyze resampled distributions characterized by increasing proportions of positive and negative realized returns. We first describe the resampling procedure and then present the results of the analyses.

Our resampling procedure is intended to shed light on how much the sampling weights of positive or negative returns must alter, relative to the respective full sample distributions, to change qualitative conclusions from association tests. We resample by month for each of 402 sample months and repeat the resampling 20 times. In each month \( t = 1 \) through 402, with \( N_t \) firms in each month, we resample with replacement \( N_t \) firms. Using these resampled data for 402 months, we repeat the association tests for all four risk factors. In the portion of Table 2 Panel A labeled “Induced Sampling Increase in Positive Excess Returns” we resample while

---

25 Prior empirical research that, like us, uses firm-specific tests as opposed to portfolio tests also finds unexpected results for the HML factor. For example, in their firm-specific tests, Core, Guay and Verdi (2008, Table 4, Panel D) document a negative (sometimes weakly significant, sometimes insignificant) relation between the HML factor beta and realized returns. In portfolio designs, the sign on the HML factor beta is generally positive in prior literature.

26 We have interpreted our results under the presumption that realized returns are a reasonable proxy for expected returns. That is, our asset pricing tests presume the construct validity of realized returns and then probe the identification of risk factors that are priced in realized returns. An alternative interpretation is possible if one starts with the assumption that a certain asset pricing model is correct; in this case, the question becomes whether a test on realized returns yields the pricing effect that should exist in the expected returns. To the extent that realized returns fail to show a significant association with risk factors, realized returns will be interpreted as an imperfect, that is, noisy and/or biased, proxy for expected returns. Under this perspective, our results can be viewed as helping to identify conditions that are informative about the construct validity of realized returns as a proxy for expected returns.
preserving the population proportions of positive and negative returns. These results, in the column labeled 0%, coincide with the Full Returns sample results; small differences result from sampling with replacement as opposed to drawing the full sample. The columns to the left, labeled -2.5%, -5%, -10% and -25%, show the full sample results when our resampling procedure decreases the portion of positive returns sampled by the specified percentages and increases the portion of negative returns sampled by the same percentages. The columns to the right, labeled +2.5%, +5%, +10% and +25%, show the results when our resampling procedure increases the portion of positive returns sampled by the specified percentages and decreases the portion of negative returns sampled by the same percentages. That is, we show the sensitivity of results of association tests to changes in the sample, using the sign of the returns as an example. The results in the Negative [Positive] Excess Returns sub-samples can be viewed as benchmark results, when the proportion of positive returns is set at 0% [100%]).

As shown in Panel A, results for asset pricing tests suggest that changing the proportion of positive returns increases the significance of results, except for the $HML$ factor. For example, the t-statistic on the implied $SMB$ factor increases from 0.38 (25% decrease in positive excess returns) to 1.67 (unbiased sample) to 2.98 (25% increase in the proportion of positive returns). Factor premia are differentially sensitive to these sampling changes, and the market factor in particular seems more robust compared to other factors. We interpret these results as suggesting that results of association tests are sensitive to the distributional properties of estimation samples, and therefore sensitive to differences in sample selection criteria, and the degree of sensitivity may differ with the nature of the selection criteria.

4.3. Association tests (correlations) between realized returns and CofE estimates using non-random (sign-based) samples

We repeat the sign-based non-random sampling analyses for the CofE estimates. For the CofE sample, the portion of negative (positive) excess returns is 47.30% (52.70%), as compared to 51.85% (48.15%) for the Full Returns sample. In contrast to results for the actual CofE sample, where all associations are either insignificant (the VL CofE metric) or reliably negative, the associations are highly significant for all CofE metrics in both the
Negative and Positive returns samples, and of opposite sign. In addition, the magnitudes of the conditional
 correlations—our measure of the strength of association—are noticeably larger than the magnitudes of the
 unconditional correlations. For example, the magnitudes of the conditional correlations between realized
 returns and the VL CofE estimate are over 30 times larger than the magnitude of the unconditional correlation;
 other differences are also striking (the smallest difference is a factor of about 2.4). When we resample to
decrease and increase the proportions of positive and negative returns, relative to the proportions in the actual
CofE sample, we find that for all CofE metrics except the VL CofE, correlations remain negative or
insignificant until the resampling increases the portion of positive returns sampled by 25% and decreases the
portion of negative returns sampled by the same percentage.

To summarize, we find that the associations between realized returns and implied costs of equity in the full
CofE sample are reliably negative for four of the five implied costs of equity we consider. Conditioning on the
sign of realized returns, we also find there are subsamples that show positive associations between CofE metrics
and realized returns. Based on these results, we conclude that the shape of the actual CofE returns distribution
could materially influence the results of association tests.

4.4. Non-randomness of the CofE sample

As previously described, we aim to shed light on how differences in the distributional properties of
estimation samples of realized returns, and, by implication, differences in sample selection criteria, affect the
results of tests of association between realized returns and both risk factor betas and implied cost of equity
estimates. In this section, we report results of analyses of the effects of sample size versus sample non-
randomness, using the Full Returns sample as the proxy for the population and the CofE sample as a potentially
non-random subsample of this population.27 As previously discussed, these two samples differ with regard to
size (a monthly average of 6,122 firms in the Full Returns sample and 955 firms in the CofE sample) and with

27 The data requirements to calculate the CofE estimates restrict this sample to firms with Value Line following, with
analyst following and with some combination of positive earnings and/or increasing earnings. Easton and Monahan (2005)
describe the CofE measures we consider.
regard to dispersion, skewness and kurtosis, with the Full Returns sample being more extreme on all three
distributional properties. We now explore the effects of sample size and non-randomness on association tests
between realized returns and risk factor betas. Because the factor betas are available for both the Full Returns
sample and the smaller CofE sample, these tests can be used to illustrate that the CofE sample differs from the
Full Returns sample with respect to the association between realized returns and risk proxies.

Turning first to the effects of sample size, Table 3 presents results of association tests between risk factor
betas and realized returns for the Full Returns sample (repeated from Table 2 Panel A); a Random Subsample of
the same size each month as the actual CofE sample (an average of 955 firms per month), but strictly randomly
drawn from the Full Returns sample; and the actual CofE sample. The Random Subsample and actual CofE
samples are the same size each month, approximately 16% of the Full Returns sample on average, but results
will differ if the actual CofE sample is not a random subsample of the Full Returns sample.

A comparison of the results for the Full Returns sample and the Random Subsample sheds light on the
effects of sample size per se. The coefficient estimates are nearly the same for the two estimations (differences
range between 1 and 4 basis points) and the t-statistics are lower by amounts between .08 (market risk
premium) and .13 (AQFactor premium). In two-sided tests, sampling error yields a size premium that is no
longer significant at the 10% level (t-statistic = 1.58, p-value = 0.1150). When we repeat the asset pricing tests
on the actual CofE sample, none of the four factor betas shows a significant association with excess returns
(results are in the rightmost column). While the book-to-market premium was insignificant in the Full Returns
sample and the size premium was significant at conventional levels in the Random Subsample in a one-sided
test, all four factor premia are now reduced by at least 50%, and no t-statistic exceeds 0.90 in magnitude.

We conclude from this analysis that even substantial reductions in sample size, by about 84% in the average
cross section in our sample period, have a relatively modest effect on the efficiency of the estimation, but that
distributional differences in either realized returns or in factor betas have substantial effects on the results. We
interpret these results as supporting the notion that the CofE sample is a non-random subsample of the Full
Returns sample for the purpose of testing associations with proxies for risk.
5. Distribution Matching on Simulated and Empirical Data

Using both simulations and the archival data of the CofE sample, in this section, we first provide further evidence that the non-randomness of the marginal distribution of one variable in our association tests (realized returns) leads to biased estimates of correlations with another variable. In the simulations of Sections 5.1 and 5.2, rather than estimating associations using actual data, we take the opposite-direction approach and induce a pre-specified correlation between two simulated variables in a population. We draw from these simulated populations both randomly and in three distinct non-random ways, whereby the selection probabilities of the non-random draws are based on the marginal distribution of one variable, and provide evidence that the correlation estimates from these non-random samples are biased. We next describe and illustrate a distribution-matching approach that resamples from a non-random sample in ways designed to adjust the non-random sample distribution to mimic a specified reference sample distribution. We validate this distribution-matching approach by showing that correlation estimates from distribution-matched samples converge to their true values, despite of consisting of only non-randomly drawn observations from the reference sample. In Section 5.3, we report the results of applying the distribution-matching approach to the actual CofE sample.

5.1. Simulation Design

Data generation and (Random and Non-random) Sampling

We simulate populations with 5,000 observations as pairs of variables \( y \) and \( x \), allowing for non-zero skewness and kurtosis in the marginal distributions. As the correlation coefficient \( \sigma_{yx} \) depends on both skewness and kurtosis, we first compute an equivalent (“intermediate”) correlation coefficient to generate preliminary, and correlated, normally distributed variables. We then transform the preliminary variables into the desired potentially non-normal variables.

More specifically, we follow these steps:

1. We choose the desired first four moments of the distributions of \( y \) and \( x \).
2. Following Fleishman (1978), we transform variables $y$ and $x$ with known skewness and kurtosis into normally distributed variables in the following way:

$$a = -c_a + b_a z_a + c_a z_a^2 + d_a z_a^3 \quad \text{with } a \in \{x, y\}$$

(2)

We determine the transformation coefficients $b_a$, $c_a$ and $d_a$ using Newton-Raphson iteration.

3. Using the three transformation coefficients ($b_a$, $c_a$ and $d_a$), we compute the intermediate correlation for two variables that are normally distributed, again using an iterative numerical procedure (following Vale and Maurelli 1983).²⁸

4. We then generate two preliminary variables $z_y$ and $z_x$ that are independent and standard-normally distributed. To introduce dependence, we multiply the $N \times 2$ matrix of $[z_y \mid z_x]$ with the $2 \times 2$ Cholesky decomposition of the intermediate correlation matrix.²⁹ To introduce non-normality, we apply the $b_a$, $c_a$ and $d_a$ transformation coefficients from Step 2 above, multiply with the desired standard deviation, and add the desired mean. The resulting variables will show the pre-specified (original) level of correlation as well as the pre-specified first four moments.

We generate 1,000 populations with either two standard normal variables (Panel A of Table 4) or two non-normally distributed variables (Panel B), whereby $y \sim (0,1,3,10)$ and $x \sim (0,1,-1,3)$, to highlight the non-parametric nature of the distribution matching approach.³⁰ We repeat the procedure for true correlations of 0.5, 0 and -0.5.

To generate the random and non-random samples, we begin by drawing a random sample from each population of size $m = 1,000$ observations. We then draw three types of non-random samples, whereby the (marginal) sample distribution of $y$ will differ from the (marginal) population distribution of $y$.

---

²⁸ For non-zero higher moments, there are generally four solutions to the polynomial in (2). We first perform a grid search over the parameter space to identify all four solutions for $b_a$, $c_a$ and $d_a$, and retain the solution with the lowest absolute distance between the intermediate correlation and the desired correlation.

²⁹ The Cholesky decomposition of the positive-definite and symmetric matrix $A$ returns the matrix $U$ that satisfies $A = U'U$.

³⁰ In the first case, the transformation coefficients simplify to $\{b_a, c_a, d_a\} = \{1,0,0\}$ and the intermediate correlation will equal the original correlation.
For the first non-random sample (‘Non-Random Sample I’ in Table 4), the selection probability of Observation $i$ is decreasing in the ranked absolute distance from the mean:

$$\text{Prob}_i = \frac{A_i}{\sum_{i=1}^n A_i} \times m \text{ where } A_i = \text{rank} \left[ \frac{1}{\text{abs}(y_i - \overline{y})} \right]$$

(4)

The motivation behind this type of non-random sampling is to select, with increasing likelihood, relatively stable observations (“firms”) with $y_i$ relatively closer to $\overline{y}$. Because we use the absolute value of the distance, neither positive nor negative values of $y_i$ are favored by the weighting scheme. Our goal is to capture the effects on the distribution of realized returns of the data requirements for calculating the five CofE measures we consider, by over-selecting large, stable firms. In terms of stability, the average cross-sectional standard deviation in the one-year-rolling beta estimates is 1.42 in the Full Returns sample versus 0.83 in the CofE sample (results not tabulated). Because there is no measure of firm size in these simulations, we proxy for an equal-weighted “market” using the cross-sectional mean, $\overline{y}$. Compared to the population or a random sample, the empirical cumulative distribution function of this non-random sample will be truncated at both extremes, and show a steeper slope in the mid-range of the distribution.

For the second type of non-random sample (‘Non-Random Sample II’), we impose an exogenously assumed marginal distribution of $y_i$, as opposed to a distribution defined using the empirical values of $y_i$ in the population. Specifically, we sample observations based on the uniform distribution over the entire population interval $[y_{min}; y_{max}]$, with higher selection probabilities for observations in the tails. We consider this distribution an example of a distribution that is unrelated to the observed empirical distribution of realized returns. Compared to the other types of non-random sampling (I and III), the high weight of the tails in the uniform distribution necessitates bootstrapping with replacement. Depending on the realizations of $y_i$ in a given population, the empirical cumulative distribution function will approximate a straight line.

The third non-random sample, ‘Non-Random Sample III’, is similar to the first non-random sample except that selection probabilities are not symmetric to the population mean. Here, selection probability is based on
the ranked distance to the maximum value of \( y_i \), such that the selection probability is strictly increasing in \( y_i \). In implementing this sampling, we replaced \( \bar{y} \) with \( y^{\text{max}} \) in Eq. (4) above. As a consequence, the empirical distribution function will be truncated at the left and lower than that of the population through most of its support, with the distance strictly decreasing with increasing values of \( y_i \).

*Distribution-Matching of Non-Random Samples*

For each of the three types of non-random samples drawn from the simulated population, we bootstrap “distribution-matched” samples. A distribution-matched sample is derived from its corresponding non-random sample, and can therefore consist only of those non-randomly drawn observations. The distribution-matched sample is constructed by bootstrapping with replacement, such that the empirical cumulative distribution of \( y_i \) in the non-random sample, \( F^{\text{NRS}}(y_i) \), mimics the empirical distribution of \( y_i \) in the population, \( F^{\text{POP}}(y_i) \). While each non-random subsample is a strict subsample of the population, values of \( y_i \) will be systematically overrepresented, underrepresented or not represented at all in the non-random sample, with the latter possibly limiting the success of the distribution matching. Put another way, distribution matching is restricted to the common support of \( F^{\text{NRS}}(y_i) \) and \( F^{\text{POP}}(y_i) \).

In these simulations, both variables are mean zero, and no observation in the population is assigned a zero selection probability. Consequently, the common support of \( F^{\text{NRS}}(y_i) \) and \( F^{\text{POP}}(y_i) \) is relatively large, there is no truncation of the distributions in the population, and truncation in a specific non-random sample is likely to be small. Furthermore, the variables are constructed from continuous distributions, and duplicate values of \( y_i \) are rare. Hence, we can apply a simplified and highly efficient approach for distribution matching to the simulated data, in which a distribution-matched sample is constructed by fitting the non-random sample probabilities to the population probabilities at each value of \( y_i \) in the non-random sample.\(^{31}\) Specifically, in these simulations, the non-random sample of \( m = 1,000 \) observations is first sorted on \( y_i \). Observation \( i = 1 \) will have \( y_i = y^{\text{min}} \),

\(^{31}\) For the application on actual returns data below, the approach follows the same principle with the following modification: the Kolmogorov-Smirnov statistic is minimized in an iterative procedure, even if insignificance in a given month cannot be achieved.
with $F_{\text{NRS}}^{\text{min}} = 1/m = 0.001$, with $F_{\text{POP}}^{\text{min}}$ likely to be different. Observation $i$ will be bootstrapped $m F_{\text{POP}}^{\text{min}}$ times to mimic the empirical cumulative distribution function in the population at $y_{\text{min}}$. More generically, observation $i$ of the non-random sample is bootstrapped $m \left( F_{\text{POP}}^{y_i} - F_{\text{POP}}^{y_{i-1}} \right)$ times. The observation $i = m$, with $y_i = y_{\text{max}}$, is drawn $m \left( 1 - F_{\text{POP}}^{y_{i-1}} \right)$ times.32

Before and after distribution matching, the difference in the empirical distributions of $y_i$ between the population and a non-random sample is assessed using the non-parametric Kolmogorov-Smirnov (KS) statistic, which computes the maximum absolute distance between two cumulative empirical distributions, in percent.

$$KS = \max_j \left| F_{\text{NRS}}^{y_j} - F_{\text{POP}}^{y_j} \right| \quad \text{where } j = 1, 2, \ldots, n$$

(5)

The KS statistic is associated with an (asymptotic) $p$-value for a test on distribution equality between the respective sample and the entire population.

The outcome variable of interest is the estimated correlation between $y_i$ and $x_i$ in the non-random samples before and after distribution matching; the population correlation is specified at three levels: 0.5, 0, and -0.5. Examining zero-correlation variables allows us to examine if distribution matching induces an apparent correlation where none exists. Examining a negative correlation as well as a positive correlation allows us to rule out that the effectiveness of distribution matching depends on the sign of the (true) association.

5.2. Simulation Results

Figure 1 shows a graphical representation of the non-random draws and the effectiveness of the distribution matching approach. The figure depicts example distributions of $y_i$ for a single simulation run, for both the unadjusted non-random samples (shown to the left) and the distribution-matched samples (shown to the right). The (constant) benchmark distribution, which appears in both the right and left graphs, is from the population in that particular run ($n = 5,000$). Numerical results of the simulations analysis are reported in Table 4. Panel A

32 Because of rounding effects, the distribution-matched sample does not necessarily reach 1,000 observations in a first pass. In those cases, the remaining observations are drawn randomly from the non-random sample.
(Panel B) contains the results for two standard-normally distributed (non-normally distributed) variables. We focus our discussion on the results in Panel A; results in Panel B are qualitatively similar and are not discussed in the text. We note from this comparison, however, that the effectiveness of our (non-parametric) distribution matching approach does not seem to depend on the shape of the marginal distributions.

We first verify that empirical correlation estimates in the population and random samples (‘CORR’) are close to the specified (true) correlations (‘CORR*’), and that there are no meaningful differences between the population and the random sample. For all three levels of true correlation, the KS statistic for random samples is about 2.4%, and p-values for the difference between the random sample and the population are about 0.69. Estimated correlations differ from true correlations by 0.005 or less, confirming that a reduction in sample size, even to 20% of the population ($m = 1,000$), is relatively unimportant for the association point estimates, as long as the sample reduction is random.

In contrast, non-random samples, by construction, have a distribution of $y_i$ that differs sharply from the population distribution. For ‘Non-random sample I’ (‘Non-random sample II’) [‘Non-random sample III’], KS statistics are about 14% (25.5%) [26.7%]. To assess the significance of the bias in correlation estimates from non-random samples, we report the percentile of the mean non-random sample correlation in the distribution spanned by the 1,000 correlations from the random samples as ‘Percentile (Random Sample)’ in the table. A small (large) percentile corresponds to a low (high) correlation estimate.

When the true correlations are 0.50 or -0.50, the estimated correlations are biased towards zero in magnitude in Non-random samples I and III and are upward biased in Non-random sample II. The bias is highly significant, with percentiles either 0.0 (i.e., below the distribution of 1,000 correlations from random samples), or 99.9 or higher (i.e., only 1 or fewer of the 1,000 correlations is higher). Distribution matching, however, is effective in mitigating the effects of non-random sampling. Across all three types of non-random samples, the corresponding distribution-matched samples show correlations that are much closer to the true correlations. Biases range from -0.0144 to 0.0169, and are insignificant in all cases, with percentiles ranging from 30.1 to 77.5.
From these results, we conclude that distribution matching is effective in reducing the bias in correlation estimates in non-random samples of $y_{ij}$, for the cases we consider, and for both positive and negative true (population) correlations. Also, the results for uncorrelated variables ($\text{CORR}^* = 0$, middle column of the table) show that distribution matching does not induce an apparent correlation where none exists, alleviating concerns about type-II errors. The result for Non-random sample I, which attempts to mimic stable observations (“firms”) that are more likely to have data on actual $\text{CofE}$ metrics in our empirical application, is of particular interest. We caveat that, because the result is based on simulated data and our imposed stability criterion in the sample construction, this finding is not dispositive. It is, however, indicative of the bias in association tests if data availability requirements bias a sample towards stable firms, as is often the case in actual research situations.

5.3. Distribution matching the actual $\text{CofE}$ sample

Results in Table 3 suggest that the realized returns of the actual $\text{CofE}$ sample are a non-random sample of the Full Returns sample returns. In addition, Table 1 shows that excess returns for the $\text{CofE}$ sample have a similar mean/median, but noticeably smaller standard deviation, skewness and kurtosis, as compared to the Full Returns sample. Combined, we interpret these findings as raising the question as to how much the result from another type of association test, specifically, the negative correlation between $\text{CofE}$ estimates and realized returns reported in Table 2, is generalizable to the reference sample.

To analyze the effect of differences in the shape of the returns distribution, we create distribution-matched $\text{CofE}$ samples using two approaches. As stated above, compared to distribution matching on simulated data, distribution matching on actual returns data faces additional difficulties arising from relatively small common support regions as well as from truncation of the distribution, both of which vary from month to month. The size of the cross sections also varies from month to month. To accommodate these factors, we modify the matching approach to make it an iterative procedure that attempts to minimize the KS-based statistic, even in months where insignificant KS statistics cannot be achieved because of large initial differences between the $\text{CofE}$ sample and Full Returns sample. Recall that the non-parametric Kolmogorov-Smirnov statistic (KS
statistic) captures any difference between two distributions, not limited to the first four moments of the distribution. At the same time, the minimization does not require us to pre-specify a sample size. This approach, while conceptually grounded, is computationally burdensome for large samples. We therefore also illustrate a second less computationally demanding approach that matches the CofE sample to the Full Returns sample using a resampling technique based on pre-determined return intervals (“bins”).

Kolmogorov-Smirnov-based distribution match.

For this approach, we start by randomly sampling 20% of the month-specific sample, or a fixed number of 100 unique firms, from the CofE sample in a given month. We compute the KS statistic\(^ {33} \) of this initial draw of firms against all returns from the reference sample in that month. The Kolmogorov-Smirnov test on this initial sample is likely to reject the null hypothesis that the Full Returns distribution is equal to the returns distribution of, for example, the 100 initially selected firms. We start our iteration to minimize the KS statistic by randomly adding one firm (Firm 101). We then re-compute the KS statistic, again versus the reference sample distribution of returns that month. If the KS statistic using 101 firms (against the reference distribution) is greater than or equal to the KS statistic using the 100 firms sample (against the reference distribution), we dismiss the 101st firm, and replace it with another randomly chosen (with replacement) 101st firm from the CofE sample. If the KS statistic using 101 firms is lower than the KS statistic using the 100 firms sample (against the reference distribution), we keep the 101st firm, draw a 102nd firm, and evaluate the inclusion of the 102nd firm. We repeat this step 1,000 times, thereby allowing for KS-based distribution-matched samples to increase by a maximum of 1,000 firms each month.\(^ {34} \) As the convergence to a minimum KS statistic will depend both on the initial 20% (or 100) observations drawn and on the order of additions to that sample, we repeat the procedure 30 times, and retain the final sample with the lowest KS statistic (the minimal difference as compared to the full returns distribution). We repeat the construction of KS-based samples for each month.

\(^{33}\) As location of the distribution has no impact on either correlation coefficients or regression coefficients, we standardize both distributions (reference and current sample distribution) to be mean 0 before the computation of the KS statistic.\(^ {34}\) The actual CofE sample contains on average 955 firms per month. With 1,000 iterations, we are effectively allowing each firm to enter the distribution-matched sample, to the extent its inclusion leads to a closer match to the returns in the reference sample.
When the iteration is initialized with 20% of the CofE sample firms, the final distribution-matched sample contains an average of 242.2 firms (about 25.4% of the 955 firms in the average CofE sample month), with a time-series standard deviation of 54 firms. When we start with a fixed number of 100 firms each month, the average cross section consists of 124 firms (with a standard deviation of 14 firms).

**Bin-based distribution match.**

As mentioned above, the iterative procedure for the KS-based matching is computationally expensive. In the interest of defining a more practical approach, we also create bin-based distribution-matched samples. To that end, we first place the returns of both the Full Returns sample and the CofE sample into bins with width of 100 basis points (bp). We calculate the sample proportion of observations in each bin, for both samples. To distribution-match, we re-weight (by resampling return-CofE pairs with replacement) each bin in the CofE sample, so that the sample proportion matches the proportion in the corresponding reference sample bin. For example, if the realized returns bin [0.10; 0.11] contains 5% of the CofE sample observations and 10% of the reference sample observations, we resample the CofE sample bin to increase its percentage to 10% of the sample size in that month.

At the extremes of the reference sample distribution, we encounter bins without corresponding observations in the CofE sample. At both the upper and lower extremes of the CofE sample, we reweight the most extreme positive and most extreme negative returns, with equal weighting at both extremes in the following form (month subscripts omitted):

---

35 The design choices in this bin-based approach are admittedly ad hoc. We are working on sensitivity checks around the chosen parameter levels.

36 This approach also sharpens both goodness-of-fit and poorness-of fit in an unbiased fashion. That is, if a given bin in the CofE sample contains realized-return/CofE pairs that fit poorly, this approach will exacerbate that poor fit when the sampling percentage increases for that bin, and vice versa if the bin contains pairs that fit well. When sampling percentages are reduced, the opposite is the case.
\[
\begin{align*}
\omega_{\text{CofE}} &= \left\{ \begin{array}{ll}
\frac{\min(i_{\text{CofE}})}{w_{\text{RS}}} \sum_{j=\min(i_{\text{RS}})}^{\min(i_{\text{CofE}})} \left[ \min(i_{\text{CofE}}) - i_{\text{RS}} + 1 \right]^{\gamma} & \forall \ i_{\text{RS}} = \min(i_{\text{CofE}}) \\
\frac{\max(i_{\text{RS}})}{w_{\text{RS}}} \sum_{j=\max(i_{\text{RS}})}^{\max(i_{\text{CofE}})} \left[ i_{\text{RS}} - \max(i_{\text{CofE}}) + 1 \right]^{\gamma} & \forall \ i_{\text{RS}} = \max(i_{\text{CofE}}) 
\end{array} \right. \\
\end{align*}
\]  

\( w_{\text{CofE}} \) is the sampling proportion of Bin \([i; i+0.01]\) in the CofE sample and the reference sample respectively. The product of sampling proportion \( w_{\text{CofE}} \) and (overall) sample size in Month \( t \) is the bin-specific number of draws that month. We then numerically solve, within the sampling procedure, for the (constant) weight parameter \( \gamma \) until we obtain a standard deviation for the distribution-matched CofE sample that is statistically indistinguishable from that of the Full Returns sample. After this calibration, the time-series average of the differences in cross-sectional standard deviations between the Full Returns sample and the distribution-matched CofE sample is -0.0008 (t-statistic = -0.56), with \( \gamma = 2.20 \). Figure 2 provides the intuition behind the approach by plotting the distribution of excess returns for the CofE sample, before and after distribution matching, as well as the reference distribution of returns.

As reported in Table 3, and in contrast to the reference sample results, implied factor premia in the unadjusted CofE sample are all insignificant. We begin our analyses by re-estimating the implied factor premia using the KS-based distribution-matched samples and the bin-based distribution matched sample (\( \gamma = 2.20 \)); results are not tabulated. In all three distribution matched samples, the implied market factor, size factor, and AQfactor premia are positive and significant at the 0.01 level or better, with t-statistics of 2.86 or higher. The book-to-market factor remains insignificant at the 0.10 level (two-sided). In sum, the results on implied factor premia in the distribution-matched samples are qualitatively similar to results in the reference sample.

The results of correlation tests between realized returns and CofE estimates for the distribution-matched samples under both the Kolmogorov-Smirnov and bin-based approaches are shown in Table 5. The average Kolmogorov-Smirnov (KS) statistic of the CofE sample is about 14%. The average \( p \)-value over 402 sample months is 0.0009, and the test rejects similarity of distributions in 401 of 402 sample months at the 0.10 level or
better. After KS-based distribution matching using 20% of firms (100 firms) initially, the average KS statistic is about 6%, with an average $p$-value of 0.5287 (0.7789), and the test rejects similarity of distributions in 42 (five) of 402 months at the 0.10 level or better. We conclude from this analysis that the KS-based approach to distribution matching is effective, although it is also computationally burdensome.

The second and third column of Table 5 report the correlations between five $CofE$ measures and realized returns, using the KS-based $CofE$ sample. For comparison, the correlations for the unadjusted $CofE$ sample are shown in the first column (repeated from Table 2 Panel B). The time-series average correlations, across 402 months, are positive and highly significant in 9 of the 10 specifications, with t-statistics between 2.23 and 6.08. The exception is the association with the CT measure with 100 firms as initial sample, for which the correlation is positive but insignificant (t-statistic = 0.73). These results indicate reliably positive associations between at least four implied costs of equity metrics and realized returns.

Unlike the KS-based approach, the bin-based approach seeks to match each bin in the $CofE$ sample’s return distribution with the corresponding bin of the Full Returns sample. Results in Table 5 show this match is effective for $\gamma = 2.15$ and $\gamma = 2.20$, where the overall difference in standard deviations is insignificant at conventional levels. At higher values of $\gamma$, the approach over-corrects as evidenced by the t-statistic of -1.69 for the test of equality of standard deviations when $\gamma = 2.25$. Because the focus is on matching standard deviations, we do not expect that the bin-based distribution match will yield an adjusted $CofE$ returns distribution that is statistically similar at other moments to the Full Returns distribution, and this is indeed the case. Results in Table 5, lower part, show differences in skewness and kurtosis that are significant at conventional levels. That said, the KS statistics for general similarity of distributions, reported in the first row of Table 5, show that the bin-based distribution match induces less dissimilarity between the adjusted $CofE$ sample and the Full Returns sample, as evidenced by reduced KS test statistics and reduced levels of significance.

Columns 3-6 of Table 5 report correlations between five $CofE$ estimates and realized returns, using the bin-based distribution-matched $CofE$ sample. In these analyses, all five implied cost of equity measures have statistically reliably positive correlations with realized returns. The magnitudes of the correlations range from
about 0.02 (for the CT measure) to approximately 0.05 (for the VL measure) and t-statistics range from 1.62 (for the CT measure, $\gamma = 2$) to over 4 (for the GLS and VL measures and larger values of $\gamma$).

We next address the concern that distribution matching the CofE sample leads an increased association between CofE metrics and excess returns at the cost of a diminished association between CofE metrics and other proxies for risk, specifically, the risk factor betas. This concern may arise, in part, from Botosan et al.’s (2011) finding that news-purged realized returns, which should measure expected returns, have either no associations or counter-intuitive associations with several risk proxies. We wish to determine if our bootstrapping approach has a similar undesired effect. We test for a decline in the associations between CofE metrics and risk factor betas using the sample composition from the KS-based matching in Table 5 with initial sample size equal to 20% of the CofE sample. As a benchmark, we draw a random sample of the same size in any given month, and regress the five CofE metrics on lagged risk factor betas from Eq. (1a). Our test is based on the time-series of the difference between the 402 month-specific KS-based sample results and the 402 month-specific results from these equal-sized random samples. We repeat the procedure 100 times, and evaluate the differences using the average Fama-MacBeth-type t-statistics across the 100 outcomes.

We find that 19 of 20 coefficient estimates (five CofE metrics times four risk factor betas) are insignificant at conventional levels, with t-statistics ranging from -0.59 to 1.46 (results untabulated). The exception is the coefficient on the market beta in the VL CofE regression, which shows a significant difference of 0.0001 ($t = 2.09$). In all cases, coefficients from the KS-based sample are numerically very similar to coefficients from the random samples. In particular, they are always of the same sign, and always significant at comparable levels.

Combined with the evidence in Table 3, we interpret the weight of the evidence in Table 5 as demonstrating that differences in the shape of the returns distribution between the Full Sample and the CofE sample have a marked effect on the results of association tests. The first inference we draw is that sample selection criteria that yield estimation samples with different returns distributions, as compared to a reference sample, may make it more difficult to detect the predicted associations between realized returns and implied CofE estimates. The second inference we draw is that adjusting the distribution of the outcome variable in the non-random sample
(in this case, realized returns) to mimic that of the reference sample provides at least a partial solution to the issues raised by non-random sampling.

6. Extensions

6.1. Relation to selection models

In this section, we clarify the relation between our distribution matching approach and a commonly-used technique for dealing with missing data, Heckman-type selection models. Both the selection model approach and the distribution matching approach aim to incorporate information into the test model that goes beyond the information in the subsample with complete data, but the two approaches differ in the kind of information that is incorporated. Distribution matching focuses on the outcome variable only, whose empirical distribution in the reference sample can be defined by the researcher or is empirically estimable. As a result, distribution matching does not require explanatory variables for a selection model, which might impose similar or even more stringent data restrictions than does the actual test model.

To be more specific, the sample restriction issue at the heart of our analysis does not arise in the approach developed in Heckman (1979) because the exogenous covariates in his first-stage selection model are attainable for all observations, or, equivalently, attainable for a random subsample of the population.\footnote{The estimation on a random subsample will suffer from a loss in efficiency, compared to the estimation in the population, but results remain unbiased (as also shown in Table 3).} Consequently, results from a Heckman model are generalizable only to the sample for which the selection model variables are available. In addition, our approach does not require an assessment of the fit of the selection model; increasing the fit of the selection model by including more explanatory variables is likely to impose increasingly stringent sample restrictions due to data requirements.

Distribution matching is by design non-parametric and based on the empirical distribution function of the outcome variable; it is not limited to normally distributed outcome variables. In contrast, the derivation of the Heckman correction for sampling biases relies on the assumption that the residuals from the selection model

\footnote{The estimation on a random subsample will suffer from a loss in efficiency, compared to the estimation in the population, but results remain unbiased (as also shown in Table 3).}
and the test model are jointly normally distributed. The normality assumption allows for a closed-form solution for the sampling bias in OLS coefficients as a function of the inverse Mills ratio, the standard deviation of the test model residual, and the correlation between test model residuals and selection model residuals. The descriptive statistics for excess returns reported in Table 1 cast doubt on this normality assumption in our setting. At a minimum, we caution that Heckman test results with realized returns as the dependent variable might be affected by violations of the normality assumption.

We implement the Heckman model subject to the constraint of avoiding, as much as possible, additional sample restrictions, at the potential cost of not maximizing the fit of the selection model. To achieve this goal, we restrict our analysis to selection models with explanatory variables that exist for every observation, or at least the vast majority of observations, in the Full Returns sample. Our selection models include some or all of the following: firm size (CRSP market capitalization at the end of the prior month), firm age (the difference, in months, between the first month on CRSP and the current month), CRSP trading volume, the book-to-market ratio from the Compustat annual file, and the four univariate risk factor betas from the asset pricing regressions (1a).38

Table 6 reports semipartial correlation coefficients between CofE metrics and excess returns and goodness-of-fit measures for the (probit) selection models we estimate. The selection model that includes only size has a Cox-Snell pseudo $R^2$ of 0.28 (0.48 when rescaled to a theoretical maximum of 1; Nagelkerke 1991), with no additional sample loss; adding more variables increases the pseudo $R^2$ to a maximum of 0.32, with a sample loss of 2.1% when the model includes the log of the book-to-market ratio.39 The reported semipartial correlations are averages of the 402 month-specific (cross-sectional) semipartial correlations, which are, in turn, obtained by

---

38 Firm age, as defined, and the factor betas are available for all observations. We use the log of all characteristics (firm age, size, volume and book-to-market). We acknowledge that CRSP does not contain volume data for NASDAQ firms prior to November 1982; therefore, the sample losses for selection models that include volume are largely due to that earlier period, while coverage afterwards is almost complete.

39 There are several accepted definitions of pseudo $R^2$. The reported figures are computed following Cox and Snell (1989), and have a theoretical maximum, even in cases of perfect fit, below 1. We also report a rescaled version of pseudo $R^2$ from the probit model that rescales the Cox-Snell estimate to a theoretical maximum of 1 as suggested by Nagelkerke (1991).
controlling for the inverse Mills ratio in the respective CofE metric first, then computing the correlation between the returns and the residualized CofE metric.

The semipartial correlations reported in Table 6, after adjusting for selection bias through the inverse Mills ratio, are all negative or, as is often the case for the VL CofE metric, not reliably different from zero. Across CofE metrics, point estimates appear slightly larger in magnitude, and are more highly statistically different from zero, as compared to those reported for the CofE sample in Table 2 Panel B. However, we again caution that the effects of the inverse Mills ratio on the semipartial correlations of interest might be due to violations of the normality assumption, an inadequate fit of the selection model or some combination of the two factors. We conclude that, in our setting, Heckman selection models do not change the conclusion from results obtained using the unadjusted CofE sample.

We also use a Heckman-type adjustment in the risk factor beta regressions for the CofE sample, as in Table 3 (results not tabulated). Specifically, we use the variables of Model IV in the selection model, and rerun the cross-sectional asset pricing tests using a factor beta and the inverse Mills ratio. Again, the effect of including the inverse Mills ratio is limited. Factor premia estimates are hardly affected (differences range from -0.0004 to -0.0001) and remain insignificant at conventional levels.

6.2. Asset pricing tests based on returns of samples that meet selection criteria used in accounting research

So far, our tests have focused on the CRSP population of firms with at least 12 consecutive monthly returns during our sample period (the Full Returns sample) and the subsample of those returns associated with firms for which the CofE measures we analyze can be calculated. We have analyzed how differences in returns properties, both within the two samples (the sign of returns) and between the two samples (the shapes of the returns distributions) affect results of association tests. In this section, we consider how results of asset pricing tests of the association between risk factor betas and realized returns are, or are not, sensitive to sample selection criteria, other than having sufficient data to calculate CofE estimates, that are likely to yield non-random samples in applied research situations.
The cross-sectional sample selection criteria we consider are membership in the S&P 500, a potential sample constraint in compensation research; NYSE listing, required when using trading data from the TAQ database; the availability of standard deviation of analysts’ earnings forecasts, required for research using analyst forecast dispersion; and stock prices equal to or greater than $5, which has been used as a control for potentially noisy and illiquid stocks, sometimes referred to as “penny stocks”. We apply each of these criteria separately to the Full Sample and re-estimate Equation (1b) for the observations meeting and not meeting the criterion separately. We also report the proportions of firms that do and do not meet each sample selection criterion and descriptive data on the distributions of excess returns (the average of the monthly cross-sectional mean excess returns, standard deviation, skewness) for the two subsamples. Results are reported in Table 7.

In general, the sample selection criteria we consider result in unequal proportions of firms in the Full Returns sample that do and do not meet each criterion. The difference in proportions is, not surprisingly, most extreme for the S&P 500 criterion (8.44% meet the criterion) and for the price-at-least-$5 criterion (75.60% meet the criterion). Also, average mean excess return, standard deviation and skewness appear to differ between the subsamples, although in the absence of statistical tests of equality this result is merely suggestive.

With this caveat, we offer the following general observations. First, the subsamples that do not meet the sample selection criteria have larger average mean excess returns (with the exception of the subsample of stocks with prices below $5 where the average mean excess return is negative), larger standard deviations of excess returns and greater positive skewness of excess returns. Given the results previously reported, this suggests that the asset pricing tests might yield results that are more consistent with theory for firms that would not be included in the sample resulting from the application of each of the selection criteria. This is the case for all four selection criteria we consider (although the differences are relatively modest for the price criterion and the

---

40 The Execucomp database covers S&P 1500 firms since 1994, but other compensation data sources can be more restrictive. See, for example, Brookman, Jandik and Rennie (2006) for an overview.
41 Both the lagged price and the current price in the realized return calculation must be at least $5 to meet the criterion.
42 In ongoing work, we intend to refine and expand the analysis in Table 7, including the sample selection criteria considered and the specific investigations of how those criteria affect both the properties of estimation samples of realized returns and the results of asset pricing tests of associations between risk factor betas and realized returns.
HML factor shows a negative or insignificant association for all subsamples). In addition, KS statistics for tests of equality of distributions between the Full Returns sample and each subsample in Table 7 show that for three of the four selection criteria considered, percentage deviations between the Full Returns sample and the subsample that meets the criterion exceed the deviations for the subsample that does not meet the criterion; the exception is the price criterion. In the context of our previous results and discussions, we draw the inference that reasonable and necessary sample selection criteria may result in samples whose realized returns distributions differ from the full sample distribution of realized returns, and to which the research results may or may not generalize.

7. Conclusions

We predict and find that associations between realized returns and popular proxies for expected returns, specifically, five implied cost of equity (CofE) estimates and risk factor betas from the Fama-French 3-factor model and an accruals quality factor, are influenced by the properties of the realized returns distribution that is used to estimate the associations. We consider how non-randomness in estimation samples of realized returns can affect inferences about whether there is in fact an association between realized returns and common measures of risk.

Our preliminary tests are predicated on the view that a positive association between realized returns and proxies for risk, such as factor betas or CofE measures, is expected in a good state, while a negative association is expected in a bad state. We proxy for good and bad states using the sign of realized returns. Because an unconditional sample of realized returns contains a mixture of positive and negative returns observations, these positive and negative associations can dampen and offset each other in unconditional broad-sample estimations. Using the sign of realized excess returns as a direct and extreme form of non-random sampling of returns, we find reliable evidence of both positive associations between realized returns and risk factors when the state of the world is “good” and negative associations between realized returns and risk factors when the state is “bad”.
We further show that small changes in the sample proportions of positive and negative returns might alter qualitative results, in particular for implied factor premia.

After confirming the existence of subsamples with differing associations between CofE metrics and realized returns, we focus more generally on how systematic differences in the overall shape of return distributions affect the results of tests of association between realized returns and measures of risk. For these tests we use the CRSP population of firms with at least 12 consecutive monthly returns during 1976-2009 as the reference sample, and compare results obtained using this sample with those obtained using the sample of firms with sufficient data to calculate five implied CofE measures. The latter sample is a substantially smaller and perfect, but potentially non-random subsample of the former.

Our results indicate that the effects of sample size reductions per se on point estimates and on standard errors (efficiency losses) in tests of association are relatively modest in our setting, provided the sample is drawn randomly. In contrast, using empirical data on implied factor premia and using simulated populations, we show that non-randomness can change qualitative results. Specifically, for the actual CofE sample, we find no statistically reliable implied factor premia, in contrast to results from the reference sample and random subsamples of the same size as the CofE sample. Using simulated data, we document reliably negative correlations in non-random (stable) subsamples, despite having induced positive correlations in the populations.

These results lead us to examine a general form of non-randomness in returns that stems from stringent data requirements, using as an example the CofE sample. We find that differences in the shape of the returns distribution in the CofE sample, as compared to the returns distribution in the reference sample, affect the results of association tests in ways that raise doubts on the external validity of conclusions based on association test results in specific samples, such as the CofE sample. When we resample observations from the CofE sample so that the resulting returns distribution mimics the returns distribution in the reference sample, we find statistically reliable positive correlations between realized returns and four of the five CofE measures we consider.
Our results suggest that sample selection criteria, including for example data availability requirements, can yield samples whose returns distributions may differ from those of a population or a reference sample and that this non-randomness can qualitatively affect the results of association tests between realized returns and both risk factor betas and \( CofE \) measures. We believe that many samples used in accounting research are likely to be characterized by intentional or unintentional non-randomness, induced by sample inclusion criteria based on non-returns variables such as exchange listing and analyst following. Applying several commonly used selection criteria, we find that tests of association between risk factor betas and realized returns are often sensitive to whether the tests are based on firms that do, or do not, meet a given selection criterion.

Overall, our analysis highlights that non-randomness of samples resulting from data requirements may lead to conclusions that do not generalize to a reference sample. We recommend assessing the impact of data requirements on the empirical distribution function of the test model variables, in particular the variable that suffers the most severe loss (in our case, returns). Furthermore, we believe that distribution matching has the potential to increase the external validity of results obtained from non-random samples and, in doing so, serve a coordination function across research studies that use varying sampling criteria. To the extent the same reference sample can be defined and reconstructed by many researchers, comparisons of results across studies are facilitated.
References


Figure 1

Panel A: Non-Random Sample I: Before (left) and (right) after distribution matching

Panel B: Non-Random Sample II: Before (left) and (right) after distribution matching

Panel C: Non-Random Sample III: Before (left) and (right) after distribution matching
Figure 1 shows the empirical cumulative distribution of $y_i$ for three types of non-random samples as described in the text and for the corresponding distribution matched samples, for one run of the results reported in Table 4, Panel A. Sample distributions are depicted with the red/lighter line and the left (right) graphs are before (after) distribution matching. The blue/darker line marks the population distribution and is constant in all six graphs.
Figure 2 shows the empirical distribution (density) of excess returns for three samples: the actual CofE sample (dashed red line), the reference sample (blue/dark bars), and the bin-based distribution-matched sample with $\gamma = 2.20$ (grey/light bars). The width of each bin is 100 basis points. Data in the figure are bin-specific average sample proportions across 402 months and are truncated at +/-50%. For purposes of this graph, distribution-matched-sample returns are taken from a single randomly chosen run of the resampling procedure.
The sample period is February 1976 to July 2009 (402 months or cross sections). The table presents average data across these 402 cross sections. The Full Returns sample (reference sample) contains on average 6,122 firms (2,460,998 firm-months), which are required to have at least 12 consecutive months of CRSP returns data. The Implied Cost of Equity (\(CofE\)) sample is a perfect subsample of the Full Returns sample, and contains an average of 955 firms each month (383,955 firm-months). Firms in this sample have sufficient data to calculate (monthly variants of) five \(CofE\) estimates based on Value Line data, denoted VL, and based on models in Claus and Thomas (2001, CT), Gebhardt, Lee and Swaminathan (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN).

### Panel A: Full Returns Sample (Asset Pricing Test Sample)

<table>
<thead>
<tr>
<th></th>
<th># Firms</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>P5</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>P95</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realized Returns</td>
<td>6,122</td>
<td>0.0130</td>
<td>0.1615</td>
<td>3.7403</td>
<td>82.7487</td>
<td>-0.7442</td>
<td>-0.1947</td>
<td>-0.0595</td>
<td>0.0020</td>
<td>0.0676</td>
<td>0.2453</td>
<td>3.2195</td>
</tr>
<tr>
<td>Excess Returns</td>
<td>6,122</td>
<td>0.0083</td>
<td>0.1615</td>
<td>3.7403</td>
<td>82.7487</td>
<td>-0.7488</td>
<td>-0.1994</td>
<td>-0.0642</td>
<td>-0.0027</td>
<td>0.0629</td>
<td>0.2406</td>
<td>3.2148</td>
</tr>
</tbody>
</table>

### Panel B: Implied Cost of Equity Sample

<table>
<thead>
<tr>
<th></th>
<th># Firms</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>P5</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>P95</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realized Returns</td>
<td>955</td>
<td>0.0121</td>
<td>0.0896</td>
<td>0.6034</td>
<td>6.1594</td>
<td>-0.3734</td>
<td>-0.1217</td>
<td>-0.0393</td>
<td>0.0087</td>
<td>0.0594</td>
<td>0.1557</td>
<td>0.5742</td>
</tr>
<tr>
<td>Excess Returns</td>
<td>955</td>
<td>0.0074</td>
<td>0.0896</td>
<td>0.6034</td>
<td>6.1594</td>
<td>-0.3781</td>
<td>-0.1263</td>
<td>-0.0439</td>
<td>0.0041</td>
<td>0.0548</td>
<td>0.1510</td>
<td>0.5696</td>
</tr>
<tr>
<td>VL CofE</td>
<td>955</td>
<td>0.0121</td>
<td>0.0070</td>
<td>0.8016</td>
<td>2.7667</td>
<td>0.0002</td>
<td>0.0032</td>
<td>0.0061</td>
<td>0.0118</td>
<td>0.0165</td>
<td>0.0236</td>
<td>0.0513</td>
</tr>
<tr>
<td>GLS CofE</td>
<td>955</td>
<td>0.0071</td>
<td>0.0036</td>
<td>5.4237</td>
<td>59.8661</td>
<td>0.0006</td>
<td>0.0034</td>
<td>0.0053</td>
<td>0.0068</td>
<td>0.0084</td>
<td>0.0110</td>
<td>0.0505</td>
</tr>
<tr>
<td>MPEG CofE</td>
<td>955</td>
<td>0.0088</td>
<td>0.0045</td>
<td>3.2930</td>
<td>27.0187</td>
<td>0.0003</td>
<td>0.0037</td>
<td>0.0060</td>
<td>0.0081</td>
<td>0.0105</td>
<td>0.0160</td>
<td>0.0538</td>
</tr>
<tr>
<td>OJN CofE</td>
<td>955</td>
<td>0.0098</td>
<td>0.0037</td>
<td>4.5876</td>
<td>46.6087</td>
<td>0.0040</td>
<td>0.0061</td>
<td>0.0077</td>
<td>0.0092</td>
<td>0.0110</td>
<td>0.0153</td>
<td>0.0526</td>
</tr>
<tr>
<td>CT CofE</td>
<td>955</td>
<td>0.0073</td>
<td>0.0049</td>
<td>4.9134</td>
<td>49.7870</td>
<td>0.0001</td>
<td>0.0020</td>
<td>0.0046</td>
<td>0.0067</td>
<td>0.0090</td>
<td>0.0133</td>
<td>0.0580</td>
</tr>
</tbody>
</table>
## Table 2
### Association Tests Estimated Using Unconditional, Conditional on Sign of Return and Biased Return Samples

**Panel A: Univariate Association between (Excess) Returns and Risk Factor Betas (Implied Factor Premia)**

<table>
<thead>
<tr>
<th></th>
<th>Full Returns Sample</th>
<th>Negative Excess Return</th>
<th>Positive Excess Return</th>
<th>Induced Sampling Increase in Positive Excess Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-25%</td>
<td>-10%</td>
<td>-5%</td>
</tr>
<tr>
<td><strong>Avg. Monthly Proportion</strong></td>
<td>n/a</td>
<td>51.85%</td>
<td>48.15%</td>
<td>23.72%</td>
</tr>
<tr>
<td>beta Market</td>
<td>0.0052</td>
<td>-0.0100</td>
<td>0.0137</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>-9.58</td>
<td>5.06</td>
<td>1.36</td>
</tr>
<tr>
<td>beta SMB</td>
<td>0.0028</td>
<td>-0.0064</td>
<td>0.0113</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>1.69</td>
<td>-10.20</td>
<td>6.77</td>
<td>0.38</td>
</tr>
<tr>
<td>beta HML</td>
<td>-0.0020</td>
<td>0.0030</td>
<td>-0.0059</td>
<td>-0.0008</td>
</tr>
<tr>
<td></td>
<td>-1.33</td>
<td>5.05</td>
<td>-3.68</td>
<td>-0.66</td>
</tr>
<tr>
<td>beta AQFactor</td>
<td>0.0077</td>
<td>-0.0133</td>
<td>0.0289</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>2.44</td>
<td>-12.44</td>
<td>9.48</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**Panel B: Correlations between (Excess) Returns and Implied Cost of Equity Estimates**

<table>
<thead>
<tr>
<th></th>
<th>Full CoE Sample</th>
<th>Negative Excess Return</th>
<th>Positive Excess Return</th>
<th>Induced Sampling Increase in Positive Excess Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-25%</td>
<td>-10%</td>
<td>-5%</td>
</tr>
<tr>
<td><strong>Avg. Monthly Proportion</strong></td>
<td>n/a</td>
<td>47.30%</td>
<td>52.70%</td>
<td>28.53%</td>
</tr>
<tr>
<td>VLI CoE</td>
<td>-0.0033</td>
<td>-0.1130</td>
<td>0.1018</td>
<td>-0.0425</td>
</tr>
<tr>
<td></td>
<td>-0.52</td>
<td>-20.03</td>
<td>17.80</td>
<td>-6.44</td>
</tr>
<tr>
<td>GLS CoE</td>
<td>-0.0096</td>
<td>-0.0707</td>
<td>0.0604</td>
<td>-0.0336</td>
</tr>
<tr>
<td></td>
<td>-2.19</td>
<td>-15.97</td>
<td>12.30</td>
<td>-7.07</td>
</tr>
<tr>
<td>MPEG CoE</td>
<td>-0.0203</td>
<td>-0.0115</td>
<td>0.0763</td>
<td>-0.0535</td>
</tr>
<tr>
<td></td>
<td>-4.52</td>
<td>-22.59</td>
<td>17.69</td>
<td>-10.63</td>
</tr>
<tr>
<td>OJN CoE</td>
<td>-0.0159</td>
<td>-0.0951</td>
<td>0.0726</td>
<td>-0.0475</td>
</tr>
<tr>
<td></td>
<td>-3.44</td>
<td>-20.63</td>
<td>15.97</td>
<td>-9.24</td>
</tr>
<tr>
<td>CT CoE</td>
<td>-0.0258</td>
<td>-0.0985</td>
<td>0.0623</td>
<td>-0.0545</td>
</tr>
<tr>
<td></td>
<td>-6.11</td>
<td>-23.71</td>
<td>14.92</td>
<td>-11.79</td>
</tr>
</tbody>
</table>
The sample period is February 1976 to July 2009 (402 months). Panel A reports implied factor premia from univariate asset pricing tests (i.e., associations of realized returns with factor betas), and Panel B reports correlations between CofE metrics and realized returns. The tabulated coefficient estimates are averages of the 402 cross-sectional regression coefficients (Pearson correlation coefficients) in Panel A (Panel B). T-statistics are based on the time-series standard error of the monthly estimates (Fama and MacBeth 1973). The column labeled “Full Returns Sample” in Panel A shows the results of estimating Equation (1b) using all monthly returns observations, 6,122 on average. The column labeled “Full CofE Sample” in Panel B shows correlations calculated using all monthly returns observations, 955 on average. The columns labeled “Negative Excess Return” and “Positive Excess Return” in both panels show results after segmenting monthly excess returns on the sign of each firm’s excess return. The columns labeled “Induced Sampling Increase in Positive Excess Returns” in both panels show results using resampled excess returns. We resample by month, for each of 402 sample months, and with replacement. The resampling alters the negative or positive proportions of excess returns, relative to the Full Sample proportions, by the specified percentages shown in the column headers. For example, in the column labeled +2.50%, we increase the proportion of positive excess returns by 2.50% , and decrease the proportion of negative excess returns by the same percentage, relative to the proportions of positive and negative returns in the Full Returns Sample.
The sample period is February 1976 to July 2009 (402 months). The average cross section in the Full Returns (reference) sample consists of 6,122 firms with at least 12 consecutive months of CRSP returns data. The Actual Cost of Equity (CofE) sample is a perfect subsample of the Full Returns sample, and contains an average of 955 firms each month with sufficient data to calculate (monthly variants of) five CofE estimates based on Value Line data, denoted VL, and based on models in Claus and Thomas (2001, CT), Gebhardt, Lee and Swaminathan (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN). The table shows the results of estimating Equation (1b) using all monthly returns observations in the Full Returns Sample, a ‘Random Subsample’ drawn from the Full Returns sample (averaged across 20 runs), in which each monthly cross section is of equal size to the CofE sample in that month. The last column contains the results for the Actual CofE sample.

<table>
<thead>
<tr>
<th></th>
<th>Full Returns Sample</th>
<th>Random Subsamples (CofE Sample Size)</th>
<th>Actual CofE Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta(_{\text{Market}})</td>
<td>0.0052</td>
<td>0.0051</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>1.95</td>
<td>0.88</td>
</tr>
<tr>
<td>beta(_{\text{SMB}})</td>
<td>0.0028</td>
<td>0.0027</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>1.69</td>
<td>1.58</td>
<td>0.20</td>
</tr>
<tr>
<td>beta(_{\text{HML}})</td>
<td>-0.0020</td>
<td>-0.0019</td>
<td>-0.0005</td>
</tr>
<tr>
<td></td>
<td>-1.33</td>
<td>-1.21</td>
<td>-0.32</td>
</tr>
<tr>
<td>beta(_{\text{AQFactor}})</td>
<td>0.0077</td>
<td>0.0073</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>2.44</td>
<td>2.31</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Panel A: Both variables (standard) normally distributed

<table>
<thead>
<tr>
<th></th>
<th>CORR* = 0.5</th>
<th></th>
<th>CORR* = 0</th>
<th></th>
<th>CORR* = -0.5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KS</td>
<td>CORR</td>
<td>KS</td>
<td>CORR</td>
<td>KS</td>
<td>CORR</td>
</tr>
<tr>
<td>Population (n = 5,000)</td>
<td>N/A</td>
<td>0.5000</td>
<td>N/A</td>
<td>0.0003</td>
<td>N/A</td>
<td>-0.4999</td>
</tr>
<tr>
<td>Random Sample (m = 1,000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.6904)</td>
<td></td>
<td>(0.6878)</td>
<td></td>
<td>(0.6944)</td>
<td></td>
</tr>
<tr>
<td>Non-Random Sample I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranked abs. distance to mean (Eq. 4)</td>
<td>0.1385</td>
<td>0.3267</td>
<td>0.1403</td>
<td>0.0006</td>
<td>0.1403</td>
<td>-0.3268</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.0000)</td>
<td>0.0</td>
<td>(0.0000)</td>
<td>50.5</td>
<td>(0.0000)</td>
<td>100.0</td>
</tr>
<tr>
<td>Distribution-Matched Sample</td>
<td>0.0302</td>
<td>0.4856</td>
<td>0.0297</td>
<td>0.0018</td>
<td>0.0299</td>
<td>-0.4877</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.5422)</td>
<td>30.1</td>
<td>(0.5486)</td>
<td>53.6</td>
<td>(0.5391)</td>
<td>71.4</td>
</tr>
<tr>
<td>Non-Random Sample II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform distribution</td>
<td>0.2550</td>
<td>0.7775</td>
<td>0.2546</td>
<td>0.0015</td>
<td>0.2533</td>
<td>-0.7780</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.0000)</td>
<td>100.0</td>
<td>(0.0000)</td>
<td>53.2</td>
<td>(0.0000)</td>
<td>0.0</td>
</tr>
<tr>
<td>Distribution-Matched Sample</td>
<td>0.0093</td>
<td>0.4953</td>
<td>0.0095</td>
<td>0.0014</td>
<td>0.0094</td>
<td>-0.4943</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.9992)</td>
<td>43.7</td>
<td>(0.9992)</td>
<td>53.2</td>
<td>(0.9987)</td>
<td>60.7</td>
</tr>
<tr>
<td>Non-Random Sample III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranked abs. distance to maximum</td>
<td>0.2665</td>
<td>0.4236</td>
<td>0.2628</td>
<td>0.0017</td>
<td>0.2645</td>
<td>-0.4233</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.0000)</td>
<td>0.0</td>
<td>(0.0000)</td>
<td>53.5</td>
<td>(0.0000)</td>
<td>99.9</td>
</tr>
<tr>
<td>Distribution-Matched Sample</td>
<td>0.0348</td>
<td>0.4873</td>
<td>0.0348</td>
<td>0.0034</td>
<td>0.0347</td>
<td>-0.4831</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.4003)</td>
<td>31.5</td>
<td>(0.4078)</td>
<td>55.0</td>
<td>(0.4073)</td>
<td>77.5</td>
</tr>
</tbody>
</table>

(continued on next page)
Table 4 presents correlation results for simulated populations of data, for random samples, and for three types of non-random samples with corresponding distribution-matched samples. Two variables $x$, $y$ are constructed as described in the text, with a given (true) correlation $\text{CORR}^* = \{0.5, 0, -0.5\}$. Panel A contains the results for two standard-normally distributed variables. Panel B relaxes the normality assumption, with $y \sim (0,1,3,10)$ and $x \sim (0,1,-1,3)$. Results are averages from 1,000 runs. Populations consist of 5,000 observations, from which samples of 1,000 observations are drawn, randomly or non-randomly. The three types of non-random samples are drawn with selection probabilities that are functions of $y$: For the ‘Non-Random Sample I’, the selection probability is decreasing in the ranked absolute distance from the mean, following Equation 3 in the text. ‘Non-Random Sample II’ is based on the exogenously given uniform distribution (increasingly higher selection probabilities for observations in the tails). ‘Non-Random Sample III’ is based on the ranked distance from the maximum value of $y$ (selection probability strictly increasing in $y$). Distribution-matched samples consist of observations from the non-random samples only, and are constructed by bootstrapping such that the empirical cumulative distribution of $y$, $F_{\text{NRS}}(y)$, in the non-random sample mimics the empirical distribution of $y$ in the population, $F_{\text{POP}}(y)$. The difference in the empirical distributions of $y$ is assessed using the Kolmogorov-Smirnov statistic (‘KS’). $p$-values are the (asymptotic) p-values from tests of distribution equality between the population and the respective sample ($F_{\text{NRS}}(y) = F_{\text{POP}}(y)$). ‘CORR’ denotes the estimated correlations. ‘Percentile (Random Sample)’ contains the percentile of the mean non-random sample correlation in the distribution spanned by the 1,000 correlations from the random samples.

<table>
<thead>
<tr>
<th></th>
<th>CORR* = 0.5</th>
<th></th>
<th>CORR* = 0</th>
<th></th>
<th>CORR* = -0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KS</td>
<td>CORR</td>
<td>KS</td>
<td>CORR</td>
<td>KS</td>
</tr>
<tr>
<td>Population (n = 5,000)</td>
<td>N/A</td>
<td>0.4999</td>
<td>N/A</td>
<td>0.0005</td>
<td>N/A</td>
</tr>
<tr>
<td>Random Sample (m = 1,000)</td>
<td>0.0245</td>
<td>0.5010</td>
<td>0.0246</td>
<td>0.0009</td>
<td>0.0244</td>
</tr>
<tr>
<td>(p-value)</td>
<td>(0.6851)</td>
<td></td>
<td>(0.6833)</td>
<td></td>
<td>(0.6921)</td>
</tr>
<tr>
<td>Non-Random Sample I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1419</td>
<td>0.3704</td>
<td>0.1450</td>
<td>-0.0003</td>
<td>0.1428</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.0000)</td>
<td>0.0</td>
<td>(0.0000)</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>0.0326</td>
<td>0.5018</td>
<td>0.0332</td>
<td>0.0003</td>
<td>0.0323</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.4691)</td>
<td>50.6</td>
<td>(0.4328)</td>
<td>48.9</td>
</tr>
<tr>
<td>Non-Random Sample II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4993</td>
<td>0.7728</td>
<td>0.5033</td>
<td>0.0024</td>
<td>0.4976</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.0000)</td>
<td>100.0</td>
<td>(0.0000)</td>
<td>51.2</td>
</tr>
<tr>
<td></td>
<td>0.0135</td>
<td>0.5023</td>
<td>0.0135</td>
<td>0.0007</td>
<td>0.0134</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.9837)</td>
<td>51.4</td>
<td>(0.9831)</td>
<td>49.0</td>
</tr>
<tr>
<td>Non-Random Sample III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2649</td>
<td>0.4305</td>
<td>0.2615</td>
<td>0.0007</td>
<td>0.2654</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.0000)</td>
<td>0.1</td>
<td>(0.0000)</td>
<td>49.0</td>
</tr>
<tr>
<td></td>
<td>0.0355</td>
<td>0.4808</td>
<td>0.0349</td>
<td>-0.0002</td>
<td>0.0355</td>
</tr>
<tr>
<td>(p-value)</td>
<td>Percentile (Random Sample)</td>
<td>(0.4070)</td>
<td>17.7</td>
<td>(0.3933)</td>
<td>48.1</td>
</tr>
</tbody>
</table>
Table 5 shows correlations between five CoE measures and excess returns for the Actual CoE sample and distribution-matched samples. For the ‘KS-based Sampling’, we construct distribution-matched samples that aim to minimize the non-parametric Kolmogorov-Smirnov (KS) statistic that captures general differences in the empirical distribution of excess returns between the Full Returns sample and the CoE sample. We perform the simulation 30 times and select the sample with the lowest KS statistic. This procedure is repeated for all 402 sample months. We preset the initial sample size for iteration either to 20% of the actual CoE sample that month (‘Initial # = 20%’), or to 100 unique firms (‘Initial # = 100’). For the ‘bin-based weighted sampling’ procedure, we divide the month-specific returns distributions of the Full Returns sample and the CoE sample into “bins” (intervals) of 100 basis points. Each month, we then redraw, with replacement, from the CoE sample to mimic the corresponding sample proportions in the Full Returns sample bin. Bins with no observations in the corresponding CoE sample are dropped. Bins in the extreme tails are weighted using Equation (4) in the text. We iterate the weighting factor γ to minimize the average difference in standard deviations between the Full Returns sample and the CoE sample. We repeat this resampling procedure 20 times. We first compute cross-sectional correlations each month and run and average correlations and associated time-series t-statistics for the 402 months in each run. The table contains the grand averages of average correlations and t-statistics across the 20 runs.
Table 6 presents semipartial correlation coefficients between excess returns and five CoE measures, controlling for the inverse Mills ratio from a Heckman-type selection model. The tabulated results are averages of monthly estimates and counts. The column headers refer to the explanatory variables in the probit selection model. The first column, labeled ‘No correction’, repeats the monthly average Pearson correlations from Table 2. Lag (MktCap) is the market capitalization from CRSP at prior month end. Volume is the CRSP trading volume in shares for the respective month. Age is the difference, in months, between the first month on CRSP and the month analyzed. B/M is the book-to-market ratio from Compustat as of the most recent fiscal year end. All characteristics variables are used in log form. ‘Factor Betas’ are the four univariate factor betas as previously defined. The ‘Combined’ selection model uses all characteristics from Model (IV) plus the four factor betas in Model (V). The upper (lower) definition of pseudo R^2 follows the computation by Cox and Snell (1989). Cox-Snell estimates are not rescaled to a theoretical maximum of 1; Nagelkerke (1991) suggests the validity of such rescaling. The row ‘Avg. Sample N’ (‘Avg. Reference Sample N’) contains the monthly average number of observations used in the selection model; the corresponding sample loss is the average monthly percentage of observations in the CoE sample (the Full Returns sample) without all necessary data for the various selection models, over the month-specific number of observations with CoE data (returns data).

<table>
<thead>
<tr>
<th></th>
<th>No correction</th>
<th>Lag (MktCap)</th>
<th>Lag (MktCap), Volume, Age</th>
<th>Lag (MktCap), Volume, Age, B/M</th>
<th>Factor Betas</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td>(II)</td>
<td>(III)</td>
<td>(IV)</td>
<td>(V)</td>
<td>(VI) = (IV) + (V)</td>
<td></td>
</tr>
<tr>
<td>VL CoE</td>
<td>-0.0033</td>
<td>-0.0079</td>
<td>-0.0060</td>
<td>-0.0067</td>
<td>-0.0171</td>
<td>-0.0055</td>
</tr>
<tr>
<td></td>
<td>-0.52</td>
<td>-1.32</td>
<td>-0.98</td>
<td>-1.10</td>
<td>-3.05</td>
<td>-0.91</td>
</tr>
<tr>
<td>GLS CoE</td>
<td>-0.0096</td>
<td>-0.0146</td>
<td>-0.0118</td>
<td>-0.0126</td>
<td>-0.0170</td>
<td>-0.0117</td>
</tr>
<tr>
<td></td>
<td>-2.19</td>
<td>-3.57</td>
<td>-2.81</td>
<td>-3.02</td>
<td>-4.23</td>
<td>-2.81</td>
</tr>
<tr>
<td>MPEG CoE</td>
<td>-0.0203</td>
<td>-0.0266</td>
<td>-0.0235</td>
<td>-0.0244</td>
<td>-0.0259</td>
<td>-0.0234</td>
</tr>
<tr>
<td></td>
<td>-4.52</td>
<td>-6.49</td>
<td>-5.59</td>
<td>-5.81</td>
<td>-6.68</td>
<td>-5.70</td>
</tr>
<tr>
<td>OJN CoE</td>
<td>-0.0159</td>
<td>-0.0215</td>
<td>-0.0190</td>
<td>-0.0200</td>
<td>-0.0221</td>
<td>-0.0191</td>
</tr>
<tr>
<td></td>
<td>-3.44</td>
<td>-5.00</td>
<td>-4.32</td>
<td>-4.55</td>
<td>-5.44</td>
<td>-4.42</td>
</tr>
<tr>
<td>CT CoE</td>
<td>-0.0258</td>
<td>-0.0299</td>
<td>-0.0279</td>
<td>-0.0286</td>
<td>-0.0301</td>
<td>-0.0278</td>
</tr>
<tr>
<td></td>
<td>-6.11</td>
<td>-7.73</td>
<td>-6.98</td>
<td>-7.24</td>
<td>-7.96</td>
<td>-7.15</td>
</tr>
<tr>
<td>Avg. Pseudo R^2 (Cox-Snell)</td>
<td>N/A</td>
<td>0.28</td>
<td>0.30</td>
<td>0.32</td>
<td>0.03</td>
<td>0.32</td>
</tr>
<tr>
<td>Avg. Pseudo R^2 (rescaled)</td>
<td>N/A</td>
<td>0.48</td>
<td>0.51</td>
<td>0.51</td>
<td>0.05</td>
<td>0.52</td>
</tr>
<tr>
<td>Avg. Sample N</td>
<td>955</td>
<td>955</td>
<td>943</td>
<td>939</td>
<td>955</td>
<td>939</td>
</tr>
<tr>
<td>Avg. Sample Loss (%)</td>
<td>N/A</td>
<td>0.0%</td>
<td>1.7%</td>
<td>2.1%</td>
<td>N/A</td>
<td>2.1%</td>
</tr>
<tr>
<td>Avg. Reference Sample N</td>
<td>6,122</td>
<td>6,117</td>
<td>5,670</td>
<td>4,883</td>
<td>6,122</td>
<td>4,883</td>
</tr>
<tr>
<td>Avg. Reference Sample Loss (%)</td>
<td>N/A</td>
<td>0.1%</td>
<td>10.0%</td>
<td>22.4%</td>
<td>N/A</td>
<td>22.4%</td>
</tr>
</tbody>
</table>
**Table 7**

Univariate Association between (Excess) Returns and Risk Factor Betas (Implied Factor Premia)

<table>
<thead>
<tr>
<th>S&amp;P500 Member</th>
<th>Listed on NYSE</th>
<th>σ(EPS forecasts) missing</th>
<th>Price at least $5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>beta&lt;sup&gt;Market&lt;/sup&gt;</td>
<td>0.0053</td>
<td>0.0017</td>
<td>0.0059</td>
</tr>
<tr>
<td></td>
<td>2.08</td>
<td>0.65</td>
<td>2.27</td>
</tr>
<tr>
<td>beta&lt;sup&gt;SMB&lt;/sup&gt;</td>
<td>0.0029</td>
<td>-0.0007</td>
<td>0.0031</td>
</tr>
<tr>
<td></td>
<td>1.74</td>
<td>-0.43</td>
<td>1.86</td>
</tr>
<tr>
<td>beta&lt;sup&gt;HML&lt;/sup&gt;</td>
<td>-0.0021</td>
<td>-0.0001</td>
<td>-0.0023</td>
</tr>
<tr>
<td></td>
<td>-1.38</td>
<td>-0.04</td>
<td>-1.48</td>
</tr>
<tr>
<td>beta&lt;sup&gt;AQFactor&lt;/sup&gt;</td>
<td>0.0078</td>
<td>0.0010</td>
<td>0.0083</td>
</tr>
<tr>
<td></td>
<td>2.47</td>
<td>0.30</td>
<td>2.62</td>
</tr>
</tbody>
</table>

| Avg. Proportion of Firms | 91.56% | 8.44% | 67.56% | 32.44% | 41.34% | 58.66% | 24.40% | 75.60% |
| Avg. Number of Firms     | 5,621  | 500   | 4,131  | 1,991  | 2,641  | 3,520  | 1,497  | 4,625  |
| Avg. KS statistic        | 0.0129 | 0.1413 | 0.0454 | 0.0955 | 0.0761 | 0.0521 | 0.1655 | 0.0540 |
| Avg. p-value             | (0.7109) | (0.0051) | (0.0264) | (0.0022) | (0.0055) | (0.0471) | (0.0000) | (0.0232) |
| Average Mean Excess Return | 0.0085 | 0.0069 | 0.0090 | 0.0072 | 0.0061 | 0.0088 | -0.0012 | 0.0113 |
| Average Std. Dev.        | 0.1663 | 0.0862 | 0.1816 | 0.1037 | 0.1180 | 0.1829 | 0.2343 | 0.1235 |
| Average Skewness         | 3.6869 | 0.7117 | 3.5655 | 1.3746 | 0.9585 | 3.9172 | 3.0356 | 2.5773 |
| Average Kurtosis         | 78.5929 | 8.1166 | 66.7037 | 20.7575 | 11.3998 | 71.4865 | 36.6523 | 55.6653 |

The sample period is February 1976 to July 2009 (402 months). The average cross section in the Full Returns sample contains 6,122 firms with at least 12 consecutive months of CRSP returns data. The table shows the results of estimating Equation (1b) using all monthly returns observations in the sample period, separated using various cross-sectional sample selection criteria. We analyze subsamples based on: month-specific membership in the S&P 500, listing on the NYS, and whether sufficient analyst earnings forecasts exist to compute forecast dispersion metrics on IBES. The rightmost column identifies all returns observations that are calculated with both lagged and current prices higher or equal to $5.