Non-Random Sampling and

Association Tests on Realized Returns and Risk Proxies

Frank Ecker*
(Duke University)

Jennifer Francis
(Duke University)

Per Olsson
(ESMT)

Katherine Schipper
(Duke University)

This paper investigates how data requirements can induce a non-random selection of observations from the reference sample to which the researcher wishes to generalize results. We illustrate the effects of non-random sampling on results of association tests in a setting with data on one variable of interest for all observations, and frequently-missing data on another variable of interest. We develop and validate a resampling approach to construct samples that approximate randomly-drawn samples, using only observations from the data-restricted subsample. Simulation tests show these distribution-matched samples yield generalizable results. We demonstrate the effects of non-random sampling in an archival setting using tests of the association between realized returns and five implied cost of equity metrics. Here, the reference sample has full information on realized returns, but only 16% of reference sample observations have data on cost of equity metrics. In contrast to inferences from the unadjusted (non-random) cost of equity sample which shows weak or negative associations, distribution-matched samples show reliable evidence of the theoretically predicted positive association between realized returns and cost of equity metrics. Additional analyses compare distribution-matching with other approaches to dealing with non-random samples, specifically, multiple imputation and selection models.

June 2015

* Corresponding author: Frank Ecker, frank.ecker@duke.edu. We appreciate financial support from Duke University's Fuqua School of Business. We thank Mary Barth, Alex Belloni, Alon Brav, Federico Bugni, Judson Caskey, Qi Chen, Shuping Chen, Dain Donelson, Ron Dye, Peter Easton, Jason Hall, Xu Jiang, Bill Kinney, Stephannie Larocque, Charles Lee (Stanford Summer Camp discussant), Fan Li (Duke University Department of Statistics), Xi Li (SMU-SOAR Symposium discussant), John McInnis, Maria Ogneva (FARS Meeting discussant), Panos Patatoukas, Jerry Reiter (Duke University Department of Statistics), Hanna Setterberg, Yong Yu and seminar participants at Dartmouth College, Stockholm School of Economics, University of Indiana, University of Iowa, University of Maryland, University of Munich, University of Notre Dame, University of Texas, Singapore Management University's SOAR Symposium 2013, FARS Midyear Meeting 2014, the Stanford University Accounting Summer Camp, and the 5th International Corporate Governance Conference at Tsinghua University for their helpful comments and suggestions. The paper was formerly circulated under the title "Association Tests of Realized Returns and Risk Proxies Using Non-Random Samples".

1. Introduction

This study examines how non-randomness in data-restricted samples affects empirical assessments of associations of interest and proposes a resampling technique to adjust for the effects of non-randomness and thereby increase the generalizability of results. We define non-randomness by comparing a data-restricted sample to a definable reference sample. Our resampling procedure ("distribution matching") aims to mimic a marginal reference distribution using only observations from the non-random sample. We validate the distribution matching technique on simulated data with known induced levels of correlations and three forms of non-randomness, and apply it to an archival setting characterized by stringent data requirements, namely the association of realized returns and implied cost of equity (*CofE*) metrics. We show the sample with data on *CofE* metrics is a non-random sample of the reference sample, defined as all listed firms with at least 12 months of realized returns. We show that tests of the *CofE*-returns association using a distribution-matched *CofE* sample lead to different conclusions compared to those based on the unadjusted (non-random) *CofE* sample. Finally, using two-stage asset pricing tests, we illustrate that non-randomness resulting from data requirements is a pervasive phenomenon.

Our study contributes both methodologically and substantively. Methodologically, we propose and validate an approach for dealing with non-random samples in empirical-archival settings, when the cause of non-random sampling is data availability requirements. We use information about the reference sample, beyond the standard "complete-case analysis" which requires complete data on all variables, to correct for the non-randomness created by data requirements. The goal is to assist researchers in constructing more powerful and less biased test samples, thereby increasing the generalizability of results. The substantive contribution is to shed light on previously documented weak or no associations between realized returns and implied cost of equity estimates based on analyst forecasts (e.g., Easton and Monahan, 2005; Botosan and Plumlee, 2005; Guay, et al., 2011). In

contrast to this research, we document reliably positive associations, once we adjust for non-randomness of the data-restricted cost of equity sample.[1]

Starting from the observation that empirical accounting research often uses "complete cases," meaning observations with complete data for all variables of interest, we examine a pairwise association (correlation or regression coefficient) in a stylized setting where a reference sample contains full information on one variable (y) and restricted data on the second variable (x). Because complete-case analyses effectively impose the data restrictions of x on y, thereby ignoring information about y in incomplete observation pairs, a complete-case samples may be a non-random sample of y, which leads to association results that are not generalizable to the reference sample. Our approach uses information lost through data requirements to increase generalizability.

We propose, validate and illustrate a resampling approach ("distribution matching") that alleviates the bias of non-randomness on association tests. We use information about the marginal distribution of y (the reference distribution), and resample pairs of observations (x and y) from the data-restricted sample to match the reference distribution as closely as possible. The goal is to construct samples that appear as if they were drawn randomly from the reference distribution, despite the data restrictions. We validate distribution matching with simulated data with known induced levels of statistical associations in the population. We draw three types of non-random samples based on the marginal distribution of one of the variables and document the resulting biases in correlations in these samples. Despite resampling *only* from these non-randomly selected observations, we show that the bias in correlations is substantially reduced or even eliminated by distribution matching.

Having shown that distribution matching works for simulated data, we apply the technique to an archival example, the associations between realized returns and cost of equity. Our reference sample is the population of U.S. listed firms with at least 12 consecutive monthly returns on CRSP during February 1976 to July 2009 (the Full Returns sample). This reference sample meets two key conditions. First, the test of interest is theoretically

[1] We do not attempt to improve on implied *CofE* models in other ways, such as correcting for biases in analyst forecasts (Guay et al., 2011) or using statistical forecasting models for earnings (Hou et al., 2012; Li and Mohanram, 2014). While we view such approaches as useful in their own right, we focus on the properties of the original models as used by much prior work (see Section 3). In addition we show that four of the five *CofE* metrics show a reliably negative association with returns in the unadjusted non-random sample, providing a conservative benchmark result.

meaningful, in the sense that results from actual research samples could be generalized to the reference sample. Second, the empirically determined (marginal) distribution of a variable of interest—realized returns—has few data requirements. By definition, all reference sample firms have a cost of equity, but the researcher cannot observe the *CofE* metric(s) for all firms because of data requirements.[2] That is, our reference sample contains complete data on one variable (*y*, returns) and not on the other variable (*x*, *CofE*).[3]

Descriptive statistics indicate the returns of the *CofE* sample are not a random subsample of reference sample returns; the *CofE* sample has substantially lower standard deviation, skewness and kurtosis. This is not surprising, given that data requirements for the *CofE* models, such as analyst following, positive earnings forecasts and positive earnings growth, typically lead to samples of larger, more stable firms (Francis et al., 2004; Easton and Monahan, 2005). As a benchmark result from this non-random sample, the *CofE* metrics are either negatively (four metrics) or insignificantly (one metric) correlated with realized excess returns, confirming prior findings concerning the associations between realized returns and (unadjusted) *CofE* measures.

To probe whether the non-randomness in returns systematically affects association tests, we first analyze the relation between realized returns and risk factor betas for the Full Returns sample (our reference sample), a random subsample and the actual *CofE* sample. Imposing an unnecessary data restriction into an association test with risk metrics (factor betas) that can be performed on the entire reference sample allows us to separate the effects of non-randomness from the effects of a reduced sample size. The coefficients on factor betas[4] for the random sample of equal size as the *CofE* sample are similar in magnitude and statistical significance to those for the reference sample, while for the actual *CofE* sample, 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 inferences; and (2) the *CofE* sample should not be assumed to be a random subsample of the reference sample.

---

[2] Four *CofE* models are based on analysts' earnings forecasts (Claus and Thomas, 2001; Gebhardt, et al., 2001; Easton, 2004; Ohlson and Jüttner-Nauroth, 2005). The fifth model uses Value Line target prices and dividend forecasts.

[3] The problem is equivalent to an "item non-response" in an otherwise complete questionnaire. The item is known to exist, but the data are not available to the researcher.

[4] These regression coefficients are interpretable as implied factor premia (for example, the coefficient in a regression of excess returns on market beta can be interpreted as the implied market risk premium).

To apply distribution matching to the actual *CofE* sample, we resample the observations in the *CofE* sample, so that the returns distribution in the distribution-matched *CofE* sample mimics the returns distribution in the reference sample. 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. We distribution-match by constructing monthly subsamples of *CofE* firms that minimize the deviation between the returns distribution of the *CofE* sample and the reference distribution of returns, based on the KS statistic. In contrast to results obtained for the unadjusted *CofE* sample, correlations between returns and *CofE* measures for the distribution-matched subsamples are positive and significant across specifications, except for one specification using the *CofE* metric based on Claus and Thomas, 2001. Our second approach addresses the practical concern that, while the KS-based resampling procedure is effective and requires few assumptions, it imposes a substantial computational burden, especially for large samples. We propose a second distribution-matching approach that sorts the returns distributions of both the reference sample and the *CofE* sample into researcher-defined strata ("bins") of the continuous variable, and then applies a form of stratified resampling that aims to match the standard deviation of the reference returns distribution. Similar to results obtained using the KS-based approach, we find that correlations between realized returns and *CofE* metrics are reliably positive in bin-based distribution-matched samples.

Our analysis shows that distribution matching can result in smaller estimation samples than the original non-random samples. In contrast with the standard approach of using the largest possible number of observations with complete data on both variables, we show that it is not necessarily the case that data-dictated samples of maximized size lead to unbiased inferences. Rather, there is a tradeoff between test power and generalizability.

We also discuss and illustrate two other approaches to dealing with missing data: (1) multiple imputation, which uses a stochastic regression framework to estimate (impute) values for the missing data and (2) Heckman-type selection models, which treat the *CofE* sample as a choice-based subsample of the reference

sample. In contrast to distribution-matching, which is designed to be non-parametric, both multiple imputation and selection models require a normality assumption, raising doubts about the specific applicability of both approaches in tests using realized returns. In addition, applying a Heckman-type approach assumes the ability to estimate the selection model on a random sample of the population, an assumption likely to be violated in practice, where there is typically a trade-off between selection model fit and data requirements for the selection model variables. In other words, a selection model approach transfers issues related to data restrictions from the test model to the selection model. With these concerns in mind, we illustrate the multiple imputation and selection-model approaches.

We believe our findings support both methodological and substantive inferences. The methodological inference is that selection criteria yielding samples with outcome distributions differing from the reference distribution can materially affect the results of association tests, including producing results that do not generalize to the reference sample to which the tested hypothesis applies. Many samples are likely to be characterized by non-randomness induced by selection criteria. Specifically, restrictive data requirements for explanatory or control 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. We provide two illustrations beyond the *CofE* setting. First, we apply several plausible selection criteria to the reference sample, including S&P 500 membership, NYSE listing, availability of a dispersion measure of analyst earnings forecasts and stock price of $5 or above. We show that application of these criteria changes the distribution of realized returns and leads to biased estimates in association tests of realized returns with risk factor premia. Second, we induce changes in the distribution of realized returns, and show the sensitivity of risk factor premia to these changes. Together with our prior findings, we conclude that results obtained using unadjusted non-random samples may not support generalizations to the reference sample. In fact, our analysis of the *CofE* sample highlights that maximizing the size of the non-random sample, after imposing data requirements, may conflict with the goal of obtaining a random sample, which is fundamental for the generalizability of the results.

Our substantive inference speaks to the generally weak or non-existent associations between realized returns and risk proxies documented in prior research. While results have led some researchers to conclude that *CofE* estimates and risk factor betas are poor proxies for (or poor determinants of) expected returns, or that realized returns are a poor proxy for expected returns, or some combination of the two, our inference is that returns-risk associations are *expected* to differ across studies that use different samples, if those samples are not randomly drawn from the population of realized returns. In particular, our findings from distribution-matched samples suggest that analyst-based *CofE* measures have greater construct validity than results from previous research would indicate.

In the next section, we introduce the problem of non-randomness for association tests, develop and validate the distribution matching approach using simulated data, and clarify the relation between distribution matching and other missing-data methods. Section 3 describes the returns-*CofE* setting of our archival application. Section 4 describes the archival data and test design, both before and after we implement distribution matching and reports results from multiple imputation for comparison. Section 5 discusses extensions to other settings and compares distribution matching with Heckman-type selection models. Section 6 concludes.

2. Motivation and Validation of Distribution Matching

2.1 Motivation and intuition

Data constraints commonly lead to situations where empirical tests can be performed only on a subsample of observations, even though the test results are intended to generalize to a population or reference set of observations, to which the tested hypothesis logically applies. We focus on association tests (regression coefficients or correlation coefficients) between two variables in a stylized research setting. One variable (y) is available for all firms in the researcher-defined reference sample, and data on the second variable (x) are often missing.[5] In this situation, a common treatment in the accounting literature is list-wise deletion, i.e., restricting the test sample to observations with complete information, yielding a data-restricted sample. The data

[5] In the empirical example described later, data on y (realized returns) are available for all firms in the reference sample while data on x (*CofE* metrics) are missing for 84% of observations in the average cross-section.

constraints on *x* are imposed on *y*, causing information about the unrestricted distribution of *y* to be lost. We hypothesize this list-wise deletion leads to non-random test samples, and that association tests based on such non-random samples yield results that may not be generalizable to the reference sample. We correct for this problem by incorporating information about the reference distribution (of the complete variable) into the association test. We resample observations from the data-restricted, non-random sample to create a test sample that mimics the reference distribution of the complete variable. Our approach creates a test sample that appears randomly drawn from the reference sample with respect to *y*, despite data constraints on *x*.

The intuition for this approach is as follows. Consider the estimate of a Pearson correlation coefficient $\hat{\rho}$ between two continuous random variables *x* and *y*:[6]

$$\hat{\rho} = \int\int x_i\, y_i\; f(x_i, y_i \mid s_i = 1)\, dxdy = \int\int x_i\, y_i\; f(x_i \mid y_i, s_i = 1)\; f(y_i \mid s_i = 1)\, dxdy \tag{1}$$

where x_i and y_i are standardized (demeaned and divided by their respective standard deviations) realizations of *x* and *y*, $f(.)$ denotes the density function, and s_i is an observation-level indicator for membership in the data-restricted test sample. For simplicity, subscripts for time *t* are suppressed.

The true correlation in the reference sample, assuming availability of complete data, is given by:

$$\rho^* = \int\int x_i\, y_i\; f(x_i, y_i)\, dxdy = \int\int x_i\, y_i\; f(x_i \mid y_i)\; f(y_i)\, dxdy \tag{2}$$

$\hat{\rho}$ is a consistent estimator of the true ρ^* only if the joint distribution in the restricted sample equals the joint distribution in the reference sample, $f(x_i, y_i \mid s_i = 1) = f(x_i, y_i)$, or equivalently, $f(x_i \mid y_i, s_i = 1)\, f(y_i \mid s_i = 1) = f(x_i \mid y_i)\, f(y_i)$. This condition implies that the unobserved data are missing

[6] In this discussion, the subsequent simulations and most of the empirical work, we focus on the correlation coefficient rather than the regression coefficient because the former is not affected by changes in the (relative) standard deviations of the two variables. Therefore, mechanical changes in standard deviations, e.g., because the reference distribution is more dispersed, or because of a reduction of the number of observations, will not confound our analysis. Examining correlation coefficients lets us demonstrate the effects of distribution matching in isolation. We discuss the (equivalent) effects on the regression coefficients in Section 4.4.2.

completely at random ('MCAR'); only then would a restricted sample (i.e., a sample after list-wise deletion of observations because of missing data) be a random subsample of the reference sample.[7]

In our stylized setting, as well as in some other accounting research settings, it is possible to assess the difference in the marginal distributions of the fully observed variable y_i between the restricted sample, $f(y_i \mid s_i = 1)$, and the reference sample, $f(y_i)$, and reject the assumption of MCAR. Differences between these marginal distributions mean the restricted sample is non-random, and consistency of $\hat{\rho}$ is less likely.

Therefore, the MCAR assumption must be replaced with a weaker assumption: Either the data are missing at random, conditional on observed variables ('MAR'), or the data are not missing at random ('NMAR'), which implies that missingness also depends on unobserved data. While it is possible to reject the MCAR condition, the unavailability of missing data makes it impossible test whether data are MAR or NMAR. Our main analyses extend the common approach of constructing test samples by list-wise deletion under the MCAR assumption, and focus on research methods under the MAR assumption. In Section 5, we assess the impact of a possible NMAR assumption using a Heckman-type selection model in our setting.

Referring to Equations (1) and (2), the distribution matching approach requires the distribution of x, conditional on y_i, is unchanged in the restricted sample compared to the reference sample:

$$f(x_i \mid y_i, s_i = 1) = f(x_i \mid y_i) \tag{3}$$

Assuming complete data on y, the MAR assumption implies that Equation (3) holds. Then $\hat{\rho}$ will converge to ρ^* as $f(y_i, s_i = 1)$ approaches $f(y_i)$ via distribution matching. In the context of our archival analysis, condition (3) implies the *CofE* metrics are not systematically biased in the restricted sample, conditional on the value of the future realized return. It seems unlikely that the probability of having the analysts' forecasts required to construct an implied *CofE* metric depends on the value of realized returns, which can only be assessed ex post. Regardless of any concerns specific to our setting, condition (3) contrasts with and is arguably

[7] We acknowledge that research can, and sometimes should, be performed on restricted or even intentionally biased samples. In those cases, results are not intended to be generalizable to a reference sample.

weaker than the more common assumption that data are missing completely at random, essentially equating $f(x_i, y_i \mid s_i = 1)$ with $f(x_i, y_i)$, particularly when differences in the marginal distributions of returns between the reference sample and the *CofE*-restricted sample, $f(y_i)$ versus $f(y_i \mid s_i = 1)$, are knowable from the data.

Under condition (3), our approach focuses on the marginal distribution of *y* in the restricted sample. We resample only from observations in the restricted and possibly non-random sample with complete data on both variables, but in a systematic way, so that, in the limit, the marginal distribution of y_i matches the marginal reference distribution of y_i:

$$f(y_i \mid s_i = 1) \longrightarrow f(y_i) \tag{4}$$

While the convergence in (4) is achievable in the limit, the effectiveness of distribution matching in a given research setting is a function of several factors, including the number of restricted sample observations and the size of the common support of the restricted sample and reference distribution. A smaller restricted sample means fewer observations to resample from and a smaller common support of the distributions means $f(y_i)$ is more severely truncated in the restricted sample. The resampling approach may also be less effective or even unnecessary if only a few observations are missing, making the restricted sample (nearly) equal to the reference sample.

Our measure of similarity in the cumulative distributions between the reference sample and the non-random data-restricted sample is the non-parametric Kolmogorov-Smirnov (KS) statistic, which computes the percentage maximum absolute distance between two cumulative empirical distributions:[8]

$$KS = \max_i \left| F^{NRS}(y_i) - F^{POP}(y_i) \right| \quad \text{where } i = 1, 2, ..., n \tag{5}$$

where $F^{NRS}(y_i), F^{POP}(y_i)$ are the cumulative distributions of *y* in the non-random sample and population, respectively. The KS statistic is associated with an asymptotic *p*-value for a significance test of distribution equality between a subsample and the reference sample. We use the KS statistic to assess the degree of

[8] Any test statistic that captures overall differences between two distributions could serve a similar function.

convergence in (4) within the KS-based distribution matching approach. The Appendix describes this approach for the simulation and Section 4.4.2 describes our illustration in an archival setting.

2.2 Relation of distribution matching to other missing-data approaches

Approaches applicable in a MAR setting. Alternative missing-data approaches under the MAR assumption include multiple imputation (MI)[9] and full-information maximum likelihood (FIML) estimation. Like distribution matching, both MI and FIML incorporate information about the marginal distribution $f(y_i)$. FIML does so by including the observations in the likelihood calculation, even if data on some variables are missing. MI uses a stochastic regression framework to impute possible values for the missing data multiple times, after which the completed ("imputed") datasets can be independently analyzed and the results aggregated. Complete variables, i.e., the marginal distribution of returns in our setting, are not imputed, but preserved from the reference sample and also considered in the analysis. MI uses the entire reference sample, so it is more efficient than distribution matching in cross-sectional analyses. MI can also be applied when data are missing for more than one variable, and can incorporate the use of auxiliary variables that are either informative about missingness or correlated with the missing data.

The approaches differ, however, with regard to underlying distributional assumptions. Distribution matching is designed to be non-parametric, and both multiple imputation and maximum likelihood estimation rely on multivariate normality. Maximum likelihood does so in the likelihood function itself. Multiple imputation restricts error sampling to normal distributions, and also samples imputation parameters from distributions that are assumed to be normal. Because the normality assumption seems unlikely to hold in our archival setting, given the descriptive statistics in Table 2, we first validate and apply distribution matching. Based on theoretical arguments in Schafer (1997), supported by simulation evidence in Demirtas et al. (2008), that multiple imputation appears to be less susceptible to deviations from multivariate normality than maximum likelihood, we repeat our main tests using various forms of multiple imputation in Section 4.5.

[9] Both the theoretical framework of MI and the validity of multiple imputation of MAR data are well established (see, e.g., Rubin 1987, Schafer 1997, and Little and Rubin 2002).

Other missing-data approaches. We do not aim to discuss the vast literature on missing-data methods, but rather to clarify the intuition of distribution matching by contrasting its features and assumptions with selected other methods. We start with a truncated regression model. Assuming the true distribution of the complete outcomes is normal, Tobin (1956) derives closed-form solutions for samples where the outcome variable is truncated at a known upper or lower bound (see also Wooldridge, 2010). We do not focus on this "tails problem," but rather emphasize that non-randomness likely manifests in a restricted sample with a different shape than the reference distribution, even if the common support is large or complete.[10] Also, we do not require assumptions about the reference distribution of the outcome variable, but rather estimate its shape from the reference sample with complete data.[11]

Our approach also draws intuition from the survey literature use of stratified sampling, or incomplete post-stratification. The latter involves reweighting observations according to their marginal weights in a reference distribution or population. The weights are typically constructed based on discrete and exogenous variables such as race or gender, not an outcome variable, for example. The intuition for this approach and our approach is similar, in that survey respondents need not be representative of the population and hence need to be weighted differentially, if the goal is to generalize results to the population.[12] To that end, both post-stratification and distribution matching import information about the marginal reference distribution. In fact, for the common support region of sample distribution and reference distribution, distribution matching is essentially a form of post-stratification that treats the variable as continuous (each y_i is its own stratum), and does not require ex-ante grouping of observations into strata. In addition, the sample distribution may not only be non-random within the common support region, but also truncated. Intuitively, the effect of truncation is mitigated by oversampling from the tails of the sample distribution.

---

[10] In our simulations we show that even when truncation is minimal, the bias in correlation coefficients in non-random samples of the outcome variable can be large.

[11] As the simulations illustrate, our procedure can also be used with a theoretically derived reference distribution, rather than an empirically estimated distribution.

[12] An alternative is to oversample from selected groups to ensure these groups are surveyed in the first place. Subsequently, observations from the selected oversampled groups are assigned the (lower) population weight.

Our approach differs from hot-deck imputations that use the entire reference sample as a test sample by filling in the missing values in incomplete observations using realized values from "donor" observations that are similar to the "recipient" observations based on a proximity metric, usually measured using complete variables for both observations. While our approach also uses only realized values of the missing variable, we resample whole observations from the restricted sample to match the known distribution of one variable. That is, we preserve pairs of the variables of interest. While distribution matching might decrease the size of the test sample, hot-deck imputations aim to maximize its size. In that regard, they are similar to multiple imputations.

Finally, distribution matching is distinct from Heckman-type selection model approaches that use a first-stage probit selection model to capture the mechanism that selects observations into the restricted sample. Under certain conditions,[13] the bias in the test model can be alleviated by incorporating the inverse Mills ratio from the selection model. In contrast, distribution matching treats the selection mechanism itself as ignorable, and uses information about its consequences by assessing and minimizing the difference of the sample distribution to a reference distribution. Section 5.1 discusses and illustrates the selection model approach in the context of our archival setting. We find that results from the restricted non-random samples, both on *CofE* and on factor betas, are little affected by including the inverse Mills ratio.

2.3 Validity tests on simulated data

We use simulated data to validate our distribution matching approach by showing that correlation estimates from distribution-matched samples converge to their true values, even though these samples consist only of non-randomly drawn observations from the reference sample.[14] For the first set of simulations (results reported in Table 1), we generate populations of data for two variables (y and x) with known correlations, and draw from these simulated populations both randomly and in three non-random ways, with selection probabilities based on

[13] Briefly, those conditions are (1) the (largely untestable) assumption of bivariate normality of selection model and test model residuals, and (2) the assumption that the selection model can be performed on a random sample of the reference sample. Many authors document the sensitivity of test results with respect to even minor departures from the normality assumption, leading to biases that may even exceed the bias from standard complete-case analyses. Due to this sensitivity, some authors go so far as to question the usefulness of selection models in practice (Enders, 2010).

[14] The Appendix details the design of our simulations, the generation of three distinct non-random samples, and our implementation of the distribution matching approach on the simulated data.

the marginal distribution of *y*. For each of the three types of non-random samples drawn from the simulated population, we then resample with replacement to create distribution-matched samples. Using only observations from the respective non-random sample, distribution matching is designed to mimic the marginal distribution of variable *y* in the population as closely as possible.

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 0.5, 0, and -0.5. We examine the zero-correlation case to rule out the possibility that distribution matching induces a correlation where none exists. We also examine both negative and positive correlations. The two symmetric non- random samples are expected to yield either a negative bias ('Non-random sample I') or a positive bias ('Non-random sample II'). Combined, the setup is intended to provide evidence that the effectiveness of distribution matching does not depend on the either sign of the true association or the sign of the bias in the association estimate. 'Non-random sample III' is a form of non-symmetric selection probability.

Figure 1 graphically represents the non-random draws and the effectiveness of distribution matching for the three types of non-random samples, drawn from normally distributed data. The figure depicts example distributions of *y* for a single simulation run, for both the unadjusted non-random samples (on the left) and the distribution-matched samples (on the right) of size $m = 1{,}000$. The benchmark distribution, which appears in both right and left graphs, is from the population in that particular run ($n = 5{,}000$). For all three non-random draws, the left-side graphs illustrate that the distributions deviate considerably from the population benchmark. After distribution matching, the right-side graphs show the sample distributions closely follow the population distribution and are indistinguishable for large regions of *y*.

Numerical results of the simulations analysis are reported in Table 1. We discuss the results in Panel A (normally distributed variables). Results in Panel B (non-normally distributed variables) are qualitatively similar, suggesting that the effectiveness of our non-parametric distribution matching approach does not depend on the shape of the marginal distributions; in particular, its effectiveness does not depend on a normality assumption. We first verify that empirical correlation estimates in the population and random samples

('CORR') are close to the specified (true) correlations ('CORR*'), and 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 selection/reduction is random.

In contrast, and by construction, non-random samples have a distribution of y that differs sharply from the population distribution. For 'Non-random sample I' ('Non-random sample II') ['Non-random sample III'], KS statistics are about 13.9% (25.5%) [26.7%]. To assess the significance of the bias in correlation estimates for these samples, we report the percentile of the mean non-random sample correlation in the distribution spanned by the 1,000 correlations as 'Percentile (Random Sample)'. A small (large) percentile corresponds to a low (high) 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 of 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 reduces or eliminates the effects of non-random sampling: across all three non-random samples, the corresponding distribution-matched samples exhibit correlations 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.

Based on these simulation results, we conclude that distribution matching is effective in reducing the bias in correlation estimates in non-random samples of y_i; results for zero correlations show that distribution matching does not induce an apparent correlation where none exists. The result for Non-random sample I, intended to mimic stable observations ("firms"; see Appendix for the definition), is of particular interest. Because the result is based on simulated data and an imposed stability criterion in the sample construction, we view this finding as suggesting the kind of bias in association tests if data availability requirements bias a

sample towards stable firms, as is often the case in actual research situations. Our next tests investigate whether this general simulation result applies to a specific empirical-archival setting well-studied in the literature.

3. Application to the Association of Realized Returns and Implied Cost of Equity Metrics

We test for a bias in association test results when samples are restricted to large stable firms, as is the case for samples of implied *CofE* metrics. Specifically, we re-examine the correlation between realized returns and analyst-based *CofE* metrics as they have been used to test for expected-returns associations. Settings include voluntary disclosure (Botosan 1997), AIMR scores (Botosan and Plumlee 2002), earnings attributes (Francis et al. 2004), earnings restatements (Hribar and Jenkins 2004), legal institutions/security regulation (Hail and Leuz 2006), shareholder taxes (Dhaliwal et al. 2007), mandatory IFRS adoption (Li 2010), earnings quality and information asymmetry (Bhattacharya et al. 2012) and financial constraints and taxes (Dai et al. 2013). We believe the construct validity of analyst-based *CofE* metrics is of interest to many researchers, so that an application of the distribution matching approach provides insights in its own right. Section 3.1 summarizes previous research and section 3.2 explains our choice of this setting to illustrate distribution matching.

3.1 Prior research on the association between realized returns and implied *CofE* metrics

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 cost of equity tests (associations of realized returns with implied *CofE* metrics). Tests of the latter are predicated on the assumption that 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. Realized returns are ex-post outcome measures that might 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: (1) realized returns are a reasonable proxy for expected returns; that is, the non-expected return component is small and/or cancels out through aggregation in broad samples or (2) even in broad samples, the non-expected return component is a key non-cancelling component of realized returns (e.g., Elton, 1999; Vuolteenaho, 2002).[15] Adopting the latter perspective, researchers have developed and analyzed several *CofE* metrics that are derived independently of realized returns (e.g., Gebhardt, et al. 2001; Claus and Thomas, 2001; Botosan and Plumlee, 2002; Easton, 2004; Brav, et al. 2005; Ohlson and Jüttner-Nauroth, 2005),[16] or, alternatively, researchers have empirically purged the realized return of its non-expected ("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, et al. (2011) and Ogneva (2012) control for a specific kind of fundamental (earnings) news in realized returns to directly identify the cash-flow news component of realized returns.[17]

Another stream of research views realized returns and *CofE* metrics as alternative, albeit imperfect, proxies for expected returns, and aims to validate *CofE* metrics jointly with realized returns in association tests. Easton and Monahan (2005) and Guay, et al. (2011) find the association between realized returns and several commonly used *CofE* estimates is often insignificant or even significantly negative. Botosan et al. (2011) find the association varies between positive and negative over time and is, on average, weak.

---

[15] Vuolteenaho (2002) concludes that cash flow news is the main driver of firm-specific realized returns. Elton (1999) observes there are periods exceeding 10 years during which realized stock returns are, on average, less than the risk-free rate, thereby questioning whether realized returns are a reasonable proxy for expected returns. He 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 exploring ex-ante cost of capital measures rather than realized returns.

[16] The *CofE* metrics are inferred from valuation models relating expectations of future cash flows, dividends or earnings to current price. By construction, these *CofE* metrics are derived from "static" valuation models and therefore are not affected by "news" over a measurement period in the same way that realized returns might be affected.

[17] 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.

The contrast of these weak results with economic intuition leads some researchers to propose and test approaches to increase the association. One approach attributes the weak association to realized returns. Using variance decomposition to control for non-expected return components in realized returns, Easton and Monahan (2005) find no, or significantly negative, association between "news-purged" realized returns and four of the seven *CofE* estimates they consider. In contrast, Botosan et al. (2011) use different empirical proxies for non-expected return components (similar to Ogneva, 2012), and find their *CofE* estimates have significant positive associations with "news-purged" returns. However, they also document their news-purged returns construct has either no association, or counter-intuitive association, with the risk-free rate, beta, book-to-market and other proxies for risk, leading them to question the validity of their "news-purged" realized returns as a proxy for expected returns, and to express caution about the approach. In terms of our research setting, we note that there seems to be a tradeoff in that adjustments to *CofE* metrics may worsen the relation between *CofE* metrics and other proxies for risk such as beta (Botosan et al., 2011). We address this potential concern in Section 4.4.2 by showing the coefficients on risk factors are hardly affected by our distribution matching approach.

Other papers attribute the weak association to the *CofE* metrics, or more specifically, to the analyst forecasts used as inputs. Guay et al. (2011) find that modifying analyst-based *CofE* metrics to account for "analyst sluggishness" improves the associations between some *CofE* proxies and realized returns.[18] Other studies move away from the firm level to a portfolio design: Gode and Mohanram (2003), for example, find positive spreads in realized returns. Hou, et al. (2012) also use portfolio-level tests and show that returns spreads increase when they replace analyst forecasts with regression-based earnings forecasts.[19] Li and Mohanram (2014) use the Hou et al. (2012) approach of deriving alternative earnings forecasts and show positive associations on the firm level as well.

[18] In their firm-specific tests, one proposed method yields t-statistics between -0.52 and 1.93 for five implied cost of capital proxies and the other method yields t-statistics between -0.50 and 1.58.

[19] Other research seeks to improve the earnings forecast regression model by modifying the explanatory variables (Li and Mohanram, 2014; Gerakos and Gramacy, 2013) and by using different regression methods (Gerakos and Gramacy, 2013).

3.2 Analysis of the *CofE* setting

We propose a different, possibly co-existing explanation for the weak and inconsistent results in tests of association between realized returns and *CofE* metrics, originally documented by Easton and Monahan (2005). In contrast to prior work that modifies these metrics, we leave their original construction in place, and propose an explanation that derives from known features of samples used to estimate *CofE* metrics. Specifically, because the data requirements for estimating *CofE* metrics eliminate firms with no analyst following, negative book value of equity, or negative or declining earnings forecasts, firms in a *CofE* sample tend to be larger and more profitable, hence likely more stable, than the CRSP population. For example, Francis et al. (2004, Table 1) report that the aggregate market capitalization of their sample of Value Line-followed firms, averaging 790 firms per year, is just over 47% of the CRSP market capitalization.[20] We posit these data requirements result in *CofE* samples that are non-random draws from the population of CRSP firms, with a returns distribution that differs from the returns distribution in that population (or a random sample thereof).[21] We further posit that association estimates based on such a non-random sample are difficult to generalize to the population, i.e., the external validity of the results is questionable. We do not dispute previous findings, but rather use distribution matching to arrive at results that can be more justifiably generalized to the population of listed firms.

We believe the *CofE*-realized returns association has the following desirable characteristics for an empirical examination of the effects of non-random sampling: (1) the full distribution of returns is available for the reference sample; (2) data on the *CofE* metric are missing for many reference sample firms, but conceptually all sample firms have a cost of equity; and (3) previous research shows the characteristics of returns for missing firms differ from the characteristics of returns for included firms.

To illustrate whether and how data requirements may result in non-random samples, we let the data requirements for five *CofE* metrics dictate the sampling bias in returns. While intuition suggests these data

[20] As discussed in Section 5, firm size is an important determinant of the selection model, even if the inclusion of the inverse Mills ratio has limited impact on the coefficients of interest.

[21] 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 do not dictate the data frequency in our tests, and disaggregated (e.g., monthly) data can be used.

requirements are likely to bias *CofE* samples towards more stable firms with less dispersion in returns than the returns of the reference sample, intuition does not necessarily suggest an effect on associations of *CofE* metrics with these returns. This intuition provides a second motivation to pursue a cost of capital question. For our reference sample, we have complete information on both realized returns and asset pricing factor betas (loadings) for all observations. In a CAPM world, cross-sectional variation in beta is equivalent to cross-sectional variation in expected returns. Hence, we can use association tests on factor loadings to gauge the non-randomness of the *CofE* sample by first performing factor loading association tests on the reference sample and then importing the *CofE* sample restriction into the same test. Differences in results would suggest the *CofE* sample is a non-random sample of the reference sample. Using Fama-MacBeth two-stage tests of the association of returns with risk factor betas, we show the *CofE* returns sample is indeed a non-random sample from the reference sample.

While the literature has proposed alternatives for improving the returns-*CofE* association, we choose the most conservative starting point (the original problem as first examined by Easton and Monahan (2005)) and use unmodified implied *CofE* definitions and unmodified realized returns. As explained later, we find that for distribution matched samples the sign of the association changes from significantly negative to significantly positive for most *CofE* metrics. The sign changes emphasize that distribution matching is not only about increasing test power, but can also yield opposing inferences, even in (relatively) large samples.

4. Test Design and Non-randomness of the *CofE* Sample

4.1 Sample and Descriptive Data

Table 2 describes the archival data used in our empirical tests. We first identify all firms with monthly CRSP returns data over the period February 1976 to July 2009 (402 months). These data are used for our cross-sectional asset pricing tests. The reference sample (the Full Returns sample) includes all firms with returns data in the current month and the preceding 11 months; a firm is required to have 12 consecutive monthly returns

observations to enter the Full Returns sample in Month *t*.[22] The Full Returns sample contains 2,460,998 firm-month observations, representing 24,657 unique firms. Table 2 of Panel A shows the average monthly cross section consists of 6,122 firms, with an average (median) monthly realized raw return of 1.30% (0.20%). Monthly 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.[23]

Panel B of Table 2 reports average cross-sectional statistics for the sample of firms with data to estimate the five *CofE* measures. On average, those cross sections contain 955 firms (383,955 monthly observations for 3,989 unique firms), a strict -- and potentially non-random -- subsample from the Full Returns sample. 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.[24] We calculate four other implied *CofE* estimates based on models in Claus and Thomas (2001, *CT*), Gebhardt, et al. (2001, *GLS*), Easton (2004, *MPEG*), and Ohlson and Jüttner-Nauroth (2005, *OJN*).[25] The firms in the *CofE* sample are expected to differ from those in the Full Returns sample because all *CofE* metrics require analyst following in general, and Value Line coverage in particular, as well as positive and increasing earnings forecasts.

As reported in Panel B of Table 2, the mean (median) values of the *CofE* estimates range from 0.0071 to 0.0121 (0.0067 to 0.0118). The mean (median) monthly realized excess returns for the *CofE* sample are 0.74% (0.41%). The Full Returns sample is more dispersed, more positively skewed and more leptokurtic than the *CofE* sample. With regard to dispersion, the standard deviation 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

---

[22] In contrast to requiring *CofE* data, the returns requirement does not lead to a non-random returns sample. Across the 402 sample months, the average KS statistic comparing our reference sample with the raw CRSP returns universe is 0.0044 (average *p*-value = 0.86).

[23] Because all tests are performed on the month-specific cross section, using realized returns instead of excess returns yields equivalent regression and correlation coefficients. We perform all tests using excess returns and do not further discuss raw returns.

[24] We calculate the monthly *CofE* as (1+ annual *CofE*)$^{(1/12)}$ - 1.

[25] 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.

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 0.6034 (a perfectly symmetric distribution has zero skewness). Finally, the Full Sample returns are leptokurtic, with a kurtosis coefficient of 82.75 on average, while the average *CofE* sample kurtosis is 6.16.

4.2 Test Methodology

To exploit the richness of the cost of capital setting we use both tests based on *CofE* metrics and tests of factor betas. Our main focus is on the *CofE* association tests, for which we estimate cross-sectional Pearson correlations and slope coefficients from regressions of realized (excess) returns on each of the five *CofE* metrics.[26] Benchmark correlations and slope coefficients are estimated for each Month *t* using all complete returns-*CofE* observations available for that month.

$$R_{i,t} - R_{f,t} = \delta_{0,t} + \delta_{1,t} CofE_t + \varepsilon_{i,t} \quad (6)$$

The averages of the monthly coefficient estimates $\delta_{1,t}$ over the sample period are our measures of association between realized excess returns and a specific *CofE* metric. Following Fama and MacBeth (1973), the test statistic for the significance of the associations is the average monthly coefficient estimate, relative to the time-series standard error of the monthly estimates over the sample period.[27]

Table 3 reports results for the unmodified *CofE* sample resulting from list-wise deletion when data on one of the *CofE* metrics is missing (analogous to Easton and Monahan 2005). Correlations show either no reliable relation between realized returns and *CofE*, in the case of VL *CofE*, or a reliably negative relation between returns and *CofE*, for the other four *CofE* metrics (t-statistics range from -0.52 to -6.11). Regression

[26] Intercepts are included in all regressions, but not tabulated for brevity.

[27] 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 excess returns to the standard deviation of the *CofE* estimate. Using the average results in Panel B of Table 2, this ratio ranges from 12.9 (VL *CofE*) to over 25 (GLS *CofE*). We use the correlation 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 magnitude of the correlation. In Section 4.4.2, we report both correlation and regression coefficients.

coefficients show a similar picture, with three of the five metrics showing significantly (at the .05 level or better) negative slope coefficients. All five coefficients are significantly different from their theoretical value of 1, with t-statistics ranging between -7.25 and -12.36. We view these results as broadly consistent with prior literature on the association between *CofE* estimates and realized returns, if not more negative.[28]

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

$$R_{i,t} - R_{f,t} = a_{i,t} + b_{i,t}^{F} F_t + \varepsilon_{i,t} \quad (7a)$$

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 (7a) is estimated over a rolling 12-month window ending in Month *t*.[29]

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 $\widehat{b_{i,t}^{F}}$ (the risk factor betas) obtained from estimating Equation (7a):

$$R_{i,t} - R_{f,t} = \gamma_{0,t} + \gamma_t^{F} \widehat{b_{i,t}^{F}} + \vartheta_{i,t} \quad (7b)$$

Equation (7b) is estimated for each Month *t*. The full sample tests use all firms with the necessary observations to estimate first stage betas. The second-stage coefficient estimates (γ_t^{F}) are interpretable as implied risk factor premia in Month *t* (implied by the first-stage loadings). Following Fama and MacBeth (1973), the test statistic for the significance of the implied risk factor premia 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 (the magnitudes of the implied factor premia). Following previous research, we test whether the γ_t^{F} estimates are reliably different from zero.

---

[28] While prior research has mostly used annual data, we use monthly versions of the *CofE* estimates because asset pricing tests reported 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.

[29] For the time-series regressions given by Equation (7a), we use the more common specification with excess returns to estimate factor betas. As all association test results are averages from cross-sectional estimations, using returns or excess returns is equivalent.

4.3 Non-randomness of the *CofE* sample

We use the samples described in Table 2 to establish benchmark associations between excess returns and factor betas. Our tests on factor betas are motivated by the idea that *CofE* metrics, like factor betas, are supposed to capture risk, and the fact that tests on factor betas can be performed on both the reference sample and on perfect subsamples, such as the *CofE* sample. Table 4, column 1, shows the second stage coefficient estimates from Equation (7b) and t-statistics based on the time-series standard error of the monthly estimates. Our interest is not in the significance of specific risk factors, but rather in using the Full Sample results as a benchmark for comparing subsample results. The association between returns and market beta is positive, with a coefficient estimate corresponding to a risk premium of 0.52% per month (t-statistic = 2.03), and a similar result for the *AQfactor* beta: risk premium of 0.77% per month (t-statistic = 2.44). The coefficient on the *SMB* beta is 0.0025, t-statistic = 1.69, significant at the .05 level, one-tailed. The association between returns and *HML* beta is not reliably different from zero at the 0.05 level.[30]

As previously described, we aim to shed light on how differences in the distributional properties of estimation samples of realized returns, and, by implication, how differences in sample selection criteria, affect the results of association tests between realized returns and both risk factor betas and *CofE* estimates. We first consider 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. With regard to sample size, there is a monthly average of 6,122 firms in the Full Returns sample and 955 firms in the *CofE* sample, a reduction of about 84%. With regard to distributional properties, as captured by dispersion, skewness and kurtosis, the Full Returns sample is more extreme on all three distributional properties.

Because factor betas are available for *both* the Full Returns sample *and* the *CofE* sample, asset pricing tests can be used to illustrate that the *CofE* sample differs from the Full Returns sample with respect to the

[30] Prior research using firm-specific tests, as opposed to portfolio tests, also finds unexpected results for the *HML* factor. For example, in their firm-specific tests in Table 4, Panel D, Core, et al. (2008) document a negative (sometimes weakly significant, sometimes insignificant) relation between the *HML* factor beta and realized returns. Similarly, Gagliardini, et al. (2014) show a significantly negative *HML* premium (their Tables 1 and 2). In portfolio designs (e.g., tests on size/book-to-market portfolio returns), the sign on the *HML* factor betas is generally positive in prior literature.

association between realized returns and risk proxies. To illustrate the effects of sample size, columns 2 and 3 of Table 4 presents results of association tests between risk factor betas and realized returns from 1,000 randomly drawn subsamples of the Full Returns sample ("Random Subsample") of the same size each month as the actual *CofE* sample (a monthly average of 955 firms).[31] Column 2 reports average slope coefficients and t-statistics, Column 3 contains the range of values across the 1,000 random draws. The coefficient estimates of the Full Returns sample (column 1) and the average random subsample (column 2) are nearly identical (differences are between 1 and 4 basis points); the Full Returns results fall well into the range of values (column 3). The reduced sample size means the monthly coefficients are estimated with less precision; as expected, the time-series t-statistics are lower by amounts between 0.08 (market risk premium) and 0.13 (*AQFactor* premium). These results suggest that sample size alone has minimal effect.

Turning to the effects of non-randomness, results of the asset pricing tests using the actual *CofE* sample are shown in the rightmost column of Table 4. None of the factor betas evidences a significant association with excess returns, and all factor premia are reduced in magnitude by at least 50%. Except for the insignificant implied *HML* premium, the results from the *CofE* sample fall outside the range of values spanned by the random subsamples.

We draw three conclusions from the results in Table 4. First, even substantial reductions in sample size (84% in the average cross section) have a modest effect on the efficiency of the estimation. Second, distributional differences in either realized returns or 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 purposes of testing associations with proxies for risk. Third, in such non-random samples, if factor betas fail to load significantly, insignificant results concerning *CofE* metrics should not be surprising.

[31] The Random Subsample results in column 2 of Table 3 are based on averages of 20 random subsamples drawn from the Full Returns sample.

4.4 Distribution matching on simulated, *CofE*-calibrated data and on empirical data

In this section, we use simulated data (Section 4.4.1) and archival data (Section 4.4.2) to demonstrate that distribution-matching can reduce or even eliminate the bias in correlation estimates from non-random samples.

4.4.1 CofE-calibrated simulations

We first show results of simulations (Table 5), where the data are intended to approximate the size and shape of the distribution of the reference sample excess returns and the Value Line *CofE* metric. By using empirically determined parameters of the actual distribution of excess returns and the actual distribution of a *CofE* metric, we create simulated data similar to the archival data. As detailed in the Appendix, these simulations are able to mimic the first four moments of the variable distributions. We induce correlations of 0.25, 0.10 and 0 and apply our distribution matching approach to the simulated data.

Table 5 reports the results for 'Non-Random Sample I', where the selection probability decreases in the absolute distance to the variable mean. The first and second lines of Table 5 show that, for the simulated Full Returns data (first line) and for random samples of the same size as the actual *CofE* sample (second line), estimated correlations are nearly identical to the induced correlations in the population. This finding buttresses our conclusion that even sharply diminished sample sizes do not obscure or shift estimates away from true correlations in the data, as long as the samples are randomly drawn from the reference sample. In contrast, the third line of Table 5 indicates that for a non-random sample constructed to be less extreme than the reference sample, the KS test statistic rejects similarity of the distributions at better than the .0001 level. Estimated associations between the simulated returns and simulated *CofE* metrics are negative and highly significant even though true correlations are positive. Specifically, when the true correlation is .25 (.10), the estimated correlation is -0.17 (-0.06), with a t-statistic of -61.8 (-33.9). When the true correlation is zero, the estimated correlation for the non-random sample is also zero (point estimate 0.0004, t-statistic 0.23). When we distribution-match the non-random sample, the KS test statistics decline to about .03 with significance level of about 0.50. When the true correlation is .25 (.10) [0], the estimated correlation is .17 (.060) [.00], with a t-statistic of 25.40 (9.07) [0.30].

We believe these simulations support two conclusions. First, sample marginal distributions, not sample size *per se*, affect the ability to empirically detect the true correlations between two variables. In particular, the sign differences in the correlations reported in Table 5 highlight the potential bias in results when the marginal distribution is non-randomly drawn from the reference sample. Second, despite the extreme difference in the characteristics of marginal distributions, it is possible to resample systematically from the non-random sample to create a distribution-matched sample with correlations similar in sign and magnitude to the true correlations.

4.4.2 Distribution matching the actual CofE sample

We now construct distribution matched samples from archival data of returns and the five *CofE* metrics. Results so far suggest the returns of the actual *CofE* sample are a non-random sample of the Full Returns sample returns, and Table 2 shows excess returns for the *CofE* sample have a similar mean/median, and noticeably smaller standard deviation, skewness and kurtosis, as compared to the Full Returns sample. We interpret these findings as raising the question of whether the negative or weak correlation between *CofE* estimates and realized returns reported in Table 3 is generalizable to the reference sample or arises from the effects of data requirements.

To accommodate the potentially severe truncation of the returns distribution in the *CofE* sample, we modify the implementation of the distribution-matching approach used in the simulation in two ways. The first implementation uses an iterative procedure that starts with a base sample, draws an additional observation, and keeps that additional observation only if the resulting sample shows a smaller KS statistic. The approach still aims to minimize the KS-based statistic, even in months where insignificant KS statistics cannot be achieved because of large initial differences between the *CofE* sample and Full Returns sample.[32] The minimization does not require a pre-specified sample size, but rather lets the iteration determine the optimal sample when the KS statistic cannot be further minimized. This approach is conceptually grounded but inefficient and computationally burdensome for large samples. The second, less computationally demanding implementation

[32] The non-parametric KS statistic captures any difference between two distributions, not limited to the first four moments.

matches the *CofE* sample to the Full Returns sample using a variant of stratified resampling (post-stratification) which tries to match the dispersion of the returns distribution. We detail both approaches next.

Method 1: Kolmogorov-Smirnov-based distribution match. We start by randomly sampling either 20% of the month-specific sample or 100 unique firms from the *CofE* sample in a given month. We compute the KS statistic for this initial draw against all returns from the reference sample in that month.[33] The KS 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 observation (# 101), re-compute the KS statistic and again compare to the reference distribution of returns that month. If the KS statistic using 101 observations (against the reference distribution) is equal to or greater than the KS statistic using the original 100 observations sample (against the reference distribution), we dismiss the 101st observation and replace it with another randomly chosen, with replacement, 101st observation from the *CofE* sample. If the KS statistic using 101 observations is lower than the KS statistic using the 100 observations sample, we keep the 101st observation, draw a 102nd observation, and evaluate the inclusion of the 102nd observation. We repeat this step 1,000 times, thereby allowing for KS-based distribution matched samples to increase by a maximum of 1,000 observations each month.[34] Because convergence to a minimum KS statistic depends both on the initial 20% (or 100) observations drawn and on the order of additions, 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. When the iteration begins 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

[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 a mean of zero before computing 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 distribution of the reference sample.

month), with a time-series standard deviation of 54 firms. When the iteration begins with 100 firms each month, the average cross section consists of 124 firms (with a standard deviation of 14 firms).[35]

Method 2: Description of bin-based distribution match. To reduce the computational burden of Method 1, we create bin-based distribution matched samples. Bin-based matching is similar to post-stratification but differs because the distribution is also truncated (some population strata are not represented in the sample), requiring an additional weighting scheme for the tails of the distribution.

For the common support region, we first place the returns of the Full Returns sample and the *CofE* sample into bins with width of 100 basis points (bp) and calculate the sample proportion of observations in each bin for both samples.[36] 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.[37] At the extremes of the reference sample distribution, we encounter bins without corresponding observations in the *CofE* sample. To address this issue, 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):

$$w_{i_{CofE}} = \begin{cases} w_{i_{RS}} \sum_{j=\min(i_{RS})}^{\min(i_{CofE})} \left[\min(i_{CofE}) - i_{j,RS} + 1\right]^{\gamma} & \forall \quad i_{RS} = \min(i_{CofE}) \\ w_{i_{RS}} & \forall \quad \min(i_{CofE}) < i_{RS} < \max(i_{CofE}) \\ w_{i_{RS}} \sum_{j=\max(i_{CofE})}^{\max(i_{RS})} \left[i_{j,RS} - \max(i_{CofE}) + 1\right]^{\gamma} & \forall \quad i_{RS} = \max(i_{CofE}) \end{cases} \tag{7}$$

[35] The KS-based distribution matching approach can, in principle, be used to construct multiple subsamples, which can, in turn, be analyzed separately and then aggregated. Such an approach would resemble multiple imputation. The benefit of such an approach might include correctly specified cross-sectional standard errors, which are of little interest in the Fama-MacBeth design we use.

[36] Although the design choices in this bin-based approach are admittedly ad hoc, bin-based sampling approaches are well-documented as well as computationally more efficient.

[37] This approach 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.

$w_{i_{CofE}}$ ($w_{i_{RS}}$) is the sampling proportion of Bin [*i;i*+0.01[in the *CofE* sample and the reference sample, respectively. The product of sampling proportion $w_{i_{CofE}}$ and overall sample size in Month *t* is the bin-specific number of draws that month. We numerically solve, within the sampling procedure, for the constant weight parameter γ until the standard deviation for the distribution-matched *CofE* sample 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 0.0009 (t-statistic = 0.62) at $\gamma = 2.15$, and -0.0008 (t-statistic = -0.56) at $\gamma = 2.20$. Figure 2 provides the intuition for the approach by plotting the average distribution of excess returns for the *CofE* sample, before and after distribution matching, as well as the reference distribution of returns. The dashed (red) continuous distribution of returns in the *CofE* sample is sorted into strata of pre-determined width and then resampled, such that the sample proportion of each stratum is equal to that stratum in the reference distribution. In the graph, the procedure is effective if the heights of light (grey) bars, representing the strata, match the heights of the dark (blue) reference strata.

Association tests using distribution matched samples. 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 6.[38] The average 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 lower. After KS-based distribution matching using either 20% of firms or 100 firms, the average KS statistic is just under 6% for both initializations, with an average *p*-value of 0.5287 (0.7789), and the test rejects similarity of distributions in 42 (5) of 402 months at the 0.10 level or better. We conclude from these results that the KS-based approach to distribution matching is effective.

[38] We also re-estimate the implied factor premia using the realized returns from the KS-based distribution-matched samples and from the bin-based distribution matched sample ($\gamma = 2.20$); results are not tabulated. The implied market factor, *SMB* factor, and *AQfactor* premia are positive and significant at the 0.01 level or better, with t-statistics of 2.86 or higher. The *HML* factor remains insignificant. In sum, the results on implied factor premia in the distribution-matched samples are qualitatively similar to results in the reference sample, as reported in Table 3.

The top portion of Table 6 reports correlations between five *CofE* measures and realized returns, using the KS-based distribution-matched samples (columns 2 and 3) and the bin-based distribution-matched samples (columns 4 and 5). With regard to the former, KS-based matching, 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 *CofE* metric with 100 firms as initial sample where the correlation is positive and insignificant). These results indicate reliably positive associations between four implied *CofE* metrics and realized returns, in contrast to the benchmark results on the unmodified *CofE* sample (reproduced in the first column), where four correlations are significantly negative. With regard to bin-based matching, correlation results are reported between five *CofE* estimates and realized returns, for $\gamma = 2.15$ and $\gamma = 2.20$, where the overall difference in standard deviations is insignificant at conventional levels. Because the focus is on matching standard deviations only, we do not expect the bin-based distribution match to be entirely effective in lowering the KS statistics for general similarity of distributions. Table 6, columns 4 and 5, shows that while the average KS statistic decreases relative to the unmodified *CofE* sample, it remains significant for 372 (375) months for $\gamma = 2.15$ ($\gamma = 2.20$). In these analyses, all five *CofE* measures have reliably positive correlations with realized returns, ranging from 0.021 (the CT *CofE* measure) to 0.054 (the VL measure), with t-statistics between 2.07 (the CT *CofE* measure, $\gamma = 2.15$) and 4.14 (the VL *CofE* measure, $\gamma=2.20$).

The bottom portion of Table 6 contains regression coefficients for the KS-based and bin-based distribution-matched samples. In contrast to the benchmark results reported in Table 3, regression coefficients are generally significantly positive, with t-statistics usually exceeding 2.0, and statistically indistinguishable from 1 in 16 of the 20 specifications. The slope coefficient on the *CT* metric is significantly smaller than 1 in two of the KS-based samples, and the *GLS CofE* metric shows significantly larger coefficients for the bin-based samples.

To examine the effect of distribution matching on the implied factor premia from asset pricing tests, we repeat the Fama-MacBeth-type tests reported in Table 4 using the KS-based distribution-matched samples with the initialization set at 20% of the *CofE* sample (results not tabulated). The factor premia from these samples

are qualitatively similar to the Full Returns sample results in Table 4 in that the market factor, SMB and AQfactor are positive and statistically significant, while the HML factor is insignificant at conventional levels.

We next address a concern that arises, 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 such as beta. In our setting, the concern is that distribution matching the *CofE* sample increases the association between *CofE* metrics and excess returns at the cost of a diminished association between *CofE* metrics and other risk proxies for risk, specifically, risk factor betas. We test for a decline in the associations between *CofE* metrics and risk factor betas, using (1) the sample composition from the KS-based matching in Table 6 with initial sample size equal to 20% of the *CofE* sample, as compared to (2) a random sample from the *CofE* sample of the same size in any given month. For both samples, we regress the five *CofE* metrics on lagged risk factor betas from Equation (7a). If distribution matching decreases the association between *CofE* metrics and risk factor betas, the associations will be smaller for the distribution-matched sample (1) than for the random sample (2). 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 the 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 for 19 of 20 coefficient estimates (five *CofE* metrics times four risk factor betas), differences between the two sets of associations are insignificant at conventional levels, with t-statistics ranging from -0.59 to 1.46 (results not tabulated). The exception is the coefficient on the market beta in the VL *CofE* regression, which shows a small and statistically 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; they are always of the same sign, and always significant at comparable levels.

Combined with previous results, we interpret the weight of the evidence in Table 6 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. We draw three inferences from these effects. First, selection criteria

that yield estimation samples with different returns distributions, as compared to a reference sample, decrease the ability to detect the predicted associations between realized returns and *CofE* estimates. Second, 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. Third, the finding that the distribution-matched sample may be smaller than the original, non-random sample suggests that attempts to achieve generalizability to a reference sample by maximizing the size of a data-restricted sample may not be effective.

4.5 Results from Multiple Imputations

Section 2 notes that, under the missing-at-random (MAR) assumption, an alternative to distribution matching is multiple imputation (MI), which aims to complete an incomplete dataset using values from stochastic regressions. Unlike distribution-matching, which is nonparametric, MI relies on the assumption of multivariate normality.[39] Specifically, we use the expectation maximization (EM) algorithm to determine the distribution of possible parameter values for the imputation of the *CofE* metrics, using the complete excess returns data. Using this solution as a starting point, we use an iterative Markov-chain Monte Carlo approach to draw from that distribution and construct multiple (m=10) completed datasets. Each of the 10 datasets has the size of the Full Returns sample, with the same (complete) excess returns data and a full vector of the *CofE* metrics, consisting of measured and imputed values. These 10 datasets can be analyzed independently, and results aggregated.[40]

We begin by formulating a month-specific imputation model for each *CofE* metric using only the (complete) excess returns data. To improve the fit of the imputation model, we split each monthly cross-section into groups based on percentiles of the returns distribution, allowing for different imputation models (intercepts and coefficients) in each group. We split at the median (g=2 in the Table 7), terciles (g=3) and quintiles (g=5). We stop at quintiles because, with a decreasing number of observations, the initial expectation maximization

[39] As discussed in Section 2, we focus on MI rather than on maximum likelihood estimation because the literature suggests that the results from MI might be less biased when the normality assumption is violated.

[40] Standard statistical software packages like SAS and Stata include commands for performing multiple imputations and for the aggregation of the test results from the imputed datasets. We used the MI procedure and the MIANALYZE procedure in SAS for these functions. The MI procedure also offers several diagnostics to check for convergence of the estimation.

(EM) algorithm for the parameter estimation will lose precision, might fail to converge, or, in the limit, may not be feasible at all due to the sparseness of non-missing data.

Table 7 presents the findings. Both correlation and regression coefficients from full-cross-section imputations (g=1, Column 2) are qualitatively similar to the actual *CofE* sample results. While coefficients tend to be more negative, and hence farther from their theoretical values, standard errors have increased and significance levels have decreased because of additional variance from the imputed *CofE* data. For group-wise imputations (g=2, 3, or 5), the signs of the associations are generally positive; 12 of the 15 regression slopes are significantly positive and indistinguishable from 1. Results for correlations are also generally positive, but weaker and insignificant in eight of the 15 specifications. The CT *CofE* metric in particular shows consistently insignificant associations with excess returns, and regression coefficients significantly lower than 1. We consider the results from these multiple imputation to be broadly consistent with the results from the distribution-matched samples, in that group-wise imputations generally yield the theoretically predicted positive associations between *CofE* metrics and excess returns.

We assess the sensitivity of the Table 7 results in two ways. First, because the basic imputation procedure does not rule out negative imputed values for the *CofE* metrics, we repeat our tests precluding the imputation of negative values by using log transformations before imputing.[41] (The log transformation can also help assess the impact of skewness in the distribution of *CofE* metrics.) Results from these tests, not tabulated, are generally comparable to the tabulated results for two and three imputation groups. When we allow for five imputation groups per cross section, regression coefficients are larger and of higher significance than the ones reported in Table 7. Qualitative inference only changes for the MPEG *CofE*, with a higher coefficient of 0.6537 (t-statistic against 0 = 1.83, t-statistic against 1 = -0.97).

Second, and analogous to the test on the distribution-matched samples, we validate the approach for our setting by imputing data for the factor tests. We construct a dataset that deletes the loadings estimates from (7a) for observations without *CofE* metrics. Equivalent to the *CofE* metrics in the main tests, we impute the now

[41] Random inspection suggests that the incidence of negative imputed *CofE* values is small in the average cross-section.

missing (by construction) values for the loadings before we estimate the implied factor premia using (7b). The results (untabulated) show that for imputations of the full cross section (g=1), implied factor premia are generally insignificant, with only the market risk premium significantly positive at t=1.92. For the percentile-wise cross-sectional imputations (g=2, 3, or 5) results are qualitatively equal to the results from the Full Returns sample, insofar as the market, size and *AQ* factors are significantly positive, while the book-to-market factor remains insignificant across all specifications.

We interpret the weight of the evidence as suggesting that multiple imputation can be a viable alternative to distribution matching, albeit one that imposes additional assumptions, specifically, normality of the data, and that may require additional adjustments in a specific research setting, e.g., precluding inadmissible imputed values.

5. Extensions

5.1 Relation to selection models

In this section, we clarify the relation between distribution matching and an alternative technique for dealing with missing data, Heckman-type selection models. Both selection models and distribution matching seek to incorporate information into the test model that goes beyond the information in the subsample with complete data, but the approaches differ in the kind of information incorporated. Distribution matching focuses on the outcome variable only, whose empirical distribution in the reference sample can either be derived by the researcher or is empirically estimable. The goal is to construct a test sample that appears randomly selected with respect to the outcome variable (whose reference distribution is known). In contrast, a selection model operates under the assumption that data are *not* missing at random, conditional on observed data, thus requiring explicit modelling of the missingness mechanism. This model requires additional explanatory variables, which often impose similar or even more stringent data restrictions than the actual test model. Thus, 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.[42] Consequently, results from a Heckman model are generalizable only to the sample for which the selection model variables are available.

In addition, increasing the fit of the selection model by including more explanatory variables is likely to impose increasingly stringent sample restrictions due to data requirements. Adding to the severity of the data availability problem is the exclusion restriction on the explanatory variables set in the test model compared to the explanatory variables set in the selection model. To avoid collinearity of the test model variables and the inverse Mills ratio, the recommended approach is to include at least one additional variable in the selection model not contained in the test model of interest and, in theory, not associated with the outcome variable.[43]

Distribution matching is, by design, non-parametric and based on a reference distribution of the outcome variable; its application is not limited to, for example, normally distributed variables. In contrast, like multiple imputation, the derivation of the Heckman correction for sampling biases relies on the assumption that the residuals from the selection model 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 normality assumption is crucial for the parameter estimates in the test model; the descriptive statistics for excess returns reported in Table 2 cast doubt on this assumption in our setting. At a minimum, we caution that Heckman test results with realized returns as the dependent variable are likely biased (in an unknown direction) by violations of the normality assumption.

Despite these concerns about applicability in our setting, 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. 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

---

[42] 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 4).

[43] Lennox, et al. (2012) illustrate the sensitivity of even qualitative test results to selection model specification.

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, calculated from the Compustat annual file, and the four univariate risk factor betas from the asset pricing regressions (7a).[44]

Table 8 reports semi-partial correlation coefficients between *CofE* metrics and excess returns and goodness-of-fit measures for the probit selection models estimated. The selection model including only size has a pseudo-R^2 of 0.48, with no additional sample loss; adding more variables increases the pseudo-R^2 to a maximum of 0.52, with a sample loss of 2.1% when the model includes the log of the book-to-market ratio. The reported semi-partial correlations are averages of the 402 month-specific (cross-sectional) semi-partial correlations, obtained by controlling for the inverse Mills ratio in the respective *CofE* metric first, then computing the correlation between the returns and the residualized *CofE* metric. Similarly, regression coefficients (bottom half of the table) are averages of 402 cross-sectional slope coefficients from regressions of excess returns on both the CofE metric in question and the inverse Mills ratio from the selection model.

The inverse-Mills ratio-adjusted semi-partial correlations and the adjusted the regression coefficients are negative or, in the case of the VL *CofE* metric, indistinguishable from zero. Across *CofE* metrics, point estimates appear slightly lower, and are more statistically different from zero, compared to unadjusted correlations or slopes. However, we again caution that the effects of the inverse Mills ratio on the semi-partial correlations and regression coefficients might be due to violations of the normality assumption, an inadequate fit of the selection model, or some combination of the two. We conclude that, in our setting, Heckman-type selection models do not change the conclusion from results obtained using the unadjusted *CofE* sample.

As a second test of the effectiveness of including an inverse Mills ratio, we use a similar adjustment in the factor beta regressions for the actual *CofE* sample, aiming to restore factor premia obtained from the Full Returns sample (results not tabulated). Specifically, we use the variables in selection Model IV and re-run the cross-sectional asset pricing tests using a factor beta and the inverse Mills ratio. When we include the inverse

[44] 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, sample losses for selection models that include volume are largely due to that earlier period, while coverage afterwards is almost complete.

Mills ratio, factor premia estimates from the *CofE* sample are hardly affected (differences range from -0.0004 to -0.0001), insignificant at conventional levels, and qualitatively different from the Full Returns sample results.

5.2 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 and the subsample of those returns associated with firms for which *CofE* measures can be calculated. We have analyzed how differences in returns distributions between the two samples affect results of association tests. In this section, we consider whether results of asset pricing tests of the association between risk factor betas and realized returns are sensitive to other selection criteria that likely yield non-random samples in applied research settings. The cross-sectional sample selection criteria we consider are S&P 500 membership, a potential screen in compensation research;[45] NYSE listing, a screen in some intra-day trading studies; the availability of the standard deviation of analysts' earnings forecasts, required for research examining forecast dispersion; price at least $5. We apply each criterion separately to the Full Sample and re-estimate Equation (7b), separately, for observations meeting and not meeting the criterion. We also report the proportions of firms that do and do not meet each sample selection criterion.

Results are reported in Table 9, Panel A. The four sample selection criteria generally 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). The KS statistics for tests of equality of distributions show that for three of the four selection criteria, percentage deviations between the Full Returns sample and the subsample meeting the criterion exceed the deviations for the subsample not meeting the criterion. Stated differently, the returns distributions of firms *not* selected by

[45] 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.

these three criteria more closely resemble the returns distribution of the Full Returns sample.[46] The exception is the price at least $5 criterion.

These findings suggest that asset pricing tests may yield results that are more theory-consistent, as well as more consistent with results for the Full Sample, for firms *not* included in the sample resulting from the application of plausible selection criteria. Specifically, Table 9 shows that both point estimates and t-statistics are more similar to Full Sample results for firms that do *not* meet the selection criteria. We view these results as indicative, but not dispositive, that the distributional issues we have identified and analyzed for the *CofE* sample generalize to other research situations where data-constrained samples consist of large, stable firms, and as a consequence, have returns distributions that are not random draws from the population.

5.3 Association tests between realized returns and factor betas using forced non-random samples

To illustrate the effects of a direct and extreme form of non-random sampling based on returns, we split the distribution of realized returns into positive and negative returns, and reweight both subsamples of the Full Returns sample differentially. This procedure is intended to shed light on how much the sampling weights of positive or negative returns alter qualitative conclusions from association tests, as compared to conclusions based on the Full Sample. A similar and perhaps less extreme reweighting of positive and negative returns may arise implicitly in a given research setting, through the data requirements for a variable of interest. 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 asset pricing factors.

Results are reported in Table 9, Panel B. The center column, labeled 0%, shows the results of association tests when we resample while preserving the population proportions of positive and negative returns. These results coincide with the Full Returns sample results as reported in Table 3; small differences result from sampling with replacement as opposed to drawing the full sample. The columns to the left, labeled -2.5%, -5%,

[46] Average mean excess return, standard deviation and skewness differ between the subsamples. Specifically, the subsamples that do not meet the sample selection criteria have larger average mean excess returns, larger standard deviations of excess returns and greater positive skewness of excess returns (results not tabulated).

-10% and -25%, show the Full Returns 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 suggest that changing the proportion of positive returns increases the significance of results of asset pricing tests.[47] For example, the t-statistic on the implied *SMB* factor premium increases from 0.38 (25% decrease in positive returns) to 1.67 (unbiased sample) to 2.98 (25% increase in the proportion of positive returns). Factor premia are differentially sensitive to these changes, with the market factor apparently more robust compared to other factors, although the trend exists also for it. We infer that results of association tests are sensitive to the distributional properties of estimation samples, and therefore sensitive to differences in sample selection criteria, with the degree of sensitivity differing with the nature of the selection criteria.

Our findings imply that analyses showing weak or no associations between realized returns and risk factor betas for samples obtained from applying selection criteria may underestimate the actual strength of these associations, depending on how the application of the criteria affects the returns distribution as a whole.[48] 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, the results in Table 9 Panel B indicate that the outcome of association tests is sensitive to characteristics of the returns distribution, such as whether the sample

[47] Recall that the *HML* beta is negative in our firm-specific setting, consistent with other studies using firm-specific returns (e.g.,Gagliardini, et al. 2014). Consequently, its t-statistic becomes more negative as the proportion of positive returns increases.

[48] We interpret 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 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 *expected* returns. To the extent 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 identify conditions that are informative about the construct validity of realized returns as a proxy for expected returns.

contains more positive returns. The characteristics may in turn be influenced by researcher-chosen selection criteria or sample partitions that shift the distribution of returns.

6. Conclusions

This paper examines a generic form of non-randomness through missing data, such as created by stringent data requirements, and describes, validates and illustrates a resampling distribution-matching technique that can be used to align the distribution of a non-random estimation sample with that of a reference sample to which the researcher would like to generalize. The foundation for this approach is resampling from the data-restricted non-random subsample to minimize the distance between the marginal sample distribution and the marginal reference distribution.

Our illustration of the distribution-matching approach in a specific empirical-archival application, associations between returns and implied cost of equity (*CofE*) estimates, is of interest in its own right given the practical and theoretical importance of the association and weak and mixed results in previous research. We show that associations between realized returns and five popular *CofE* metrics are influenced by the properties of the realized returns distribution used to estimate the associations. We use the CRSP population of firms with at least 12 consecutive monthly returns during 1976-2009 as the reference sample, and compare results from this sample with those obtained using the sample of firms with sufficient data to calculate the *CofE* measures. The latter sample is a substantially smaller and non-random subsample of the former. After distribution matching, so the resulting returns distribution mimics the returns distribution in the reference sample, we find reliably positive correlations between realized returns and most *CofE* measures, as predicted by theory. This result suggests that several implied *CofE* measures used in the accounting literature have greater construct validity than prior results suggest.[49]

[49] We emphasize that we implement the *CofE* metrics as originally developed. The fact that the metrics are positively correlated with realized returns in our distribution-matched sample does not mean they cannot be improved upon, either by developing new metrics altogether, by adjusting input variables, or by developing alternative empirical implementations of these metrics. For a thorough discussion and analysis, see Easton (2007).

Viewed broadly, 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 demonstrate how to use available information about a marginal reference distribution of one variable of interest (in our setting, realized returns) to construct samples that mimic a reference distribution more closely than can an unmodified sample whose composition is dictated by data requirements. Our demonstration highlights the important trade-off between a random research sample and maximizing the size of the research sample under data constraints. Viewed broadly, our analysis suggests maximization of a data-constrained sample may not be goal-congruent with increasing the generalizability from such a sample.

We also discuss and illustrate both multiple imputation (which performs well as a potential alternative to distribution matching in our setting, albeit at the cost of additional assumptions) and selection-type models (which do not perform well in our setting). Given their flexibility, we believe that distribution matching or multiple imputation can be applied in a variety of research settings with data-constrained, non-random samples.

Our findings suggest that researchers might benefit, in terms of increasing the generalizability of their results, from examining the impact of data requirements on the empirical distribution function of the test model variables, in particular the variable whose distribution is most affected by the availability of other variables of interest. We hope that the approaches taken here—possibly modified to suit the specific research context—will help future work by providing an explanation for weak or counter-intuitive initial results from restricted samples. Distribution matching might also benefit future research by helping to coordinate across studies that address either similar questions using different samples, or tries to build on prior work that used different samples. To the extent many researchers can define and construct the same reference sample, comparisons of results across studies are facilitated.

Appendix - Simulation Design

A.1. 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 σ_{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. We follow these steps:

1. We choose the desired first four moments of the distributions of *y* and *x* (subscripts suppressed).[50]
2. Following Fleishman (1978), we transform variables *y* and *x* with known skewness and kurtosis into normally distributed variables:

$$a = -c_a + b_a z_a + c_a z_a^{\,2} + d_a z_a^{\,3} \quad \text{with} \quad a \in \{x, y\} \tag{A1}$$

   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 an intermediate correlation for two variables that are normally distributed, again using an iterative numerical procedure (following Vale and Maurelli, 1983).[51]
4. We 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 2×2 Cholesky decomposition of the intermediate correlation matrix.[52] To introduce non-normality, we apply the b_a, c_a and d_a transformation coefficients from Step 2, multiply by the desired standard deviation, and add the desired mean. The resulting variables exhibit the pre-specified (original) level of correlation as well as the pre-specified first four moments for both variables.

---

[50] The non-parametric KS statistic we use to measure differences in distributions is not limited to capturing differences in the first four moments only.

[51] For non-zero higher moments, there are generally four solutions to the polynomial in (A2). We perform a grid search over the parameter space to identify all four solutions for b_a, c_a and d_a. We retain the solution with the lowest absolute distance between the intermediate correlation and the desired correlation.

[52] The Cholesky decomposition of the positive-definite and symmetric matrix **A** returns the matrix **U** that satisfies **A** = **U**'**U**.

We generate 1,000 populations with two standard normal variables (Panel A of Table 1) 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.[53] We repeat the procedure for true correlation levels of 0.5, 0 and -0.5.

To generate random and non-random subsamples from these populations, we first draw 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.

For the first non-random sample ('Non-Random Sample I' in Table 1), the selection probability of Observation i is decreasing in the ranked absolute distance from the mean:

$$\text{Prob}_i = \frac{A_i}{\sum_{i=1}^{m} A_i} \times m \text{ where } A_i = rank\left[\frac{1}{abs\left(y_i - \bar{y}\right)}\right] \quad \text{(A2)}$$

This type of non-random sampling selects, with increasing likelihood, relatively stable observations ("firms") with y_i relatively closer to $\bar{y}$. We use the absolute value of the distance that favors neither positive nor negative values of y_i. Our goal is to capture the effects on the distribution of realized returns by over-selecting large, stable firms.[54] Because there is no measure of firm size in these simulations, we proxy for an equal-weighted "market" using the cross-sectional mean, $\bar{y}$. 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

[53] 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.

[54] For example, 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.

interval [y_{min}; y_{max}], with higher selection probabilities for observations in the tails. We consider this 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 resampling with replacement. Depending on the realizations 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, such that the selection probability is strictly increasing in y. In implementing this sampling, we replaced $\bar{y}$ with y^{max} in Equation (A2). 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.

A.2. Distribution-Matching of Non-Random Samples

For each of the three non-random samples drawn from the simulated population, we resample "distribution-matched" samples. A distribution-matched sample is derived from its corresponding non-random sample, and can therefore consist only of those non-random observations. The distribution-matched sample is constructed by resampling with replacement, such that the empirical cumulative distribution of y_i in the non-random sample, $F^{NRS}(y_i)$, mimics the empirical distribution of y_i in the population, $F^{POP}(y_i)$. 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 $F^{NRS}(y_i)$ and $F^{POP}(y_i)$.

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 $F^{NRS}(y_i)$ and $F^{POP}(y_i)$ 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 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. Specifically, in these simulations, the non-random sample of $m = 1{,}000$ observations is first sorted on y_i. Observation $i = 1$ will have $y_i = y^{min}$, with $F^{NRS}(y^{min}) = 1/m = 0.001$, with $F^{POP}(y^{min})$ likely to be different. Observation i will be resampled $mF^{POP}\left(y^{\min}\right)$ times to mimic the empirical cumulative distribution function in the population at y^{min}. More generically, observation i of the non-random sample is resampled $m\left(F^{POP}\left(y_i\right) - F^{POP}\left(y_{i-1}\right)\right)$ times. The observation $i = m$, with $y_i = y^{max}$, is drawn $m\left(1 - F^{POP}\left(y_{i-1}\right)\right)$ times.[55]

[55] Because of rounding the number of draws to full integers, 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. Results are not sensitive to the inclusion of this second random draw.

References

Bhattacharya, N., F. Ecker, P. Olsson and K. Schipper. 2012. Direct and mediated associations among earnings quality, information asymmetry, and the cost of equity. *The Accounting Review* 87, 449-482.

Botosan, C. 1997. Disclosure level and the cost of equity capital. *The Accounting Review* 72, 323–350.

Botosan, C., and M. Plumlee. 2002. A re-examination of disclosure levels and expected cost of equity capital. *Journal of Accounting Research* 40, 21-40.

Botosan, C., and M. Plumlee. 2005. Assessing alternative proxies for the expected risk premium. *The Accounting Review* 80, 21‐54.

Botosan, C., M. Plumlee and X. Wen. 2011. The relation between expected returns, realized returns, and firm risk characteristics. *Contemporary Accounting Research* 28, 1085-1122.

Brav, A., R. Lehavy, and R. Michaely. 2005. Using expectations to test asset pricing models. Financial *Management* 34, 31‐64.

Brookman, J., T. Jandik and C. Rennie. 2006. A comparison of CEO compensation data sources. Working paper, University of Nevada and University of Arkansas.

Campbell, J. 1991. A variance decomposition for stock returns. *Economic Journal* 101: 157-179.

Campbell, J., and R. Shiller. 1988. The dividends-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195-228.

Claus, J., and J. Thomas. 2001. Equity risk premium as low as three percent? Evidence from analysts' earnings forecasts for domestic and international stocks. *Journal of Finance* 56, 1629‐1666.

Core, J., W. Guay and R. Verdi. 2008. Is accruals quality a priced risk factor? *Journal of Accounting and Economics* 46, 2-22.

Dai, Z., D. Shackelford, H. Zhang and C. Chen. 2013. Does financial constraint affect the relation between shareholder taxes and the cost of equity capital? *The Accounting Review* 88, 1603–1627.

Demirtas, H., S. Freels and R. Yucel. 2008. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation,* Vol. 78 (1): 69–84.

Dhaliwal, D., L. Krull, and O. Li. 2007. Did the 2003 tax act reduce the cost of equity capital? *Journal of Accounting and Economics* 43, 121–150.

Easton, P. 2004. PE ratios, PEG ratios, and estimating the implied expected rate of return on equity capital. *The Accounting Review* 79, 73‐96.

Easton, P. 2007. Estimating the cost of capital implied by market prices and accounting data. Foundations and Trends in Accounting 2, 241–364.

Easton, P., and S. Monahan. 2005. An evaluation of accounting-based measures of expected returns. *The Accounting Review* 80, 501-538.

Elton, E. 1999. Expected return, realized return, and asset pricing tests. *Journal of Finance* 54, 1199‐1220.

Enders, C. 2010. Applied missing data analysis. Guilford Press, New York.

Fama, E., and K. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3-56.

Fama, E., and J. MacBeth. 1973. Risk, return and equilibrium: empirical tests. *Journal of Political Economy* 81, 607-636.

Fleishman, A. 1978. A method for simulating non-normal distributions. *Psychometrika* 43, 521-532.

Francis, J., R. LaFond, P. Olsson and K. Schipper. 2004. Costs of equity and earnings attributes. *The Accounting Review* 79, 967-1010.

Francis, J., R. LaFond, P. Olsson and K. Schipper. 2005. The market pricing of accruals quality. *Journal of Accounting and Economics 39,* 295-327.

Gagliardini, P., E. Ossola and O. Scaillet. 2014. Time-varying risk premium in large cross-sectional equity datasets. Working paper.

Gebhardt, W., C. Lee, and B. Swaminathan. 2001. Towards an ex-ante cost of capital. *Journal of Accounting Research* 39, 135‐176.

Gerakos, J., and R. Gramacy. 2013. Regression-based earnings forecasts. University of Chicago working paper.

Gode, D., and P. Mohanram. 2003. Inferring the cost of capital using the Ohlson-Juettner model. *Review of Accounting Studies* 8, 399-431.

Guay, W., Kothari, S., Shu, S. 2011. Properties of implied cost of capital using analysts' forecasts. *Australian Journal of Management* 36, 125-149.

Hail, L., and C. Leuz. 2006. International differences in the cost of equity capital: Do legal institutions and securities regulation matter? *Journal of Accounting Research* 44, 485–531.

Hecht, P., and T. Vuolteenaho. 2006. Explaining returns with cash-flow proxies. *Review of Financial Studies* 19, 159-94.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47, 153-161.

Hou, K., M. VanDijk, Y. Zhang. 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics* 53, 504-526.

Hribar, P. and P. Jenkins, P. 2004. The effect of accounting restatements on earnings revisions and the estimated cost of capital. *Review of Accounting Studies* 9, 337–356.

Lakonishok, J. 1993. Is Beta dead or alive? In: the CAPM controversy: policy and strategy implications for investment management. New York, NY: Association for Investment Management and Research.

Lennox, C., J. Francis and Z. Wang. 2012. Selection models in accounting research. *The Accounting Review* 87 (2), 589-616.

Li, K., and P. Mohanram. 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies*, forthcoming.

Little, R., and D. Rubin. 2002. Statistical analysis with missing data. 2nd edition. Wiley & Sons, New Jersey.

Ogneva, M. 2012. Accrual quality, realized returns, and expected returns: the importance of controlling for cash flow shocks. *The Accounting Review* 87, 1415-1444.

Ohlson, J., and B. Jüttner-Nauroth. 2005. Expected EPS and EPS growth as determinants of value. *Review of Accounting Studies* 10, 349‐365.

Rubin, D. 1987. Multiple imputation for nonresponse in surveys. Wiley & Sons, New Jersey.

Schafer, J. 1997. Analysis of incomplete multivariate data. Chapman and Hall, Boca Raton.

Tobin, J. 1956. Estimation of relationships for limited dependent variables. Econometrica 26, 24-36.

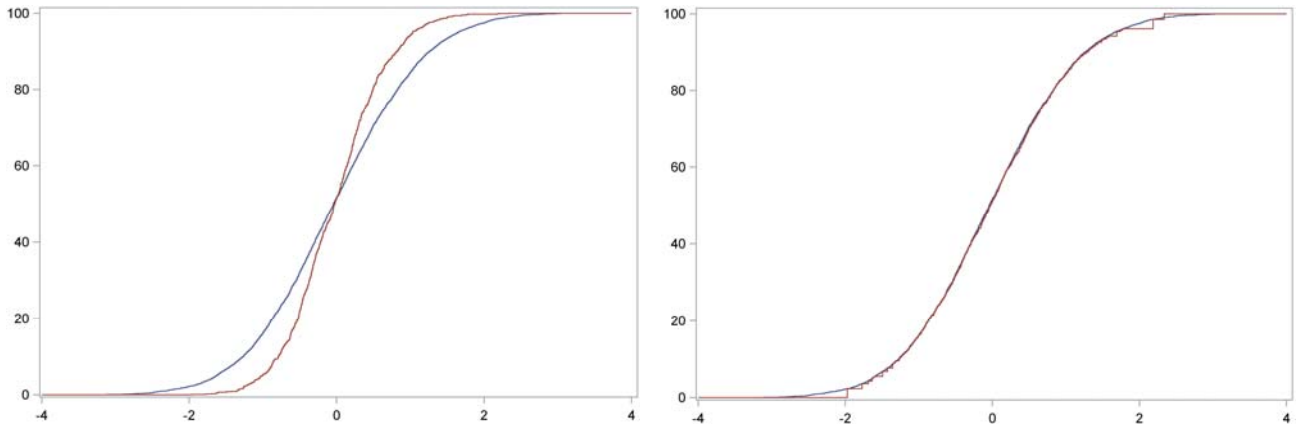
Vale, C., and V. Maurelli. 1983. Simulating multivariate nonnormal distributions. *Psychometrika* 48, 465-471.

Vuolteenaho, T. 2002. What drives firm-level stock returns? *Journal of Finance* 57, 233‐264.

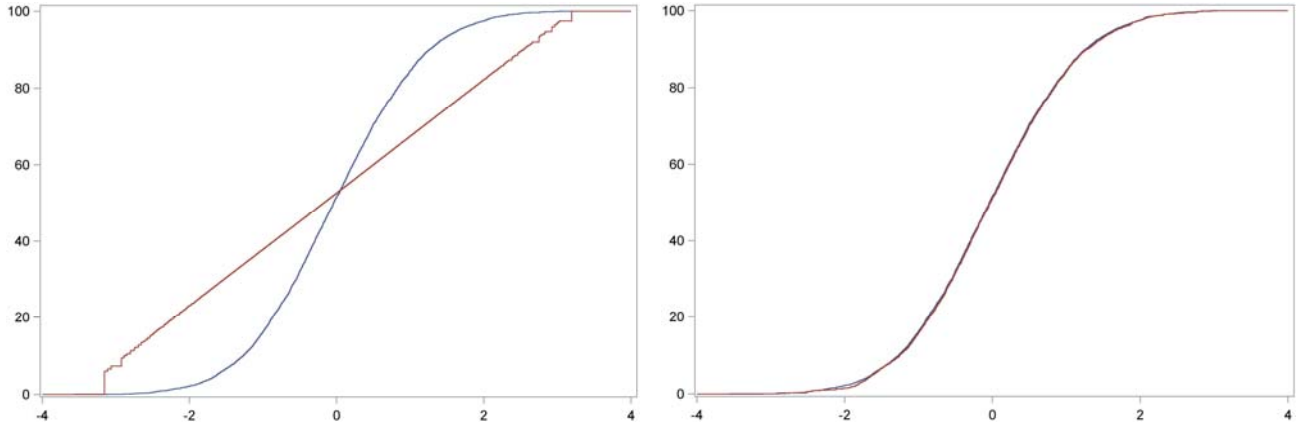
Wooldridge, J. 2010. Econometric analysis of cross section and panel data. MIT Press, Cambridge.

Figure 1 – Simulated (Univariate) Cumulative Distributions Before and After Distribution Matching

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



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



Panel C: Non-Random Sample III: Before (left) and after (right) distribution matching

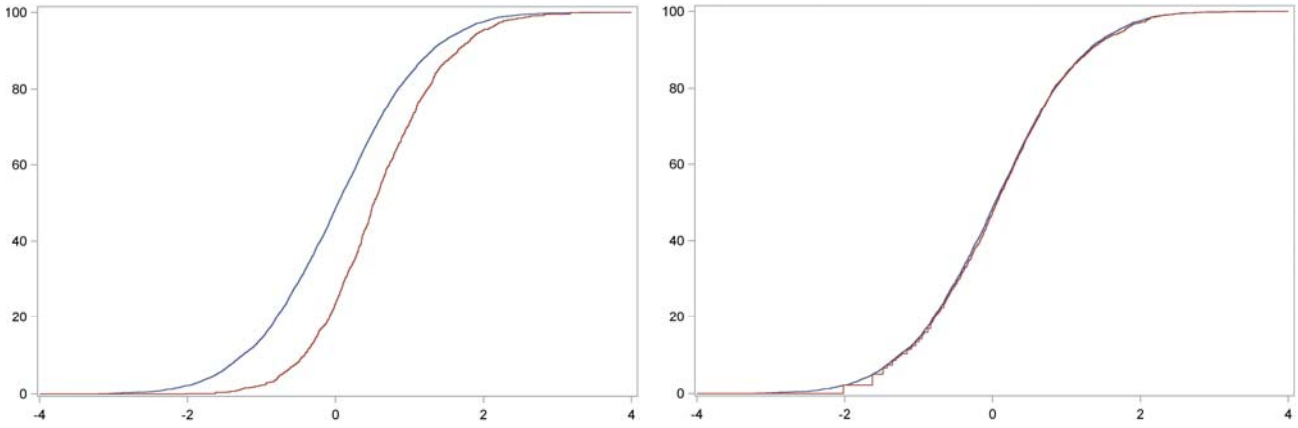


Figure 1 shows the empirical cumulative distribution of y_i for three types of non-random samples as described in the Appendix and for the corresponding distribution-matched samples, for one run of the results reported in Table 1, Panel A. Sample distributions are depicted with the red/lighter line and 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 – Average Monthly Distribution Before and After Bin-based Distribution Matching

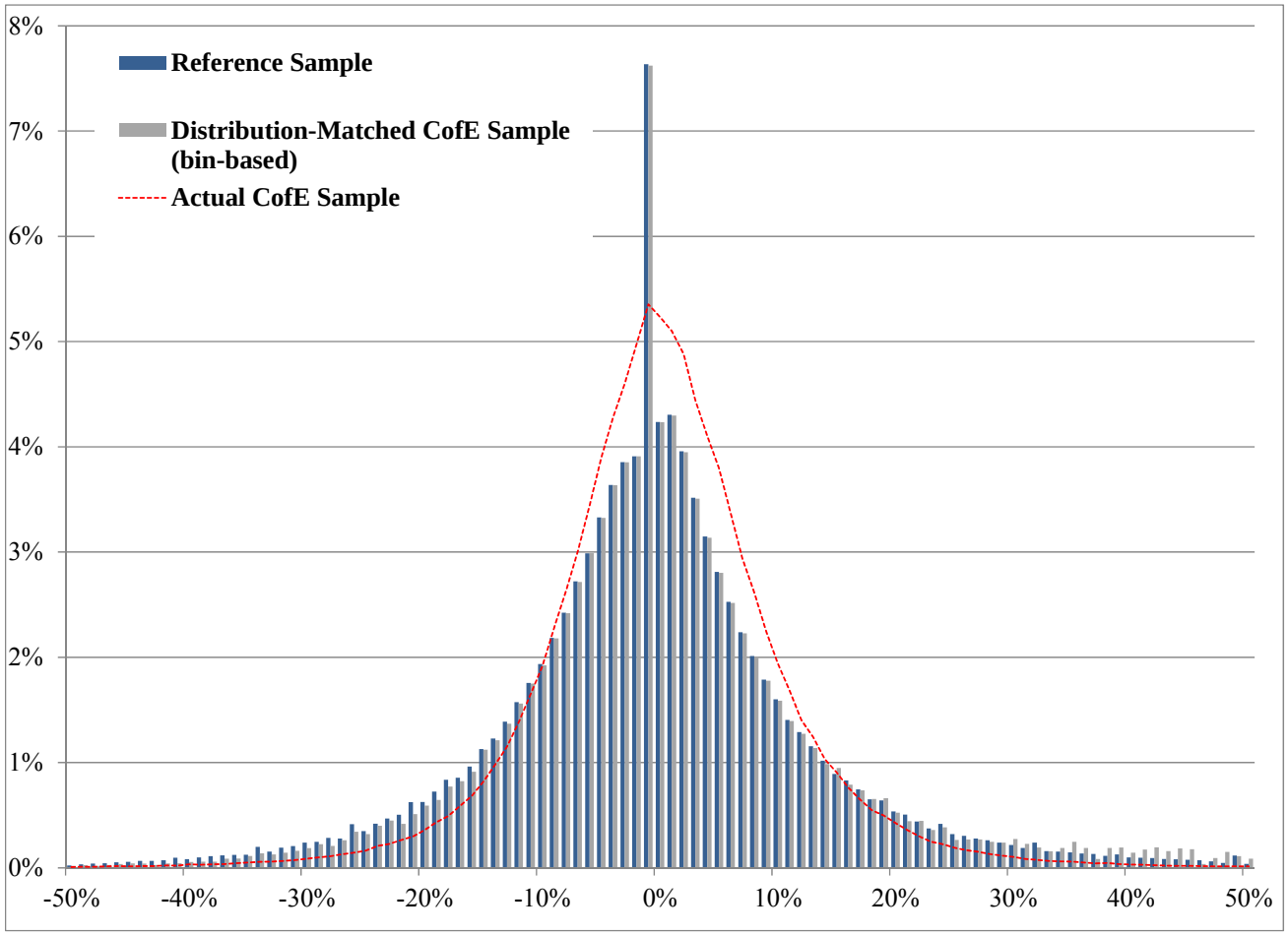


Figure 2 shows the empirical distribution (density) of excess returns for three samples: the actual *CofE* sample (dashed red line), the reference (Full Returns) sample (blue/dark bars), and the bin-based distribution-matched sample with $\gamma = 2.20$ (grey/light bars). The width of each bin is 100 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.

Table 1
Simulation Results From Forced Non-Random Samples and Corresponding Distribution-Matched Samples

Panel A: Both variables (standard) normally distributed

	CORR* = 0.5		**CORR* = 0**		**CORR* = -0.5**	
	KS	**CORR**	**KS**	**CORR**	**KS**	**CORR**
Population (n = 5,000)	N/A	0.5000	N/A	0.0003	N/A	-0.4999
Random Sample (m = 1,000)	0.0244	0.4991	0.0245	-0.0001	0.0243	-0.5005
(p-value)	*(0.6904)*		*(0.6878)*		*(0.6944)*	
Non-Random Sample I						
Ranked abs. distance to mean (Eq. A2)	0.1385	0.3267	0.1403	0.0006	0.1403	-0.3268
(p-value) | Percentile (Random Sample)	*(0.0000)*	0.0	*(0.0000)*	50.5	*(0.0000)*	100.0
Distribution-Matched Sample	0.0302	0.4856	0.0297	0.0018	0.0299	-0.4877
(p-value) | Percentile (Random Sample)	*(0.5422)*	30.1	*(0.5486)*	53.6	*(0.5391)*	71.4
Non-Random Sample II						
Uniform distribution	0.2550	0.7775	0.2546	0.0015	0.2533	-0.7780
(p-value) | Percentile (Random Sample)	*(0.0000)*	100.0	*(0.0000)*	53.2	*(0.0000)*	0.0
Distribution-Matched Sample	0.0093	0.4953	0.0095	0.0014	0.0094	-0.4943
(p-value) | Percentile (Random Sample)	*(0.9992)*	43.7	*(0.9992)*	53.2	*(0.9987)*	60.7
Non-Random Sample III						
Ranked abs. distance to maximum	0.2665	0.4236	0.2628	0.0017	0.2645	-0.4233
(p-value) | Percentile (Random Sample)	*(0.0000)*	0.0	*(0.0000)*	53.5	*(0.0000)*	99.9
Distribution-Matched Sample	0.0348	0.4873	0.0348	0.0034	0.0347	-0.4831
(p-value) | Percentile (Random Sample)	*(0.4003)*	31.5	*(0.4078)*	55.0	*(0.4073)*	77.5

(continued on next page)

Panel B: Both variables non-normally distributed

	CORR* = 0.5		CORR* = 0		CORR* = -0.5	
	KS	**CORR**	**KS**	**CORR**	**KS**	**CORR**
Population (n = 5,000)	N/A	0.4999	N/A	0.0005	N/A	-0.5005
Random Sample (m = 1,000)	0.0245	0.5010	0.0246	0.0009	0.0244	-0.5001
(p-value)	*(0.6851)*		*(0.6833)*		*(0.6921)*	
Non-Random Sample I						
Ranked abs. distance to mean (Eq. A2)	0.1419	0.3704	0.1450	-0.0003	0.1428	-0.3422
(p-value) | Percentile (Random Sample)	*(0.0000)*	0.0	*(0.0000)*	49.2	*(0.0000)*	100.0
Distribution-Matched Sample	0.0326	0.5018	0.0332	0.0003	0.0323	-0.4879
(p-value) | Percentile (Random Sample)	*(0.4691)*	50.6	*(0.4328)*	48.9	*(0.4588)*	67.8
Non-Random Sample II						
Uniform distribution	0.4993	0.7728	0.5033	0.0024	0.4976	-0.7500
(p-value) | Percentile (Random Sample)	*(0.0000)*	100.0	*(0.0000)*	51.2	*(0.0000)*	0.0
Distribution-Matched Sample	0.0135	0.5023	0.0135	0.0007	0.0134	-0.4993
(p-value) | Percentile (Random Sample)	*(0.9837)*	51.4	*(0.9831)*	49.0	*(0.9846)*	52.0
Non-Random Sample III						
Ranked abs. distance to maximum	0.2649	0.4305	0.2615	0.0007	0.2654	-0.4253
(p-value) | Percentile (Random Sample)	*(0.0000)*	0.1	*(0.0000)*	49.0	*(0.0000)*	99.6
Distribution-Matched Sample	0.0355	0.4808	0.0349	-0.0002	0.0355	-0.4887
(p-value) | Percentile (Random Sample)	*(0.4070)*	17.7	*(0.3933)*	48.1	*(0.4034)*	67.1

Table 1 presents correlation results for simulated populations of data, for random samples, and for three non-random samples with corresponding distribution-matched samples. Two variables x, y are constructed with a given (true) correlation 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: The selection probability for 'Non-Random Sample I' is decreasing in the ranked absolute distance from the mean, following Equation A2 in the Appendix; '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 corresponding non-random samples only, and are constructed by resampling such that the empirical cumulative distribution of y, $F^{NRS}(y_i)$, in the non-random sample mimics the empirical distribution of y in the population, $F^{POP}(y_i)$. The difference in the empirical distributions of the y variable between either a random sample and the population or a non-random sample and the population 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 ($F^{NRS}(y_i) = F^{POP}(y_i)$). 'CORR' denotes the estimated correlations. 'Percentile (Random Sample)' reports the percentile of the mean non-random sample correlation in the distribution spanned by the 1,000 correlations from the random samples.

Table 2
Descriptive Statistics of Monthly (Cross-Sectional) Distributions

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

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

Panel B: Implied Cost of Equity Sample

	# Firms	Mean	Std. Dev.	Skewness	Kurtosis	Min	P5	Q1	Median	Q3	P95	Max
Realized Returns	955	0.0121	0.0896	0.6034	6.1594	-0.3734	-0.1217	-0.0393	0.0087	0.0594	0.1557	0.5742
Excess Returns	955	0.0074	0.0896	0.6034	6.1594	-0.3781	-0.1263	-0.0439	0.0041	0.0548	0.1510	0.5696
VL CofE	955	0.0121	0.0070	0.8016	2.7667	0.0002	0.0032	0.0061	0.0118	0.0165	0.0236	0.0513
GLS CofE	955	0.0071	0.0036	5.4237	59.8661	0.0006	0.0034	0.0053	0.0068	0.0084	0.0110	0.0505
MPEG CofE	955	0.0088	0.0045	3.2930	27.0187	0.0003	0.0037	0.0060	0.0081	0.0105	0.0160	0.0538
OJN CofE	955	0.0098	0.0037	4.5876	46.6087	0.0040	0.0061	0.0077	0.0092	0.0110	0.0153	0.0526
CT CofE	955	0.0073	0.0049	4.9134	49.7870	0.0001	0.0020	0.0046	0.0067	0.0090	0.0133	0.0580

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), required to have at least 12 consecutive months of CRSP returns data. The *CofE* sample is a perfect subsample of the Full Returns sample, containing an average of 955 firms each month (383,955 firm-months). Firms in the *CofE* sample have sufficient data to calculate five *CofE* estimates based on Value Line data, denoted VL, and based on models in Claus and Thomas (2001, CT), Gebhardt, et al. (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN).

Table 3
Average Correlation and Regression Coefficients of
Realized Returns and *CofE* Metrics

	Actual *CofE* Sample	
	Correlation Coefficients	Regression Coefficients
VL CofE	-0.0033	-0.0142
t-stat	*-0.52*	*-0.15*
GLS CofE	-0.0096	-0.2273
t-stat	*-2.19*	*-1.34*
MPEG CofE	-0.0203	-0.4121
t-stat	*-4.52*	*-3.61*
OJN CofE	-0.0159	-0.3961
t-stat	*-3.44*	*-2.68*
CT CofE	-0.0258	-0.4723
t-stat	*-6.11*	*-3.79*
KS	0.1389	
p-value	*(0.0009)*	

The sample period is February 1976 to July 2009 (402 months). The (actual) Cost of Equity (*CofE*) sample is a perfect subsample of the Full Returns sample, containing an average of 955 firms each month with sufficient data to calculate five *CofE* estimates based on Value Line data, denoted VL, and based on models in Claus and Thomas (2001, CT), Gebhardt, et al. (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN). Table 3 contains average cross-sectional correlation coefficients and regression coefficients between five *CofE* metrics and realized excess returns over the 402 sample months for the actual (unmodified) *CofE* sample. 't-stat' denotes the Fama-MacBeth-type test statistic on the average cross-sectional correlation coefficients or regression coefficients.

Table 4
Association Tests on Reference Sample, Random Subsamples and CofE Subsamples

	Full Returns Sample	**1,000 Random Subsamples (of month-specific *CofE* sample size)**		**Actual *CofE* Sample**
		Mean	Range	
beta$^{\text{Market}}$	0.0052	0.0050	[0.0032 ; 0.0069]	0.0022
t-stat	*2.03*	*1.93*	*[1.27 ; 2.56]*	*0.88*
beta$^{\text{SMB}}$	0.0028	0.0027	[0.0015 ; 0.004]	0.0003
t-stat	*1.69*	*1.62*	*[0.93 ; 2.34]*	*0.20*
beta$^{\text{HML}}$	-0.0020	-0.0020	[-0.0030 ; -0.0010]	-0.0005
t-stat	*-1.33*	*-1.27*	*[-1.86 ; -0.63]*	*-0.32*
beta$^{\text{AQFactor}}$	0.0077	0.0074	[0.0051 ; 0.0090]	0.0023
t-stat	*2.44*	*2.32*	*[1.67 ; 2.79]*	*0.73*

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 *CofE* sample is a perfect subsample of the Full Returns sample, containing an average of 955 firms each month with sufficient data to calculate five *CofE* estimates based on Value Line data, denoted VL, and based on models in Claus and Thomas (2001, CT), Gebhardt, et al. (2001, GLS), Easton (2004, MPEG), and Ohlson and Jüttner-Nauroth (2005, OJN). The table shows average univariate implied factor premia, obtained from month-specific estimations of Equation (7b). The first column uses all monthly returns observations from the Full Returns Sample. The second column contains averages and ranges of coefficient estimates from 1,000 'Random Subsamples' drawn from the Full Returns sample. The sample size of each monthly cross-sectional draw is equal to the actual *CofE* sample size in that month. The rightmost column contains results for the Actual *CofE* sample.

Table 5
Simulation Results From CofE-calibrated Samples

	CORR* = 0.25		CORR* = 0.10		CORR* = 0	
	KS	**CORR**	**KS**	**CORR**	**KS**	**CORR**
Full "Returns" Samples	N/A	0.2503	N/A	0.1000	N/A	0.0001
Random Samples (m = CofE sample size)	0.0262	0.2503	0.0265	0.1005	0.0263	0.0005
(p-value) | t-stat	*(0.6580)*	*134.98*	*(0.6504)*	*58.46*	*(0.6590)*	*0.31*
Non-Random Sample (m = CofE sample size)						
Ranked abs. distance to mean (Eq. A2)	0.1847	-0.1662	0.1844	-0.0627	0.1841	0.0004
(p-value) | t-stat	*(0.0000)*	*-61.80*	*(0.0000)*	*-33.91*	*(0.0000)*	*0.23*
Distribution-Matched Sample	0.0319	0.1737	0.0319	0.0659	0.0318	0.0023
(p-value) | t-stat	*(0.5017)*	*25.40*	*(0.4974)*	*9.07*	*(0.5015)*	*0.30*

Table 5 presents correlations for simulated populations, for random samples, and for non-random samples (Non-random Sample I of Table 1) with its corresponding distribution-matched samples. The variables in the simulations are calibrated such that, each month, the distribution of the *y* variable and the distribution of the *x* variable approach the pooled empirical distribution of excess returns and the Value Line *CofE* metric (restricted to the first four moments). True (induced) correlations (CORR*) are 0.25, 0.10 and 0. The generated samples contain 6,122 observations in the average of 402 simulated cross sections. The random (non-random) samples have the same size as the actual *CofE* sample in any given month. The non-random sampling uses sampling weights according to Equation (A2). The difference in the empirical distribution of the *y* variable between either a random sample and the full returns sample or a non-random sample and the full returns sample 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 ($F^{NRS}(y_i) = F^{POP}(y_i)$). 'CORR' denotes the estimated correlations. 't-stat' denotes the Fama-MacBeth-type test statistic on the average cross-sectional correlation coefficients. The table presents grand average KS statistics, correlation coefficients and Fama-MacBeth-type test statistics from 20 independent runs.

Table 6
Association Tests in Distribution-Matched Samples

	Actual *CofE* Sample *(from Table 3)*	Distribution-Matched *CofE* Samples			
		KS-Based Sampling		Bin-based Weighted Sampling	
		Initial # = 20%	Initial # = 100	γ = 2.15	γ = 2.20
Avg. KS Statistic	0.1389	0.0581	0.0589	0.0992	0.1031
Avg. p-value	*0.0009*	*0.5287*	*0.7789*	*0.0280*	*0.0248*
# Months with p ≤0.10	401	42	5	372	375
Average cross-sectional correlation coefficients					
VL CofE	-0.0033	0.0524	0.0496	0.0514	0.0537
t-stat	*-0.52*	*6.08*	*5.35*	*4.14*	*4.28*
GLS CofE	-0.0096	0.0312	0.0278	0.0399	0.0413
t-stat	*-2.19*	*4.53*	*3.52*	*3.85*	*3.95*
MPEG CofE	-0.0203	0.0270	0.0274	0.0286	0.0305
t-stat	*-4.52*	*3.98*	*3.59*	*2.54*	*2.67*
OJN CofE	-0.0159	0.0316	0.0289	0.0350	0.0370
t-stat	*-3.44*	*4.36*	*3.57*	*3.10*	*3.23*
CT CofE	-0.0258	0.0153	0.0057	0.0213	0.0227
t-stat	*-6.11*	*2.23*	*0.73*	*2.07*	*2.18*
Average cross-sectional regression coefficients					
VL CofE	-0.0142	0.9000	0.9160	1.0828	1.1394
t-stat (against 0)	*-0.15*	*5.46*	*5.07*	*3.42*	*3.51*
t-stat (against 1)	*-10.68*	*-0.61*	*-0.46*	*0.26*	*0.43*
GLS CofE	-0.2273	1.3155	1.0740	2.2834	2.3356
t-stat (against 0)	*-1.34*	*3.93*	*2.70*	*3.30*	*3.30*
t-stat (against 1)	*-7.25*	*0.94*	*0.19*	*1.85*	*1.89*
MPEG CofE	-0.4121	0.7904	0.7075	1.0727	1.1192
t-stat (against 0)	*-3.61*	*3.62*	*2.83*	*2.10*	*2.13*
t-stat (against 1)	*-12.36*	*-0.96*	*-1.17*	*0.14*	*0.23*
OJN CofE	-0.3961	1.1476	0.8862	1.6353	1.6930
t-stat (against 0)	*-2.68*	*3.98*	*2.63*	*2.59*	*2.62*
t-stat (against 1)	*-9.45*	*0.51*	*-0.34*	*1.01*	*1.07*
CT CofE	-0.4723	0.5082	-0.0143	1.0092	1.0418
t-stat (against 0)	*-3.79*	*1.99*	*-0.05*	*2.00*	*2.02*
t-stat (against 1)	*-11.83*	*-1.93*	*-3.45*	*0.02*	*0.08*

Table 6 shows correlations and regression coefficients between five *CofE* measures and excess returns for the Actual *CofE* 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 *CofE* 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 *CofE* 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 *CofE* sample into "bins" (intervals) of 100 basis points. Each month, we redraw, with replacement, from the *CofE* sample to mimic the corresponding sample proportions in the Full Returns sample bin. Bins with no observations in the corresponding *CofE* sample are dropped. Bins in the extreme tails are weighted using Equation (3) in the text. We iterate the weighting factor γ to minimize the average difference in standard deviations between the Full Returns sample and the *CofE* sample. We repeat this resampling procedure 20 times. We compute cross-sectional correlations and regression coefficients each month and run and average the correlations and regression coefficients and their related time-series t-statistics for the 402 months in each run. The table contains the grand averages of average correlations and average regression coefficients and their related t-statistics across the 20 runs.

Table 7
Estimates from Blockwise Multiple Imputations

	Actual CofE Sample	**Multiple Imputations (M=10)** Full (g=1)	g=2	g=3	g=5
Average cross-sectional correlation coefficients					
VL CofE	-0.0033	-0.0023	0.0393	0.0427	0.0334
t-stat	*-0.52*	*-0.22*	*3.71*	*4.18*	*3.33*
GLS CofE	-0.0096	-0.0108	0.0116	0.0175	0.0206
t-stat	*-2.19*	*-1.39*	*1.45*	*2.12*	*2.37*
MPEG CofE	-0.0203	-0.0321	0.0085	0.0139	0.0149
t-stat	*-4.52*	*-4.00*	*1.00*	*1.62*	*1.63*
OJN CofE	-0.0159	-0.0248	0.0126	0.0168	0.0158
t-stat	*-3.44*	*-3.04*	*1.46*	*1.90*	*1.75*
CT CofE	-0.0258	-0.0401	-0.0103	-0.0031	-0.0077
t-stat	*-6.11*	*-5.52*	*-1.32*	*-0.38*	*-0.88*
Average cross-sectional regression coefficients					
VL CofE	-0.0142	0.2134	1.0890	1.0024	0.7062
t-stat (against 0)	*-0.15*	*0.79*	*4.16*	*4.04*	*2.96*
t-stat (against 1)	*-10.68*	*-2.91*	*0.34*	*0.01*	*-1.23*
GLS CofE	-0.2273	-0.1750	1.3541	1.5917	1.0096
t-stat (against 0)	*-1.34*	*-0.33*	*2.53*	*2.99*	*2.07*
t-stat (against 1)	*-7.25*	*-2.19*	*0.66*	*1.11*	*0.02*
MPEG CofE	-0.4121	-0.8350	0.6595	0.9592	0.4518
t-stat (against 0)	*-3.61*	*-2.20*	*1.78*	*2.54*	*1.28*
t-stat (against 1)	*-12.36*	*-4.83*	*-0.92*	*-0.11*	*-1.55*
OJN CofE	-0.3961	-0.7138	0.9631	0.9292	0.9411
t-stat (against 0)	*-2.68*	*-1.46*	*2.01*	*1.95*	*2.10*
t-stat (against 1)	*-9.45*	*-3.52*	*-0.08*	*-0.15*	*-0.13*
CT CofE	-0.4723	-1.1052	0.2082	0.3558	0.1450
t-stat (against 0)	*-3.79*	*-2.93*	*0.54*	*0.97*	*0.40*
t-stat (against 1)	*-11.83*	*-5.57*	*-2.05*	*-1.75*	*-2.34*

Table 7 presents average cross-sectional correlation coefficients and regression coefficients between excess returns and five *CofE* measures, for the actual *CofE* sample and for four examples of multiple-imputation adjustments: g=1 refers to the case of a single imputation model for the full cross section; g=2 divides the monthly cross sections at the median into two blocks and g=3 and g=5) refer to terciles (three blocks) and quintiles (five blocks). Multiple imputation analysis completes the incomplete (relative to the Full Returns sample) *CofE* sample using values from stochastic regressions. 't-stat' denotes the Fama-MacBeth-type test statistic on the average cross-sectional correlation coefficients or regression coefficients.

Table 8
Results From Heckman-type Selection Models

	No correction	Lag (MktCap)	Lag (MktCap), Volume, Age	Lag (MktCap), Volume, Age, B/M	Factor Betas	Combined
	(I)	(II)	(III)	(IV)	(V)	(VI) = (IV) + (V)
Average cross-sectional semipartial correlation coefficients						
VL CofE	-0.0033	-0.0079	-0.0060	-0.0067	-0.0371	-0.0055
t-stat	*-0.52*	*-1.32*	*-0.98*	*-1.10*	*-8.79*	*-0.91*
GLS CofE	-0.0096	-0.0146	-0.0118	-0.0126	-0.0254	-0.0117
t-stat	*-2.19*	*-3.57*	*-2.81*	*-3.02*	*-8.34*	*-2.81*
MPEG CofE	-0.0203	-0.0266	-0.0235	-0.0244	-0.0335	-0.0234
t-stat	*-4.52*	*-6.49*	*-5.59*	*-5.81*	*-11.25*	*-5.70*
OJN CofE	-0.0159	-0.0215	-0.0190	-0.0200	-0.0309	-0.0191
t-stat	*-3.44*	*-5.00*	*-4.32*	*-4.55*	*-10.15*	*-4.42*
CT CofE	-0.0258	-0.0299	-0.0279	-0.0286	-0.0342	-0.0278
t-stat	*-6.11*	*-7.73*	*-6.98*	*-7.24*	*-11.98*	*-7.15*
Average cross-sectional regression coefficients						
VL CofE	-0.0142	-0.0760	-0.0583	-0.0667	-0.3470	-0.0806
t-stat (against 0)	*-0.15*	*-0.84*	*-0.63*	*-0.73*	*-5.23*	*-0.93*
t-stat (against 1)	*-10.68*	*-11.85*	*-11.51*	*-11.67*	*-20.29*	*-12.46*
GLS CofE	-0.2273	-0.3867	-0.3321	-0.3478	-0.6980	-0.3807
t-stat (against 0)	*-1.34*	*-2.35*	*-2.03*	*-2.15*	*-5.63*	*-2.47*
t-stat (against 1)	*-7.25*	*-8.43*	*-8.15*	*-8.33*	*-13.69*	*-8.96*
MPEG CofE	-0.4121	-0.5387	-0.5011	-0.5195	-0.7455	-0.5494
t-stat (against 0)	*-3.61*	*-5.11*	*-4.67*	*-4.87*	*-9.56*	*-5.62*
t-stat (against 1)	*-12.36*	*-14.59*	*-13.99*	*-14.24*	*-22.39*	*-15.84*
OJN CofE	-0.3961	-0.5362	-0.4991	-0.5213	-0.8506	-0.5527
t-stat (against 0)	*-2.68*	*-3.84*	*-3.54*	*-3.71*	*-8.05*	*-4.25*
t-stat (against 1)	*-9.45*	*-11.01*	*-10.62*	*-10.84*	*-17.52*	*-11.93*
CT CofE	-0.4723	-0.5529	-0.5308	-0.5464	-0.7797	-0.5642
t-stat (against 0)	*-3.79*	*-4.73*	*-4.48*	*-4.65*	*-8.48*	*-5.13*
t-stat (against 1)	*-11.83*	*-13.28*	*-12.92*	*-13.17*	*-19.35*	*-14.23*
Auxiliary Information						
Avg. Pseudo R^2	N/A	0.48	0.51	0.51	0.05	0.52
Avg. Sample N	955	955	943	939	955	939
Avg. Sample Loss (%)	N/A	0.0%	1.7%	2.1%	N/A	2.1%
Avg. Reference Sample N	6,122	6,117	5,670	4,883	6,122	4,883
Avg. Reference Sample Loss (%)	N/A	0.1%	10.0%	22.4%	N/A	22.4%

Table 8 presents semipartial correlation coefficients between excess returns and five *CofE* 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 3. 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 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 *CofE* sample (the Full Returns sample) without all necessary data for the various selection models, over the month-specific number of observations with *CofE* data (returns data).

Table 9
Univariate Associations between (Excess) Returns and Risk Factor Betas (Implied Factor Premia)

Panel A: Cross-Sectional Sample Selection Criteria

	S&P500 Member		**Listed on NYSE**		**σ(EPS forecasts) missing**		**Price at least $5**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
beta^{Market}	0.0053	0.0017	0.0059	0.0029	0.0023	0.0061	0.0061	0.0042
	2.08	*0.65*	*2.27*	*1.14*	*0.96*	*2.33*	*2.28*	*1.71*
beta^{SMB}	0.0029	-0.0007	0.0031	0.0008	0.0011	0.0034	0.0031	0.0029
	1.74	*-0.43*	*1.86*	*0.45*	*0.70*	*2.02*	*1.89*	*1.74*
beta^{HML}	-0.0021	-0.0001	-0.0023	-0.0007	-0.0020	-0.0023	-0.0023	-0.0018
	-1.38	*-0.04*	*-1.48*	*-0.45*	*-1.31*	*-1.50*	*-1.46*	*-1.21*
beta^{AQFactor}	0.0078	0.0010	0.0083	0.0030	0.0035	0.0089	0.0099	0.0079
	2.47	*0.30*	*2.62*	*0.93*	*1.13*	*2.80*	*3.05*	*2.57*
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)*

(continued on next page)

Panel B: Induced Variation in the (Marginal) Distribution of Excess Returns

	Induced Sampling Change in Positive Excess Returns (Right Tail)								
	-25%	**-10%**	**-5%**	**-2.50%**	**0%**	**+2.50%**	**+5%**	**+10%**	**+25%**
Avg. Monthly Proportion	23.72%	38.21%	43.17%	45.66%	48.15%	50.65%	53.15%	58.15%	73.07%
beta^{Market}	0.0028	0.0045	0.0047	0.0050	0.0052	0.0053	0.0054	0.0059	0.0072
	1.36	*1.92*	*1.93*	*1.98*	*2.03*	*2.02*	*2.03*	*2.13*	*2.50*
beta^{SMB}	0.0005	0.0020	0.0024	0.0026	0.0028	0.0030	0.0033	0.0037	0.0054
	0.38	*1.29*	*1.48*	*1.59*	*1.67*	*1.77*	*1.93*	*2.13*	*2.98*
beta^{HML}	-0.0008	-0.0017	-0.0018	-0.0019	-0.0020	-0.0022	-0.0022	-0.0024	-0.0031
	-0.66	*-1.16*	*-1.19*	*-1.27*	*-1.32*	*-1.37*	*-1.40*	*-1.50*	*-1.83*
beta^{AQFactor}	0.0018	0.0052	0.0064	0.0071	0.0076	0.0082	0.0089	0.0103	0.0148
	0.66	*1.74*	*2.07*	*2.28*	*2.42*	*2.56*	*2.75*	*3.12*	*4.42*

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 tabulated coefficient estimates are average implied factor premia from univariate asset pricing tests (i.e., associations of realized returns with factor betas) over the 402 cross-sectional regression coefficients. T-statistics are based on the time-series standard error of the monthly estimates (Fama and MacBeth 1973). Panel A shows the results using all monthly returns observations in the sample period, separated using four cross-sectional sample selection criteria, as well as corresponding KS statistics on the difference between the in-sample returns distribution and the reference sample returns distribution. We analyze subsamples based on: month-specific membership in the S&P 500, listing on the NYSE, and whether sufficient analyst earnings forecasts exist to compute forecast dispersion metrics on IBES. Panel B reports implied factor premia for samples where we induced sampling biases in the (marginal) distribution of excess returns by first splitting the distributions into positive and negative subsamples, and then resampling from these subsamples with varying sampling proportions. Again, we draw samples of the month-specific size for each of 402 sample months with replacement. The resampling alters the negative or positive proportions of excess returns (i.e., the left versus the right tail of the distribution), 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 row 'Avg. Monthly Proportion' contains the monthly average of the effective proportion of positive excess returns.