Review of Accounting Studies (2021) 26:772–814
https://doi.org/10.1007/s11142-021-09581-0

Non-random sampling and association tests on realized returns and risk proxies

Frank Ecker[1] · Jennifer Francis[2] · Per Olsson[3] · Katherine Schipper[2]

Accepted: 8 January 2021/ Published online: 9 March 2021

Abstract

This paper investigates how data requirements often encountered in archival accounting research can produce a data-restricted sample that is 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 that uses only observations from the data-restricted sample to construct distribution-matched samples that approximate randomly-drawn samples from the reference sample. Our simulation tests provide evidence that distribution-matched samples yield generalizable results. We demonstrate the effects of non-random sampling in tests of the association between realized returns and five implied cost of equity metrics. In this setting, the reference sample has full information on realized returns, while on average only 16% of reference sample observations have data on cost of equity metrics. Consistent with prior research (e.g., Easton and Monahan *The Accounting Review* 80, 501–538, 2005), analysis using the unadjusted (non-random) cost of equity sample reveals weak or negative associations between realized returns and cost of equity metrics. In contrast, using distribution-matched samples, we find reliable evidence of the theoretically-predicted positive association. We also conceptually and empirically compare distribution-matching with multiple imputation and selection models, two other approaches to dealing with non-random samples.

Keywords Non-random sampling · Distribution matching · Multiple imputation · Resampling · Cost of equity · Realized returns

✉ Frank Ecker
Frank.Ecker@fs.de

[1] Frankfurt School of Finance and Management, Frankfurt, Germany

[2] Duke University, Durham, NC, USA

[3] ESMT Berlin, Berlin, Germany

JEL Classification M41

1 Introduction

In this paper, we develop, validate, and illustrate a practicable approach to deal with non-random samples whose underlying cause is data requirements, a pervasive issue in accounting research. Examples include requirements for analyst following, database inclusion (e.g., Execucomp includes S&P 1500 firms), and stock price above a threshold, such as $5. We examine how non-randomness[1] of the dependent variable in data-restricted samples affects results of empirical association tests, propose and validate a nonparametric resampling technique ("distribution-matching") to adjust for the effects of non-randomness, and thereby increase the generalizability of results and apply the technique to tests of associations between realized returns and implied cost of equity *(CofE)* metrics. Our goal is to assist accounting researchers in constructing more powerful and less biased test samples, thereby increasing the generalizability (decreasing the sample-dependency) of results.

As discussed later, distribution-matching differs from selection models and multiple imputations, which are sometimes used in the analysis of data-restricted samples. A Heckman-type selection model approach assumes the selection model can be reliably estimated on a random sample of the population, but practical research settings typically involve a trade-off between selection model fit and data requirements for the selection model variables. A selection model approach may therefore simply transfer the data-restrictions issue from the test model to the selection model. In our setting, we find that introducing a selection model does not resolve the issue of non-randomness that we address using distribution matching; in contrast, applying multiple imputation yields results that converge to the distribution-matching results.

The starting point of the distribution-matching technique is a common requirement in archival accounting research that the sample contains only observations with complete data for all variables of interest ("complete cases"). We first examine a pairwise association in a stylized simulation setting consisting of a reference sample with full information on one variable (y) and restricted data on the second variable (x). Complete-case analyses effectively impose any data restrictions of x on y and do not use information about y in incomplete observation pairs. If missingness of x is even weakly correlated with values of y, complete-case samples are non-random samples of y and association-test results would not generalize to the reference sample. Distribution-matching uses information about the marginal distribution of y (the reference distribution) and resamples complete-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, and using only the observations of the data-restricted sample. Using simulated data with known induced levels of statistical associations and three types of non-randomness, we show that association tests yield

[1] We define non-randomness by comparing a data-restricted sample to a specified reference sample to which the researcher would like to generalize results. Our reference sample is the population of CRSP firms with at least 12 consecutive monthly returns during February 1976 to July 2009 (the Full Returns sample).

biased results in the non-random samples; applying distribution-matching substantially reduces or even eliminates the bias. We view the results of these simulation analyses as providing evidence that distribution-matching can address the issue of non-random sampling in a generic research setting.

To illustrate distribution-matching in a specific research setting, we apply the technique to an archival setting characterized by stringent data requirements and inconsistent/counterintuitive results (e.g., Easton and Monahan 2005), namely, tests of the association between realized returns and *CofE* metrics. By definition, all firms *have* a cost of equity, but a researcher cannot *observe* the cost of equity for some firms, typically because of data requirements.[2] The theory linking realized returns and *CofE* metrics applies to the reference sample defined in note 1 (CRSP firms with at least 12 months of realized returns during February 1976–July 2009). This reference sample contains complete data on one variable (y, returns) and not on the other variable (x, *CofE*).[3] We show (1) the unadjusted sample with data on *CofE* metrics is a non-random sample of the reference sample (the *CofE* sample returns have substantially lower standard deviation, skewness and kurtosis)[4]; and (2) tests of the *CofE*-returns association based on the unadjusted *CofE* sample—that is, a sample that would typically be used in accounting research—produce weak or negative associations between realized returns and implied cost of equity metrics, consistent with previous empirical research (e.g., Easton and Monahan 2005) and inconsistent with theory.

In contrast, tests using a distribution-matched *CofE* sample produce statistically reliable evidence of the theoretically-predicted positive association between *CofE* metrics and realized returns. Distribution-matching does not involve creating new data; the technique resamples *CofE*-sample observations (*CofE* and returns pairs) so that the returns distribution in the distribution-matched *CofE* sample mimics the returns distribution in the reference sample. We apply two approaches. The first uses the non-parametric Kolmogorov-Smirnov (KS) test of general sample differences. The KS statistic rejects, at the 0.10 (0.05) [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 using only *CofE*-sample observations (*CofE* and returns pairs) so that the deviation between the returns distribution of the resulting distribution-matched *CofE* sample and the reference distribution of returns, as captured by the KS statistic, is minimized. This resampling procedure is effective and requires few assumptions but imposes a substantial computational burden. To address this practical concern, the second approach sorts the returns distributions of both the reference sample and the *CofE* sample into researcher-defined strata ("bins") of the continuous variable and applies a form of stratified resampling that aims to match the standard deviation of the reference returns distribution. We find that distribution-matching can result in smaller samples than the

[2] Four *CofE* models are based on analysts' earnings forecasts (Claus and Thomas 2001; Easton 2004; Gebhardt et al. 2001; 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] This result 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 (Easton and Monahan 2005; Francis et al. 2004).

original non-random samples; despite a possible loss of power, correlations between realized returns and *CofE* metrics are with one exception reliably positive in distribution-matched samples obtained using both approaches.[5]

The methodological inference from our results 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. We extend this inference in several ways. To illustrate the inference in other settings with restrictive data requirements, we show that imposing several plausible selection criteria on the reference sample (S&P 500 membership, NYSE listing, availability of a dispersion measure of analyst earnings forecasts, and stock price at least $5) can change the distribution of realized returns and lead to biased estimates in association tests of realized returns with risk factor premia. We reach the same inference when we directly induce changes in the distribution of realized returns and show the sensitivity of risk factor premia to these changes. To illustrate that applying distribution-matching does not produce false results, we apply the technique to Richardson et al.'s (2005) analysis of the association between returns and accruals, a setting in which previous research shows results consistent with theory, and do not overturn their inferences. To separate the effects of reduced sample size from the effects of non-randomness, we compare the association between realized returns and asset-pricing factor betas for a random subsample from the full returns sample (to capture the effects of reduced sample size) and the actual *CofE* sample (to capture the combined effects of reduced sample size and non-randomness). We reason that imposing an unnecessary data restriction on samples used in 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[6] 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) in our setting, efficiency losses due to reduced sample sizes 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. Finally, we show that inferences from analysis of our main *CofE* sample, restricted by the data requirements of all five *CofE* metrics we consider, are qualitatively similar when we re-do our main association tests on a purely IBES-based *CofE* sample.

We believe our findings support a conclusion that results obtained using unadjusted non-random samples may not support generalizations to a researcher-selected 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 results. We also believe our analyses provide a practical solution to this issue, in the form of distribution-matching.

[5] That is, we find a tradeoff between sample size/test power and generalizability. 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.

[6] 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.)

2 Motivation for and validation of distribution-matching

2.1 Motivation and intuition

In accounting research settings, data constraints often mean that analyses can be performed only on a subsample of observations, even though the results are intended to generalize to a population or reference sample to which the tested hypothesis logically applies.[7] The setting we consider is association tests (regression coefficients or correlation coefficients) between two variables, of which one (y) is available for all firms in the researcher-defined reference sample, while the second variable (x) is often missing.[8] A common treatment in the accounting literature is restricting the test sample to observations with complete information, yielding a data-restricted sample. The data constraints on x are imposed on y, causing information about the unrestricted distribution of y to be lost. We hypothesize (1) this deletion leads to non-random test samples and (2) association tests using these samples yield results that may not be generalizable to the reference sample. We address this problem by incorporating information about the reference distribution (of the complete variable y) into the association test. We resample observations from the data-restricted non-random sample to create an adjusted sample that mimics the reference distribution of the complete variable and 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 $\widehat{\rho}$ between two continuous random variables x and y[9]:

$$\widehat{\rho} = \iint x_i y_i f(x_i, y_i | s_i = 1)\, dxdy = \iint x_i y_i f(x_i | y_i, s_i = 1) f(y_i | s_i = 1)\, dxdy, \quad (1)$$

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:

$$\rho^* = \iint x_i y_i f(x_i, y_i)\, dxdy = \iint x_i y_i f(x_i | y_i) f(y_i)\, dxdy. \quad (2)$$

[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. We also acknowledge that, if the researcher's test sample is known to resemble the researcher's reference sample with respect to the dependent variable, the issue we consider does not arise.

[8] 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.

[9] In this discussion, the subsequent simulations and most of the empirical work, we focus on the correlation coefficient not 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, for example, 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.

$\widehat{\rho}$ 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, $f(x_i, y_i|\, s_i = 1) = f(x_i, y_i)$ or, equivalently, $f(x_i|\, y_i, s_i = 1)\, f(y_i|\, s_i = 1) = f(x_i|\, y_i)\, f(y_i)$. This condition implies that the unobserved data are missing completely at random (MCAR); only then would a restricted sample (i.e., a sample after deletion of observations because of missing data) be a random subsample of the reference sample.

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, $f(y_i|\, s_i = 1)$, 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 $\widehat{\rho}$ is less likely.

If the MCAR assumption is rejected, it must be replaced with a weaker assumption: either the data are missing at random, conditional on observed variables (MAR), including y (realized return in our application), 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 precludes testing 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. We begin by showing differences in the marginal distributions of realized returns between a full-returns sample and a *CofE* subsample and that results of association tests (with factor betas) also differ qualitatively between these two samples. We therefore focus on methods under the MAR assumption, specifically, distribution-matching and multiple imputations. In Section 5, we assess the impact of a possible NMAR assumption using a Heckman-type selection model.

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

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

Assuming complete data on y, the MAR assumption implies Eq. (3) holds.[10] Then $\widehat{\rho}$ 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 be assessed only ex post.

Regardless of concerns specific to our application, condition (3) contrasts with and is arguably weaker than the assumption that data are missing at random, essentially

[10] Figure 2b contains visual evidence that this equality is maintained after distribution matching in the specific empirical example discussed later in the paper. The results show a very small difference in economic terms in averages for the Value Line-based *CofE* metric before and after bin-based distribution matching. The difference (visually) increases towards the extreme realized returns because of the paucity of observations in the tails. In none of the 101 returns bins is the difference significant at the 0.01 level (results not tabulated). Analogous differences using the other *CofE* metrics are, if anything, generally smaller than the differences for the Value Line-based *CofE* metric.

equating $f(x_i, y_i|\, s_i = 1)$ 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 $f(y_i|\, s_i = 1)$, are knowable from the data. We use the marginal distribution of the dependent variable of the reference sample as opposed to simply deleting observations with incomplete data. That is, under condition (3), our approach focuses on the marginal distribution of y in the data-restricted and possibly non-random sample. We resample systematically only from observations in this sample with complete data on both variables, so that, in the limit, the marginal distribution of y_i matches the marginal reference distribution of y_i:

$$f(y_i|s_i = 1) \rightarrow f(y_i). \tag{4}$$

While the convergence in (4) is achievable in the limit, the effectiveness of distribution-matching in a given research setting is a function of, among other things, the number of restricted-sample observations and the size of the common support of the restricted-sample and reference distributions. 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. In addition, the resampling approach may be unnecessary if only a few observations are missing, making the restricted sample (nearly) equal to the reference sample.

To measure similarity in the cumulative distributions between the reference sample and the non-random data-restricted sample, we use the non-parametric Kolmogorov-Smirnov (KS) statistic, which computes the percentage maximum absolute distance between two cumulative empirical distributions.

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

where $F^{NRS}(y_i)$, $F^{POP}(y_i)$ are the cumulative distributions of y in the non-random sample and population, respectively. We use the KS statistic and its associated asymptotic p value for a test of distribution equality between a subsample and the reference sample and to assess the degree of convergence in (4) within the KS-based distribution-matching approach. Information on the setup of the simulation is available from the corresponding author.

2.2 Relation of distribution-matching to traditional matching approaches and other missing-data approaches

Traditional matching approaches The distribution-matching approach differs substantially in goal and implementation from traditional matching techniques that might be used to address issues of endogeneity and sample selection on observable determinants. Traditional matching relies on an observation-by-observation comparison (for example, matched pairs), while distribution-matching aims to reweight the observations in a sample distribution so that the resulting distribution approximates an (empirical or theoretical) reference distribution. Distribution-matching focuses on the outcome variable, while traditional matching focuses on (possibly multiple) independent variables, for example, through sorting or propensity scoring observations.

Approaches applicable in a MAR setting Missing-data approaches under the MAR assumption include multiple imputation (MI) and full-information maximum likelihood (FIML) estimation.[11] FIML incorporates information about the marginal distribution $f(y_i)$ 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, that is, the marginal distribution of returns in our setting, are preserved from the reference sample and considered in the analysis. MI uses the entire reference sample, so it is more efficient than distribution-matching in cross-sectional analyses; it can be applied when data are missing for more than one variable; and it can incorporate the use of auxiliary variables that are either informative about missingness or correlated with the missing data.[12] However, distribution-matching is non-parametric while both MI and FIML rely on multivariate normality. Descriptive statistics in Table 2 suggest the normality assumption is unlikely to hold in our example setting with realized returns. Based on theoretical arguments in Schafer (1997), supported by simulation evidence of Demirtas et al. (2008), that MI appears to be less susceptible to deviations from multivariate normality than maximum likelihood, we repeat our main tests using subsamples-based forms of MI and find results similar to those obtained using distribution-matching.

Other missing-data approaches We clarify the intuition of distribution-matching by contrasting it with three other approaches: estimations that take account of truncation; incomplete post-stratification and selection models. First, assuming the true distribution of the complete outcomes is normal, Tobin (1956) derives closed-form solutions when the outcome variable is truncated at a known upper or lower bound (see also Wooldridge 2010). Rather than focus on truncation, we 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.[13] Also, rather than making assumptions about the reference distribution of the outcome variable, we estimate its shape from the reference sample with complete data.[14] Second, the survey literature uses incomplete post-stratification, which 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 gender, not an outcome variable. The similarity to distribution-matching arises because

[11] The theoretical framework of MI and its validity for MAR data are well-established (e.g., Rubin 1987; Schafer 1997; Little and Rubin 2002).

[12] Hot-deck imputations similarly 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 and thereby preserve *pairs* of the variables of interest. While distribution-matching might decrease the size of the test sample, hot-deck imputations (similar to multiple imputations) aim to maximize its size.

[13] In simulations, we show that even minimal truncation can induce large bias in correlation coefficients in non-random samples of the outcome variable.

[14] As the simulations illustrate, our procedure can also be used with a theoretically derived reference distribution.

survey respondents need not be representative of the population, necessitating re-weighting responses if the goal is to generalize results to the population.[15] To that end, both post-stratification and distribution-matching import information about the marginal reference distribution. In fact, for the common support region of the sample and reference distributions, 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 when compared to the reference distribution. Intuitively, the effect of truncation is mitigated by oversampling from the tails of the sample distribution. In short, distribution-matching tries to combine the notions of stratified sampling and overcoming biases from truncation.

Third, distribution-matching assumes data are missing at random, conditional on observables, while Heckman-type selection models assume data are *not* missing at random, necessitating a first-stage probit selection model to capture the mechanism that selects observations into the restricted sample. Under certain conditions,[16] the bias in the test model can be alleviated by incorporating the inverse Mills ratio from the selection model. In contrast, distribution-matching disregards the selection-mechanism and uses information about its consequences by assessing and minimizing the difference of the sample distribution to a reference distribution. In Section 5.1, we apply the selection model approach and find that results using 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.[17] 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 the marginal distribution of y. For each of the three types of non-random samples drawn from the simulated populations, we 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.

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

[16] 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 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 (e.g., Enders 2010).

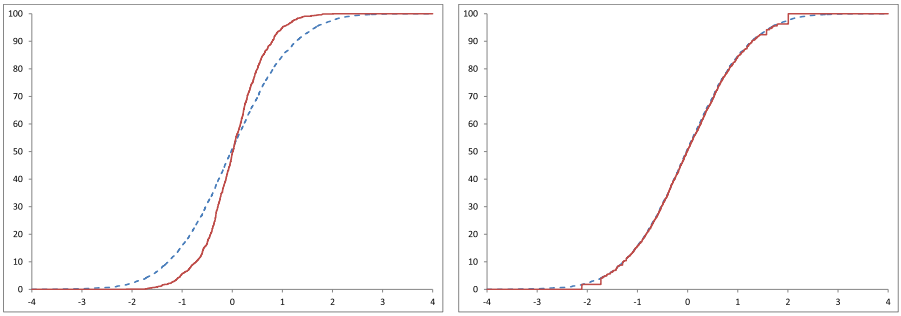
[17] Details about the design of our simulations and the generation of three distinct non-random samples from the simulated population data are available from the corresponding author upon request.

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. We examine population correlations specified at 0.5, −0.5, and 0 (to rule out the possibility that distribution-matching induces a correlation where none exists). We examine both negative and positive true correlations to provide evidence that the effectiveness of distribution-matching does not depend on the either the sign of the true association or the sign of the bias in the association estimate. We draw three kinds of non-random samples. In Non-random sample I, the selection probability of a given observation is decreasing in the absolute distance from the mean, leading to fewer observations that include extreme y values. Non-random sample II samples observations based on the uniform distribution over the entire interval of y observations, leading to higher selection probabilities for observations in the tails. These 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) in the estimates of associations. Non-random sample III is a form of non-symmetric selection probability, in that selection probability of an observation is increasing in y.

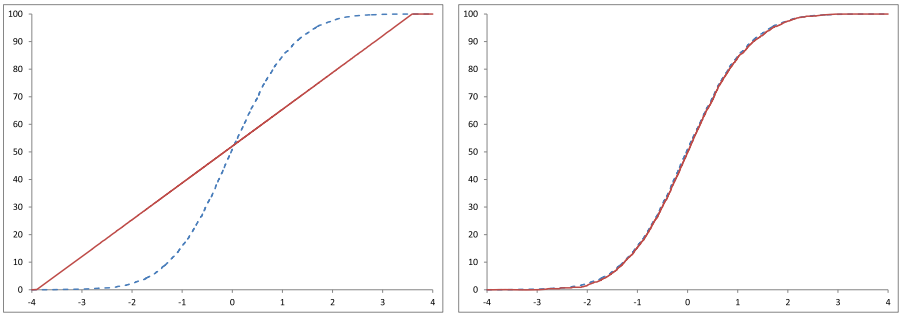
Figure 1 presents visual evidence of the effectiveness of distribution-matching for the three types of non-random samples, drawn from normally distributed population data. The figure depicts example distributions of y for a single randomly-chosen simulation run, for the unadjusted non-random samples (on the left), and the distribution-matched samples (on the right) of size $m = 1000$. The benchmark distribution, which appears in both right and left graphs, is from the population in that particular run ($n = 5000$). For all three non-random draws, the left-side graphs illustrate that the distributions deviate 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.

Table 1 reports numerical results of the simulations analysis. We focus on the results in Panel A (normally distributed variables). Results in Panel B (non-normally distributed variables) are qualitatively similar, suggesting the effectiveness of the non-parametric distribution-matching approach does not depend on the shape of the marginal distributions, in particular, 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 ($m = 1000$), is unimportant for the association-test point estimates as long as the sample selection is random.

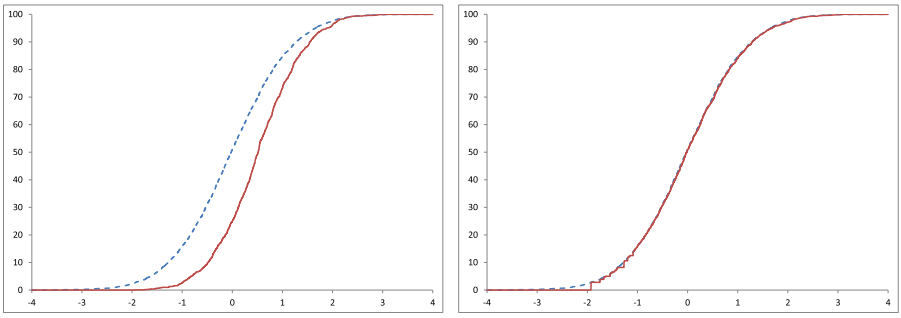
In contrast, and by construction, non-random samples have a distribution of y that differs sharply from the population distribution. For Non-random sample I (Non-random sample II) [Non-random sample III], KS statistics are about 13.9% (25.5%) [26.7%]. To assess the significance of the bias in correlation estimates for these samples, we report the percentile of the mean non-random sample correlation in the distribution spanned by the 1000 correlations as 'Percentile (Random Sample).' A



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

Fig. 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 and for the corresponding distribution-matched samples, for one run of the results reported in Table 1, Panel A. The dashed line marks the population distribution (benchmark) and is constant in all six graphs. Sample distributions are depicted with a continuous line, whereby the left (right) graphs are before (after) distribution-matching

small (large) percentile corresponds to a low (high) estimate. When the true correlations are 0.5 or −0.5, 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 1000

Table 1 Simulation results from forced non-random samples and corresponding distribution-matched samples

	CORR*=0.5		CORR*=0		CORR*=−0.5	
	KS	CORR	KS	CORR	KS	CORR
Panel A: Both variables (standard) normally distributed						
Population (n = 5000)	N/A	0.5000	N/A	0.0003	N/A	−0.4999
Random Sample (m = 1000)	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	0.1385	0.3267	0.1403	0.0006	0.1403	−0.3268
(p value) | %ile (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) | %ile (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) | %ile (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) | %ile (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) | %ile (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) | %ile (Random Sample)	*(0.4003)*	31.5	*(0.4078)*	55.0	*(0.4073)*	77.5
Panel B: Both variables non-normally distributed						
Population (n = 5000)	N/A	0.4999	N/A	0.0005	N/A	−0.5005
Random Sample (m = 1000)	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	0.1419	0.3704	0.1450	−0.0003	0.1428	−0.3422
(p value) | %ile (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) | %ile (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) | %ile (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) | %ile (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) | %ile (Random Sample)	*(0.0000)*	0.1	*(0.0000)*	49.0	*(0.0000)*	99.6

Table 1 (continued)

	CORR*=0.5		CORR*=0		CORR*=−0.5	
	KS	CORR	KS	CORR	KS	CORR
Distribution-Matched Sample	0.0355	0.4808	0.0349	−0.0002	0.0355	−0.4887
(p value) | %ile (Random Sample)	*(0.4070)*	17.7	*(0.3933)*	48.1	*(0.4034)*	67.1

Table 1 presents correlation results for simulated populations, 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 1000 runs. Populations consist of 5000 observations, from which samples of 1000 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. '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. %ile (Random Sample) reports the percentile of the mean non-random sample correlation in the distribution spanned by the 1000 correlations from the random samples

correlations from random samples) or 99.9 or higher (i.e., only one or fewer of the 1000 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 are close to zero (never exceeding 0.0169 in magnitude) 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 is of particular interest. Because the result is based on simulated data and an imposed 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 toward firms with typically-less-extreme returns realizations (described by prior research as relatively "stable" firms; see Footnote 4), as is often the case in accounting research situations. Our next tests investigate whether this simulation result applies to a specific well-studied empirical-archival setting.

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 firms with available data to calculate analyst-based implied *CofE* metrics. Specifically, we re-examine the correlation between realized returns and *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), 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 in this setting provides insights in its own right. Section 3.1 summarizes previous research, and Section 3.2 explains why we chose this setting to illustrate distribution-matching.

3.1 Research on the association between realized returns and implied CofE metrics

Realized returns are 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). The latter are predicated on the view 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, affected by the arrival of information during the return measurement period and therefore contain an expected return component and a potentially non-zero unexpected return component caused by news about cash flows and news about the expected return itself (Campbell and Shiller 1988; Campbell 1991).

Researchers typically assume *either* that (1) realized returns are a reasonable proxy for expected returns (that is, the unexpected return component is small, cancels out through aggregation, or both) *or*, (2) even in broad samples, the unexpected return component is a key non-cancelling component of realized returns (e.g., Elton 1999; Vuolteenaho 2002).[18] Adopting the latter perspective, researchers have developed several *CofE* metrics 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),[19] or, alternatively, researchers have

[18] 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.

[19] 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.

empirically purged the realized return of its unexpected ("news") component. For example, Campbell (1991) and Vuolteenaho (2002) propose a variance decomposition method that pre-specifies expected returns 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 identify the cash-flow news component of realized returns.[20] A stream of research that views realized returns and *CofE* metrics as alternative and imperfect proxies for expected returns 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.

The contrast of these 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, associations 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 unexpected 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 measure has either no association or counter-intuitive associations 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 there seems to be a trade-off 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 little affected by our distribution-matching approach.

Other papers attribute the weak association to the analyst forecast-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 but does not always result in statistically reliable associations.[21] Other studies adopt a portfolio design: Gode and Mohanram (2003), for example, find positive spreads in realized returns. In portfolio-level tests, Hou et al. (2012) show that returns spreads increase when they replace analyst forecasts with regression-based earnings forecasts.[22]

[20] 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 the 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.

[21] 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.

[22] 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).

Li and Mohanram (2014) use the Hou et al. (2012) approach to derive earnings forecasts and show positive associations on the firm level as well.

3.2 Analysis of the CofE setting

We analyze a different and possibly co-existing explanation for the weak and inconsistent results in tests of association between realized returns and *CofE* metrics. Rather than modify the metrics or consider other alternatives proposed in the literature, we leave the original *CofE* metric constructions in place and propose an explanation that derives from known features of samples used to estimate those metrics. That is, we choose the most conservative starting point (the problem as first examined by Easton and Monahan (2005)) and use unmodified implied *CofE* definitions and unmodified realized returns.

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. 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).[23] We further posit that association estimates based on such a non-random sample are difficult to generalize to the population, that is, the external validity of the results is questionable. We do not dispute previous findings but rather use distribution-matching to arrive at results that we believe can be more justifiably generalized to the population of listed firms. To summarize, 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 how data requirements may result in non-random samples, we let the requirements for five *CofE* metrics dictate the sampling bias in returns. While intuition suggests these data 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. We triangulate the effects of non-random returns on associations by using implied risk factor premia that are estimable for the entire reference sample. Specifically, our data contain complete information on both realized returns and asset pricing factor betas (loadings) for all observations in the reference sample. In a CAPM world, cross-sectional variation in beta is equivalent to cross-sectional variation in expected

[23] Relative to a variance decomposition approach or a news-purging approach, we require no assumption about either the determinants of the expected return component or 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.

returns. Therefore 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 artificially imposing 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 of the reference sample.

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 during 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 and preceding 11 months; a firm is required to have 12 consecutive monthly returns observations to enter the Full Returns sample in Month t.[24] The Full Returns sample contains 2,460,998 firm-month observations (24,657 unique firms). Table 2, Panel A, shows the average monthly cross section contains 6122 firms, with an average (median) monthly realized raw return of 1.30% (0.20%). Monthly excess returns, (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.[25]

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 3989 unique firms), a potentially non-random subsample of 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.[26] We calculate four other implied *CofE* estimates following Claus and Thomas (2001, *CT*); Gebhardt et al. (2001, *GLS*); Easton (2004, *MPEG*), and Ohlson and Jüttner-Nauroth (2005, *OJN*).[27] The *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

[24] The 12-monthly-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 CRSP returns universe is 0.0044 (average p value = 0.86).

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

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

[27] We follow Easton and Monahan (2005) and Botosan et al. (2011) and include only observations with positive values for all five *CofE* metrics in our *CofE* sample.

Table 2 Descriptive statistics of monthly (cross-sectional) distributions

	# Firms	Mean	Std. Dev.	Skewness	Kurtosis	Min	P5	Q1	Median	Q3	P95	Max
Panel A: Full returns sample (asset pricing test sample)												
Realized Returns	6122	0.0130	0.1615	3.7403	82.7487	−0.7442	−0.1947	−0.0595	0.0020	0.0676	0.2453	3.2195
Excess Returns	6122	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												
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 6122 firms (2,460,998 firm-months), required to have at least 12 consecutive months of CRSP returns data. The *CofE* sample is a 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)

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 excess returns of the *CofE* sample is 9.87%, a reduction of about 22% relative to the Full Returns sample. The Full Sample returns are positively skewed, with skewness coefficient of 3.74; 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 Benchmark results of associations between CofE metrics and realized returns

We first estimate cross-sectional Pearson correlations and slope coefficients from regressions of realized (excess) returns on each of the five *CofE* metrics; intercepts (not tabulated, in the interest of brevity) are included in all regressions. We use Eq. (6) to estimate slope coefficients for each Month t using all complete returns-*CofE* observations available for that month.

$$R_{i,t}-R_{f,t} = \delta_{0,t} + \delta_{1,t}CofE_{t-1} + \varepsilon_{i,t}. \tag{6}$$

The averages of the monthly coefficient estimates $\delta_{1,t}$ over the sample period measure the 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.[28] Results are reported in Table 3 for the unmodified *CofE* sample resulting from deletion when data on any of the *CofE* metrics is missing (analogous to Easton and Monahan 2005). These results are broadly consistent with prior literature on the association between *CofE* estimates and realized returns, if not more negative.[29] Correlations show either no reliable relation between realized returns and *CofE*, in the case of VL *CofE*, or a negative relation, in the case of the other four *CofE* metrics (t-statistics range from −0.52 to −6.11). Regression 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 reliably different from their theoretical value of 1, with t-statistics (not tabulated) between −7.25 and −12.36.

[28] 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 unrelated to the magnitude of the correlation. In Section 4.4.2, we report both correlation and regression coefficients.

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

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 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. The KS statistic captures the maximum absolute distance between the (cumulative) returns distributions in the reference sample and in the *CofE* sample. The associated *p* value is the probability of rejecting a true H_0 of both distributions being indistinguishable

4.3 Benchmark results of factor beta tests

To exploit the richness of the cost of capital setting we use tests of factor betas using two-stage asset-pricing tests; one purpose is to serve as an input to our demonstration that the *CofE* sample is a non-random sample from the reference sample. In the first stage, we estimate slope coefficients (factor betas) in a firm-specific time-series regression of excess returns on each risk factor:

$$R_{i,t} - R_{f,t} = a_{i,t} + b^{F}_{i,t} F_t + \varepsilon_{i,t}, \tag{7a}$$

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); $b^{F}_{i,t}$ = 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.[30]

[30] 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.

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 Eq. (7a):

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

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.

We use the samples described in Table 2 to establish benchmark associations between excess returns and factor betas. The tests are motivated by the idea that both *CofE* metrics and 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 subsamples, such as the *CofE* sample. Table 4, column 1, shows the second stage coefficient estimates from Eq. (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 (the coefficient estimate corresponds to a risk premium of 0.52% per month; t-statistic = 2.03) as is the association 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 0.05 level, one-tailed. The association between returns and *HML* beta is not reliably different from zero at the 0.05 level.[31]

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, the monthly average is 6122 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,

[31] 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. (2016) 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.

Table 4 Association tests on reference sample, random subsamples, and *CofE* subsamples

	Full returns sample	1000 Random subsamples (of month-specific *CofE* sample size)		Actual *CofE* sample
		Mean	Range	
$beta^{Market}$	0.0052	0.0050	[0.0032; 0.0069]	0.0022
t-stat	*2.03*	*1.93*	*[1.27; 2.56]*	*0.88*
$beta^{SMB}$	0.0028	0.0027	[0.0015; 0.0040]	0.0003
t-stat	*1.69*	*1.62*	*[0.93; 2.34]*	*0.20*
$beta^{HML}$	−0.0020	−0.0020	[−0.0030; −0.0010]	−0.0005
t-stat	*−1.33*	*−1.27*	*[−1.86; −0.63]*	*−0.32*
$beta^{AQFactor}$	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 6122 firms with at least 12 consecutive months of CRSP returns data. The *CofE* sample is a 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 Eq. (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 1000 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

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 association between realized returns and risk proxies. To illustrate the effects of sample size, columns 2 and 3 of Table 4 present results of association tests between risk factor betas and realized returns from 1000 randomly drawn subsamples of the Full Returns sample (Random Subsample) of the same size each month as the actual *CofE* sample (monthly average of 955 firms).[32] Column 2 reports average slope coefficients and t-statistics, and column 3 contains the range of values across the 1000 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 0.10 (market risk premium) and 0.12 (*AQFactor* premium). 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

[32] The Random Subsample results in column 2 of Table 4 are based on averages of 20 random subsamples drawn from the Full Returns sample.

betas evidences a significant association with excess returns, and all factor premia are reduced in magnitude by at least 50%. 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.

4.4 Demonstration of distribution-matching on simulated *CofE*-calibrated data and on archival data

4.4.1 *CofE*-calibrated simulations

Table 5 shows simulation results when the data approximate the size and shape of the distribution of the reference sample excess returns and the Value Line *CofE* metric. We use empirically determined parameters of the actual distributions of excess returns and of a *CofE* metric to create simulated data similar to the archival data. We are able to mimic the first four moments of the variable distributions. We induce correlations of 0.25, 0.10, and 0. We report results for Non-Random Sample I, constructed to be less extreme than the random sample by setting the selection probability to decrease in the absolute distance to the variable mean. For the simulated Full Returns data (first line of Table 5) and for random samples of the same size as the actual *CofE* sample (second line), estimated correlations are similar to the induced population correlations. 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 indicates that for the intentionally less-extreme non-random sample, the KS test statistic rejects similarity of the distributions at better than the 0.0001 level. Estimated associations between the simulated returns and simulated *CofE* metrics are negative and highly significant even though true correlations are positive: when the true correlation is 0.25 (0.10), 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 0.03 with significance levels of about 0.50. When the true correlation is 0.25 (0.10) [0], the estimated correlation is 0.17 (0.07) [0.00], 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 reported in Table 5—negative correlation estimates when true correlations are positive—highlight the potentially substantial bias in results when the marginal distribution is non-randomly drawn from the reference sample. Second, despite extreme differences in the characteristics of marginal distributions, distribution-matching yields an adjusted sample with correlations similar in sign and magnitude to the true correlations.

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	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 6122 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 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

4.4.2 Distribution-matching the actual *CofE* sample

We construct distribution-matched samples from data on realized 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. 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 substantial differences in the shape of the *CofE* sample returns distribution, as compared to the reference sample returns distribution, we change the implementation of the distribution-matching approach used in the simulation in two ways. We first apply 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.[33] 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 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 describe both approaches next.

Method 1: Kolmogorov-Smirnov-based distribution-match We start by randomly sampling either 20% of the month-specific sample or, separately, 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.[34] Our previous analyses suggest the KS test on this initial sample is likely to reject the 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), recompute 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 1000 times, thereby allowing for KS-based distribution-matched samples to increase by a maximum of 1000 observations each month.[35] 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 (i.e., with 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 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).[36]

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

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

[34] 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.

[35] The *CofE* sample contains on average 955 firms per month. Therefore 1000 iterations effectively allows 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.

[36] 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, resembling 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.

matching is similar to post-stratification except that 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.[37] 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.[38] 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 re-weight the most extreme positive and most extreme negative returns, with equal weighting at both extremes in the following form (month subscripts omitted):

$$w_{i_{CofE}} = \begin{cases} w_{i_{RS}} \sum_{j=\min(i_{RS})}^{\min(i_{CofE})} \left[min\left(i_{CofE}\right) - i_{j,RS} + 1 \right]^{\gamma} & \forall \; i_{RS} = \min\left(i_{CofE}\right) \\ w_{i_{RS}} & \forall \; \min\left(i_{CofE}\right) < i_{RS} < \max\left(i_{CofE}\right) \\ w_{i_{RS}} \sum_{j=\max\left(i_{CofE}\right)}^{\max(i_{RS})} \left[i_{j,RS} - \max\left(i_{CofE}\right) + 1 \right]^{\gamma} & \forall \; i_{RS} = \max\left(i_{CofE}\right) \end{cases} \quad (8)$$

$w_{i_{CofE}}(w_{i_{RS}})$ 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 γ = 2.15, and −0.0008 (t-statistic = −0.56) at γ = 2.20. Figure 2a illustrates 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 continuous distribution of returns in the *CofE* sample is sorted into bins of pre-determined widths and then resampled, such that the sample proportion of each bin is equal to that bin in the reference distribution. The procedure is effective if the heights of the resulting light bars, representing the strata, match the heights of the dark reference strata. As previously mentioned (note 10), Fig. 2b plots the average Value

[37] 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.

[38] 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.

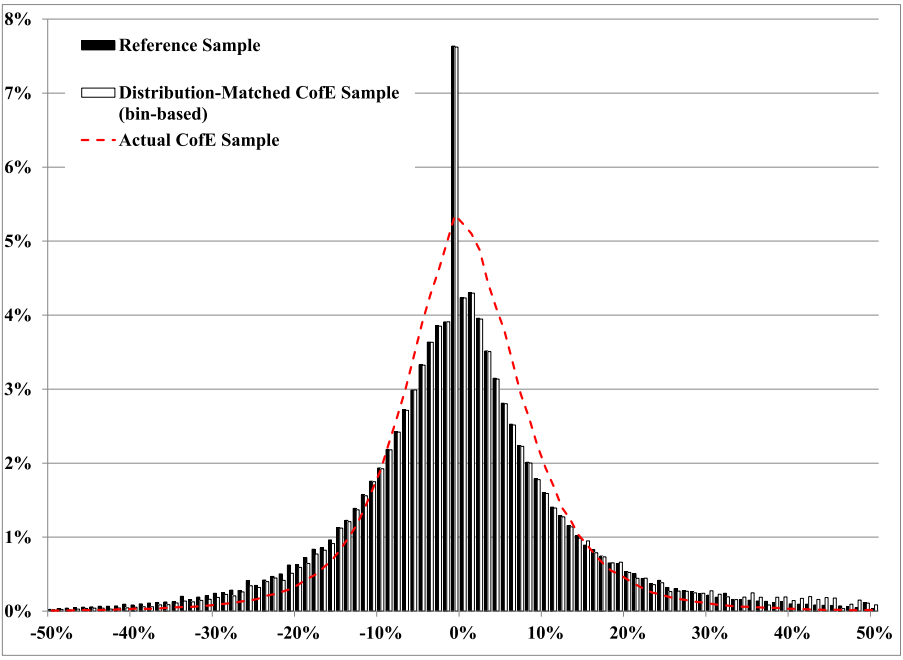
Line-based *CofE* metric by returns-bin, before and after bin-based distribution matching; across the 402 sample months, in none of the 101 returns bins is the difference significantly different from zero at the 0.01level (results not tabulated). Differences using the other *CofE* metrics (not plotted) are, if anything, generally smaller in absolute magnitude.

Association tests using distribution-matched samples Table 6 reports results of correlation tests between realized returns and *CofE* metrics for the distribution-matched samples under both the Kolmogorov-Smirnov and bin-based approaches. The average KS statistic (average *p* value) of the unadjusted *CofE* sample is about 14% (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 20% of firms (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.

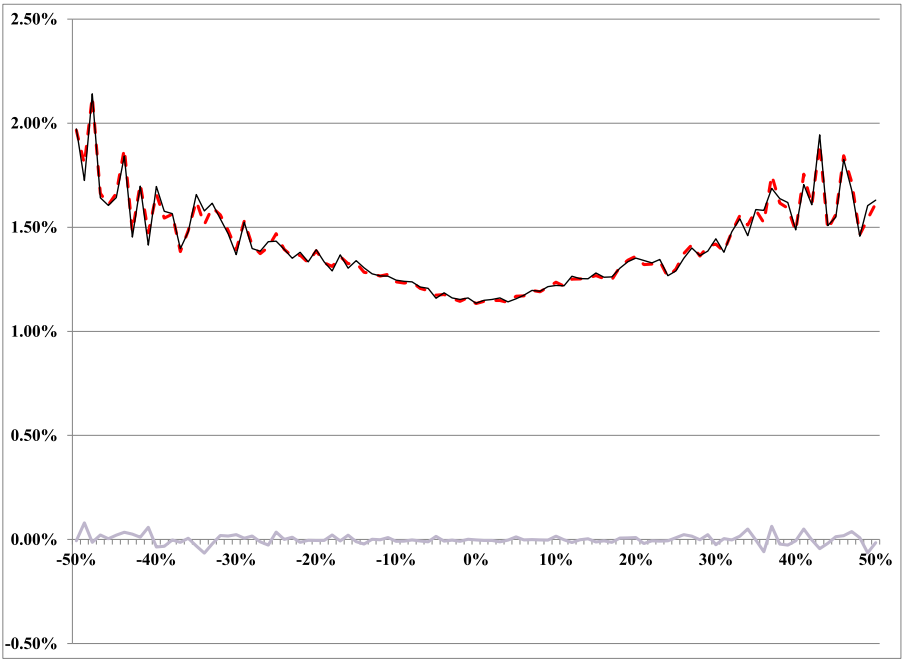
The top portion of Table 6 reports correlations between five *CofE* measures and realized returns, using the KS-based distribution-matched samples (columns 2 and 3) and the bin-based distribution-matched samples (columns 4 and 5). For KS-based matching, the time-series average correlations across 402 months are reliably positive 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. For bin-based matching, we report correlations for γ = 2.15 and γ = 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 the average KS statistic decreases, relative to the unmodified *CofE* sample, and remains significant for 372 (375) months for γ = 2.15 (γ = 2.20). In these analyses, all five *CofE* measures have reliably positive correlations with realized returns, ranging from 0.021 (CT *CofE* measure) to 0.054 (the VL measure), with t-statistics between 2.07 (CT *CofE* measure, γ = 2.15) and 4.28 (VL *CofE* measure, γ=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 results reported in Table 3, the coefficients are generally significantly positive, with t-statistics usually exceeding 2.0, and statistically indistinguishable from 1 in 16 of the 20 specifications. The 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*



Panel A: Average Monthly Returns Distribution Before and After Bin-based Distribution-matching



Panel B: Average (Value Line) CofE Before and After Bin-based Distribution-matching

sample (results not tabulated) and using one bin-based distribution-matched sample (γ = 2.20; results not tabulated). The factor premia from these samples are qualitatively

◀ **Fig. 2** Graphical evidence of distributional properties before and after bin-based distribution-matching. Panel A: Average Monthly Returns Distribution Before and After Bin-based Distribution-matching. Panel B: Average (Value Line) CofE Before and After Bin-based Distribution-matching. Fig. 2, Panel A, shows the empirical distribution (density) of excess returns for three samples: the actual *CofE* sample (dashed line), the reference (Full Returns) sample (dark bars), and the bin-based distribution-matched sample with $\gamma = 2.20$ (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%. The figure shows distribution-matched-sample returns from a single randomly chosen run of the resampling procedure. Panel B depicts the average Value Line *CofE* metric, by realized-returns bin, before (dark solid line) and after distribution-matching (dashed line), and the associated differences (light solid line)

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

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 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, 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 Eq. (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. In untabulated results, 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 between −0.59 and 1.46. The exception is the coefficient on the market beta in the VL *CofE* regression, which shows a small, reliably positive difference of 0.0001 ($t = 2.09$). In all cases, coefficients from the KS-based sample are numerically 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 results. First, selection criteria that yield estimation samples with different returns distributions, as compared to a reference sample, decrease the ability to detect theoretically-predicted associations between realized returns and *CofE* estimates. Second, adjusting the distribution of the outcome variable (in this case, realized returns) in the non-random sample to mimic that of the

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	$\gamma = 2.15$	$\gamma = 2.20$
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 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 as described 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 ("runs"). For each run, we compute cross-sectional correlations and regression coefficients each month and average the correlations and regression coefficients, computing the time-series t-statistics over the 402 months. The table contains the grand averages of average correlations and average regression coefficients and their related t-statistics across the 20 runs

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.

5 Extensions

5.1 Relating distribution-matching to selection models and multiple imputation

Selection models Both selection models and distribution-matching seek to incorporate information into the test model beyond the information in the complete-data subsample. The latter focuses only on the outcome variable, whose empirical distribution in the reference sample can be derived by the researcher or is empirically estimable, and aims to construct a test sample that appears randomly selected with respect to the outcome variable. In contrast, a selection model (1) operates under the assumption that data are *not* missing at random, conditional on observed data, and (2) requires explicit modelling of the missingness mechanism using additional explanatory variables, which might impose 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 by Heckman (1979), because the exogenous covariates in the first-stage selection model are attainable for all observations, or, equivalently, attainable for a *random* subsample of the population.[39] Consequently, results from a Heckman-type model are generalizable only to the sample for which the selection model variables are available, and increasing selection-model fit by including more explanatory variables is likely to impose increasingly stringent sample restrictions due to data requirements. Exacerbating 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.[40]

[39] 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).

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

Distribution-matching is, by design, non-parametric and based on a reference distribution of the outcome variable. In contrast, 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; 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, 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 selection model fit. We restrict our analysis to selection models with explanatory variables available for all, or at least the vast majority of, observations in the Full Returns sample and include some or all of the following: firm size (CRSP market capitalization at the end of the prior month), firm age (number of 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, eq. (7a).[41]

Table 7 reports semi-partial correlation coefficients between *CofE* metrics and excess returns and goodness-of-fit measures for the probit selection models estimated. The model including only size has a pseudo-R^2 of 0.48, with no additional sample loss; adding 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 and then computing the correlation between the returns and the residualized *CofE* metric. Similarly, regression coefficients (bottom portion 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 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, as compared to unadjusted correlations or slopes.[42] With the caveat that the effects of the inverse Mills ratio on the semi-partial correlations and

[41] 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 afterward is almost complete.

[42] 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 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.

Table 7 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	6122	6117	5670	4883	6122	4883
Avg. Reference Sample Loss (%)	N/A	0.1%	10.0%	22.4%	N/A	22.4%

Table 7 presents semi-partial 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 'No correction' column 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 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)

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 Heckman-type selection models do not change the conclusion from results obtained using the unadjusted *CofE* sample.

Multiple imputations [43]A standard implementation of multiple imputation will fail to recognize differences in the functional form connecting returns and *CofE*. Specifically, correlation and regression coefficients from a *single* imputation model for the entire cross-section of returns are qualitatively similar to the actual *CofE* sample results reported in Table 3, except that coefficients tend to be more negative (farther from the theoretical value) and standard errors are larger because of additional variance from the imputed *CofE* data. However, when we modify the approach to allow for group-wise imputation models (two groups divided at the monthly cross-sectional median; three groups or five groups) to improve the overall model fit, 12 of the 15 regression slopes (five *CofE* metrics times 3 different sample groupings) are positive, significant at the 0.10 level or better and indistinguishable from 1. Results for correlations are generally positive but weaker and insignificant in eight of the 15 specifications and especially for the *CT CofE*.

We assess the sensitivity of these results in two ways. First, we preclude the imputation of negative values[44] by using log transformations before imputing and find qualitatively comparable results for two and three imputation groups and stronger results for five imputation groups per cross section. Qualitative inference

[43] Standard statistical software packages like SAS and Stata include commands for performing multiple imputations, for diagnostics to check for the convergence of the estimation, 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. We used 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 10 datasets, each of 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. The 10 datasets can be analyzed independently and results aggregated. We formulate a month-specific imputation model for each *CofE* metric using only the (complete) excess returns data. To improve the model fit, we split each monthly cross-section at the median, into terciles and into quintiles, resulting in two, three or five groups, allowing for different intercepts and coefficients in each group.

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

changes only for the MPEG *CofE*, with a higher coefficient of 0.65 (t-statistic against 0 = 1.83, t-statistic against 1 = −0.97). Second, we impute data for the factor tests. We construct a dataset that deletes the loadings estimates from eq. (7a) for observations without *CofE* metrics and then impute the now missing (by construction) values for the loadings before we estimate the implied factor premia using eq. (7b). Untabulated results show that, for imputations of the full cross section, only the market risk premium is significantly positive (t=1.92). For imputations using two, three and five groups results are qualitatively similar to the results from the Full Returns sample. We interpret the weight of this evidence as suggesting that multiple imputation can be a viable alternative to distribution-matching, albeit one that imposes a normality assumption and that may require additional adjustments in a specific research setting, for example, precluding inadmissible imputed values.

5.2 Asset pricing tests on returns of samples that meet selection criteria used in accounting research

Using 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. We next consider whether results of asset pricing tests of the association between risk factor betas and realized returns are sensitive to the following cross-sectional selection criteria that likely yield non-random samples: S&P 500 membership, a potential screen in compensation research[45]; NYSE listing, a screen in some intra-day trading studies; availability of the standard deviation of analysts' earnings forecasts, required for research examining forecast dispersion; or a stock price of at least $5. We apply each criterion separately to the Full Sample, report the proportions of firms that do and do not meet the criterion and re-estimate Eq. (7b), separately, for observations meeting and not meeting the criterion.

Results are reported in Table 8, Panel A. The 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, unsurprisingly, most extreme for the S&P 500 criterion (8.44% meet the criterion). 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. That is, the returns distributions of firms *not* selected by these three criteria more closely resemble the returns distribution of the Full Returns sample.[46] The exception is the price of at least $5 criterion.

[45] The Execucomp database covers S&P 1500 firms since 1994, but other compensation data sources can be more restrictive. See, for example, Brookman et al. (2006) for an overview.

[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).

These findings suggest 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 8 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 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 sensitivity of association test results to (small) changes in non-random sampling, we split the Full Sample realized returns distribution into positive and negative returns and reweight both subsamples differentially to varying degrees. This test is motivated by the conjecture that data requirements might lead (implicitly) to a similar and less extreme reweighting of positive and negative returns. We resample 20 times, by month, for each of 402 sample months. In each month, with N_t firms in each month, we resample 20 times with replacement N_t firms. We use these 402 months of resampled data to illustrate the effect on the returns-factors betas association, as these data are available for the entire reference sample.

Results are reported in Table 8, Panel B. The 0% column shows results when we resample preserving the population proportions of positive and negative returns. These results coincide with the Table 4 Full Returns sample results; small differences result from sampling with replacement as opposed to using the full sample. The columns labeled −2.5%, −5%, −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 labeled +2.5%, +5%, +10%, and +25% show results when resampling increases (decreases) the portion of positive (negative) returns sampled by the specified percentages. The results suggest that increasing 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 relatively 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

[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. 2016). Consequently, its t-statistic becomes more negative as the proportion of positive returns increases.

sensitive to differences in sample selection criteria, with the degree of sensitivity differing with the nature of the selection criteria.

Overall, the results in Table 8, Panel B, illustrate that the outcome of all four returns-betas association tests is sensitive to a pre-specified characteristic of the returns distribution, namely whether the sample contains more positive returns. We directly manipulated the sample distribution; however, this sample characteristic may also be implicitly influenced by researcher-chosen selection criteria, data availability, or sample partitions.

5.4 Applying distribution-matching to Richardson et al. (2005)

We apply distribution-matching to Richardson et al.'s (2005) analysis of the association between *annual* returns and accruals to illustrate empirically that distribution-matching does not produce false results.[48] Richardson et al. find results consistent with prior research (notably Sloan 1996) and with their theory that lower-reliability accruals lead to lower earnings persistence that investors appear to not fully understand. We therefore expect that applying distribution-matching should similarly produce results consistent with prior research and theory; that is, distribution-matching should not falsify or bias these results. We follow procedures outlined in Section 3 and footnote 8 of Richardson et al. to obtain a sample as close as practicable to theirs[49] (the unadjusted accruals sample) and replicate the analysis presented in their Table 8, Panel B. Before distribution-matching, the average cross-sectional KS statistic rejects similarity of the $5-price-filtered reference distribution of returns (as described in footnote 49) and the accruals-sample distribution of returns at the 0.0917 level; differences are significant in 30 of the 40 sample years. After applying the KS-based distribution-matching approach, the average annual KS statistic is reduced to 0.0525, with an average p value of .6377. Thus the sample of returns obtained by requiring specified accounting data appears to be non-random relative to the $5-price-filtered reference distribution of returns, and distribution-matching lets the distribution approximate a randomly-drawn sample distribution. The distribution-matched sample has fewer observations (approximately 23,000 as opposed to over 105,000 in the unadjusted accruals sample).

The key test variables reported in Richardson et al.'s Table 8, Panel B, are ROA (coefficient = 0.09, t = 1.69), change in working capital (ΔWC; coefficient = −0.30, t = −7.54), change in net non-current operating assets (ΔNCO; coefficient = −0.27, t = −6.77) and change in net financial assets (ΔFIN; coefficient = −0.054, t = −1.94). After distribution matching, we find the following (not tabulated): the coefficient on ROA is

[48] We thank an anonymous referee for suggesting this test, which provides an opportunity to apply distribution matching in an annual returns setting, to complement the monthly returns setting analyzed in most of this paper. We believe there is no *ex ante* reason to predict whether the data-restrictions in the Richardson et al. paper (including requiring certain accounting data) will or will not be consequential in terms of affecting the distribution of returns for the filtered (data-requirements-constrained) sample.

[49] Our initial sample is about 37% larger than Richardson et al.'s, regardless of whether we use the current version of Compustat or a legacy version intended to approximate the version available in the early 2000s. Eliminating low-priced stocks (price less than $5 at the end of Month +3 after the fiscal year-end) results in a sample whose size is similar to that in Richardson et al. Therefore we discuss (untabulated) results using this price-filtered returns sample as the reference sample of returns as our main results. We obtain qualitatively similar results using the larger, unfiltered-returns reference sample (i.e., a sample analogous to the sample in the monthly returns setting analyzed in this paper).

Table 8 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	5621	500	4131	1991	2641	3520	1497	4625
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)*

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

Table 8 (continued)

	−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 6122 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

0.16, t = 2.27; the coefficient on ΔWC is −0.31, t = −4.28; the coefficient on ΔNCO is −0.17, t = −3.01; the coefficient on ΔFIN is −0.02, t = −0.29. Using two-sample t-tests of differences in slope coefficient estimates from the full sample and the distribution-matched sample, we do not reject equality at lower than the 0.18 level. In other words, the distribution-matched sample yields results that are statistically indistinguishable from the full-sample results of Richardson et al.'s returns test. We believe these results complement our previous analyses by supporting an inference that distribution matching does not produce false results.

5.5 Eliminating the requirement of Value Line data

We re-estimate the Table 6 association tests after dropping the requirement that *CofE* firms be followed by Value Line (results not tabulated). For these tests, the *CofE* sample contains firms with the necessary IBES data to calculate four *CofE* metrics; the monthly average is 1980 firms, a substantial increase from the monthly average of 955 firms when we impose the Value Line requirement.[50] However, the larger IBES sample returns distribution remains reliably different from the reference sample of returns: the KS statistic for the unadjusted IBES *CofE* sample is 11.21% with average p value = 0.0018; 400 of 402 months have a p value of 0.10 or less. After distribution-matching using the 20% initial draw setting, the KS statistic is 0.0585 (average p value = 0.35; 135 months have a p value = 0.10 or less); when the initial draw is 100 firms the average KS statistic is 0.0496 (average p value = 0.89; one month has a p value = 0.10 or less). Average cross-sectional regression coefficients for all four IBES *CofE* metrics in the unadjusted IBES *CofE* sample are reliably negative (t-statistics between −2.96 and −5.03) and reliably different from 1 (t-statistics between −2.30 and −4.86). After distribution-matching using either 20% of the sample or 100 firms, average cross-sectional regression coefficients are reliably positive (t-statistics between 2.20 and 2.91) and, with the exception of the *CT CofE* metric, are not reliably different from 1. In summary, dropping the requirement of Value Line data increases the sample size, does not eliminate the non-randomness of the resulting *CofE* sample returns and does not systematically alter the associations of the *CofE* metrics with realized returns in the unadjusted sample.

6 Conclusions

This paper proposes and illustrates a practical solution to a pervasive issue in empirical-archival accounting research, namely, data-restricted samples that are non-randomly drawn from the reference sample to which the researcher would like to generalize results. Paired with an objective of maximizing the number of observations with values for all variables ("complete cases"), these non-random samples are effectively dictated by data availability. We describe, validate, and illustrate a distribution-matching

[50] When we impose data requirements separately for each of four IBES-based *CofE* metrics and the VL metric, the monthly average number of observations are: 2130.7 (*CT CofE)*; 2169.7 (*OJN CofE*); 2063.1 (*MPEG CofE*); 2226.0 (*GLS CofE*); and 1495.4 (*VL CofE*). In all five cases, the KS statistic rejects at the 0.0016 level or better the hypothesis that the realized returns distributions of the *CofE* samples are similar to the realized returns distribution of the reference sample.

technique that can be used to align the distribution of a non-random estimation sample with that of a reference sample. 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.

We illustrate the effectiveness of the distribution-matching approach in tests of associations between returns and five popular implied cost of equity (*CofE*) estimates. This setting is of interest in its own right, given the practical and theoretical importance of the association and the weak and mixed results in previous research. Our analysis shows that associations between realized returns and *CofE* metrics are influenced by the properties of the realized returns distribution used to estimate the associations. Our reference sample is CRSP firms with at least 12 consecutive monthly returns during 1976–2009; our test sample is firms with sufficient data to calculate the *CofE* measures. The latter sample is a substantially smaller, non-random subsample of the former. We first show that associations between realized returns and *CofE* metrics are weak or negative, as in prior research. After distribution-matching, so that 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.[51] We also discuss two alternative approaches: 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).

Viewed broadly, our analysis implies that non-randomness of samples resulting from data requirements may lead to conclusions that do not generalize to a reference sample selected by the researcher. 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. Highlighting an important caveat to the goal of maximizing the size of a data-constrained research sample, our analysis suggests maximization of a data-constrained sample may not be goal-congruent with increasing the generalizability from such a sample.

Our findings suggest researchers might benefit, in terms of increasing the generalizability of results, from examining the impact of data requirements on the empirical distribution of the test model variables, in particular the variable whose distribution is most affected by the availability of other variables of interest. We believe the approach we discuss, modified to suit the specific research context, will assist future research by providing an explanation for weak or counter-intuitive results from data-restricted samples. Distribution-matching might also benefit future research by helping to coordinate across studies that address either similar questions using different samples. Comparisons of results across studies would be facilitated to the extent many researchers can define, construct and analyze a common reference sample.

[51] 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).

Acknowledgements We appreciate financial support from Duke University's Fuqua School of Business, ESMT and the Frankfurt School of Finance and Management. We thank the Editor, Peter Easton, and two anonymous referees for their helpful guidance as well as Robert Bartels (MEAFA conference discussant), Mary Barth, Alex Belloni, Alon Brav, Federico Bugni, Judson Caskey, Qi Chen, Shuping Chen, Dain Donelson, Ron Dye, Jason Hall, Xu Jiang, Bill Kinney, Stephannie Larocque, Charles Lee (Stanford Summer Camp discussant), Fan Li, Xi Li (SMU-SOAR Symposium discussant), John McInnis, Maria Ogneva (FARS Meeting discussant), Panos Patatoukas, Jerry Reiter, 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, the 9th MEAFA Research Meeting at the University of Sydney, and the 5th International Corporate Governance Conference at Tsinghua University for their comments and suggestions. Early drafts of the paper were circulated under the title "Association Tests of Realized Returns and Risk Proxies Using Non-Random Samples."

Funding Open Access funding enabled and organized by Projekt DEAL.

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., 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* 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*. New York: Guilford Press.
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.
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. 2016. Time-varying risk premium in large cross-sectional equity datasets. *Econometrica* 84: 985–1046.
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., S. Kothari, and S. Shu. 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: 159–194.
Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.
Hou, K., M. VanDijk, and Y. Zhang. 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics* 53: 504–526.
Hribar, P., and P. Jenkins. 2004. The effect of accounting restatements on earnings revisions and the estimated cost of capital. *Review of Accounting Studies* 9: 337–356.
Lennox, C., J. Francis, and Z. Wang. 2012. Selection models in accounting research. *The Accounting Review* 87 (2): 589–616.
Li, S. 2010. Does mandatory adoption of international financial reporting standards in the European Union reduce the cost of equity capital? *The Accounting Review* 85: 607–636.
Li, K., and P. Mohanram. 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies* 19: 1152–1185.
Little, R., and D. Rubin. 2002. *Statistical analysis with missing data*. 2nd ed. New Jersey: Wiley & Sons.
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.
Richardson, S., R. Sloan, M. Soliman, and I. Tuna. 2005. Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics* 39: 437–485.
Rubin, D. 1987. *Multiple imputation for nonresponse in surveys*. New Jersey: Wiley & Sons.
Schafer, J. 1997. *Analysis of incomplete multivariate data*. Boca Raton: Chapman and Hall.
Sloan, R. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review* 71: 289–315.
Tobin, J. 1956. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.
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*. Cambridge: MIT Press.