Journal of Banking and Finance 190 (2026) 107775

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Lost in the multiverse: Methodological uncertainty in studying global equity returns

Nusret Cakici [a], Christian Fieberg [b,h,i], Gabor Neszveda [c], Vanja Piljak [d], Adam Zaremba [e,f,g,*]

[a] Gabelli School of Business, Fordham University, New York, USA
[b] HSB Hochschule Bremen - City University of Applied Sciences, Bremen, Germany
[c] Magyar Nemzeti Bank, Budapest, Hungary
[d] University of Vaasa, Vaasa, Finland
[e] Finance and Accounting Department, MBS School of Business, Montpellier, France
[f] Institute of Finance, Poznan University of Economics and Business, Poznan, Poland
[g] Monash Centre for Financial Studies, Monash University, Melbeurne, Australia
[h] Concordia University, Montreal, Canada
[i] University of Luxembourg, Luxembourg

ARTICLE INFO

Keywords:
Methodological uncertainty
Nonstandard errors
Return predictability
International stock markets
Multiverse analysis
Country equity returns

ABSTRACT

We examine methodological uncertainty in studies of the cross-section of country equity returns, analyzing 15 predictors across up to 13,824 research implementations. Varying nine key design choices, we find that many classic signals—such as momentum, beta, or idiosyncratic risk—are surprisingly fragile. Setups emphasizing small, segmented countries enhance performance, while those focused on liquid, investable markets tend to weaken it. Applying bootstrap and out-of-sample tests, only a few factors, such as market size, dividend yield, short-term momentum, and sovereign risk, consistently emerge as robust. Our evidence calls for cautious interpretation of country-level return patterns and for robustness checks across alternative research designs.

1. Introduction

Just like stocks, country equity markets may exhibit predictable return patterns. For instance, markets with high local risk, low valuation ratios, and strong momentum have been shown to outperform their safer, higher-priced, and low-past-return counterparts (e.g., Bali and Cakici, 2010; Avramov et al., 2012; Asness et al., 2013; Baltussen et al., 2021). Researchers frequently capture these phenomena using portfolio sorts, which group markets based on their underlying characteristics. Yet country-level tests involve a small cross-section and therefore may be especially sensitive to research design choices. While firm-level research shows that seemingly trivial methodological decisions can have a significant impact (Menkveld et al., 2024; Soebhag et al., 2024; Cakici et al., 2024), country-level tests present their own unique challenges. Studies vary in terms of data sources, country index representations, geographical scope, and hyperinflation treatment, and these choices can shape the resulting evidence.

In this study, we examine the role of methodological uncertainty in the studies of cross-sectional variation in country equity returns. Building on the recent nonstandard errors literature (Menkveld et al., 2024; Soebhag et al., 2024; Walter et al., 2024), we run a multiverse analysis (Steegen et al., 2016; Simonsohn et al., 2020) of global portfolio sorts. To this end, we compile 15 return-predictive signals from the literature and transform them into long-short country allocation strategies. We account for nine key methodological decisions concerning data sources, sample preparation procedures (scope, study period, hyperinflation treatment), and portfolio implementation (number and size of portfolios, weighting schemes, rebalancing frequency, and implementation lag). We benchmark these choices against a systematic review of 52 studies to document the most commonly used procedures. Altogether, we generate 13,824 unique research designs, which we analyze to understand the impact of methodological choices on observed return patterns.

Our principal findings can be summarized as follows. First, we document substantial variation in the performance of country portfolios due to methodological design. Fig. 1 illustrates the Sharpe ratios for two

* Corresponding author.
E-mail addresses: a.zaremba@mbs-education.com, adam.zaremba@ue.poznan.pl (A. Zaremba).

https://doi.org/10.1016/j.jbankfin.2026.107775
Received 14 August 2025; Accepted 4 July 2026
Available online 6 July 2026

prominent cross-sectional patterns: value, measured with the dividend yield (DY), and momentum. The dispersion in risk-adjusted payoffs is striking. Depending on specific research designs, the Sharpe ratios for the momentum-sorted strategies typically range from -0.5 to +0.5. For the dividend yield, this span is even slightly wider, with some scores falling even below -1. Notably, different implementations lead not only to uncertainty about the magnitude of profits but also to the question of whether any economic gains exist at all. The observed variation proves so substantial that it may cast doubt on the actual structure of the cross-section of country returns. Numerous research design choices may critically affect the conclusions.

Second, we demonstrate that even the most established cross-country return patterns are astonishingly fragile. Most generate significant profits on spread portfolios only in a modest fraction of implementations, typically not exceeding 15-20%. This includes even the most prominent country characteristics, such as momentum, reversal, or valuation measures. Furthermore, the mere size of the economic gains is relatively small. The average monthly return across all tested long-short portfolios is only about 0.17%, with the values for particular characteristics ranging from -0.13% to 0.41%.

To formally verify the robustness of return patterns across different research designs, we propose two tests. First, we develop a bootstrap-based test by aggregating average returns from all implementations into an overall "mean of means." Using block bootstrapping to preserve cross-sectional dependencies, we generate a distribution of bootstrapped means and alphas to assess statistical significance. Second, we implement an out-of-sample design selection procedure that ranks implementations by past performance. Using different sorting windows and design pools, we examine whether the best past implementations continue to generate significant premia.

The results may seem surprising: they cast doubt on whether many cross-country premia exist in a robust sense. Traditional momentum, beta, and idiosyncratic risk fail to confirm their significance across research designs. Only a handful of signals—also prominent in the practitioner literature—emerge as particularly reliable. Specifically, we identify four factors that prove robust and significant across the first two tests: sovereign rating, short-term momentum, dividend yield, and market size. Among these, market size stands out as the most profitable and reliable. These four characteristics are important not only in the full multiverse but also in most filtered subsets defined by different decision nodes.

Our third finding concerns the impact of specific methodological decisions on the overall results. Using Anderson-Darling tests, regression analysis, and filtering the multiverse across decision nodes, we pinpoint the research design choices with the most critical impact on portfolio performance. First and foremost, the choice of data source matters. Relatively strong return patterns are documented in Compustat and Global Financial Data-based indices, which provide particularly broad global coverage and exposure to small markets. Conversely, focusing on more tradable and realistic securities, such as ETFs or futures, significantly reduces the observed payoffs. The returns on anomalies implemented in these assets are even three times lower than in regular indices, raising doubts about whether some premia survive in more investable settings.

Next, the geographic scope of the sample plays an important role. Studies focusing only on developed markets show almost no significant return patterns, and average factor portfolio returns are close to zero. In contrast, expanding exposure to emerging and frontier markets leads to a remarkable performance boost. Third, the portfolio weighting scheme also proves to be particularly important. Weighing assets by market capitalization—so popular at the company level almost wipes out all gains. In contrast, accentuating the role of smaller and more "extreme" markets—either through equal-weighting or rank-weighting—significantly increases returns. On the other hand, issues such as the inclusion of hyperinflationary periods in the sample or the minimum number of countries in a portfolio turn out to be practically irrelevant to the main results. This echoes firm-level evidence that anomaly payoffs often weaken when tests emphasize larger and more investable assets.

Overall, research design choices that enhance the impact of small, hard-to-trade, and likely (at least partially) segmented countries lead to stronger results. On the contrary, limiting the study to larger and more liquid markets—or emphasizing their role through appropriate research techniques—weakens the final outcomes. As a consequence, designs with broad, equal- or rank-weighted exposure may overstate the premia available in large, liquid, and directly investable markets. However, not all factors are created equal, and what strengthens one can also weaken another. For example, while dividend yield and size patterns are most pronounced in samples that include emerging and frontier markets, momentum effects prevail in developed countries.

Finally, we uncover variations in the magnitude of nonstandard errors (NSEs)—i.e., the variability of the results due to methodological uncertainty—across anomalies and research design choices. Overall, the average level of NSEs in cross-sectional, country-level studies exceeds the standard errors (SEs) by 19% to 59%, depending on the calculation methodology. While this figure may seem substantial, it remains lower than in other asset classes, such as stocks, futures, or cryptocurrencies,

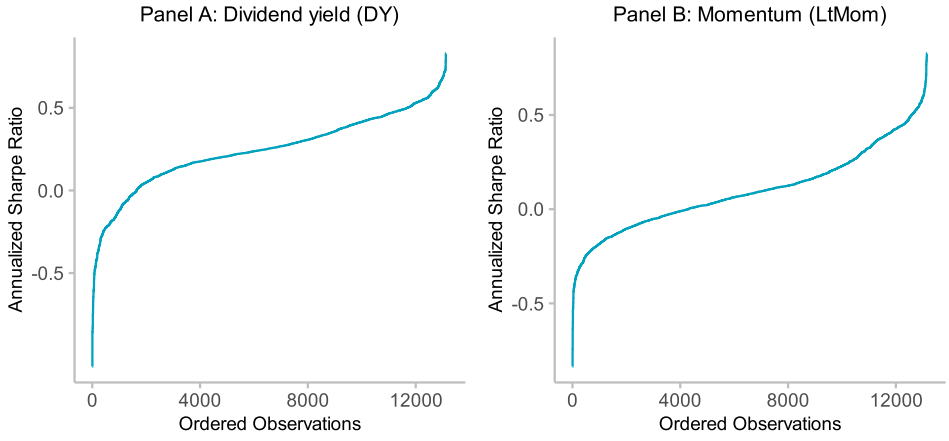

Fig. 1. Sharpe ratios on country value and momentum strategies.
The figure presents the annualized Sharpe ratios for a set of one-way sorted portfolios, constructed based on two prominent market characteristics: dividend yield (*DY*) and long-term momentum (*LtMom*), where the latter is calculated as the total return over months t-12 to t-2. The sets include 13,152 and 13,120 implementations, respectively. The Sharpe ratios are arranged in ascending order, from lowest to highest. The study period spans from January 1973 to December 2022. A detailed description of the sorting variables can be found in Table 1.

where NSEs exceed SEs by 10% to 60% (Menkveld et al., 2024; Soebhag et al., 2024; Walter et al., 2024; Fieberg et al., 2024). However, NSEs are not uniform across all scenarios. For instance, they are relatively higher for economic risk, long-term reversal, and dividend yield, but lower for market beta.

A closer examination of different decision nodes reveals ways to reduce NSEs. While variations across branches of the multiverse are not dramatic, certain methodological choices noticeably enhance result stability. Using country-level data from Compustat or Datastream, focusing on developed markets, and reducing rebalancing frequency tend to lower NSEs. In contrast, relying on ETFs or GFD data, covering non-developed markets, and requiring many countries per portfolio increases sensitivity to research design choices. This pattern points to a simple reporting standard: large anomaly payoffs should be shown alongside lower-NSE and more investable specifications.

Related literature. Our study aligns with two principal strands of finance literature. First, we contribute to the growing body of evidence on the role of nonstandard errors in asset pricing research. The seminal experimental study by Menkveld et al. (2024) uncovered significant variability in empirical findings across different research teams. Similarly, Mitton (2022) identified this issue in corporate finance research. Our work closely relates to studies by Coqueret (2023), Soebhag et al. (2024), Walter et al. (2024), Ciruli et al. (2025), and Coqueret and Pérignon (2025), which examined research designs in portfolio sorts and highlighted the significant sensitivity of conclusions to seemingly minor methodological choices. While few studies have extended the concept of nonstandard errors to other asset classes, such as cryptocurrencies (Fieberg et al., 2024, 2025a) and the carbon premia context (Bauckloh and Beyer, 2026), several have considered them as a robustness check in the context of market timing (Cakici et al., 2024). Additionally, Chen et al. (2024) and Lalwani et al. (2025) applied similar methodologies to explore methodological variation in machine learning literature. To our knowledge, we are the first to examine nonstandard errors in the context of cross-country variation in expected stock returns. Notably, this setting is novel not only because it introduces the concept to a new asset class, but also because it applies it to a low-dimensional cross-section, where the limited number of units (countries) may amplify the sensitivity of empirical conclusions to methodological choices.

Notably, a related strand of literature examines the robustness of cross-sectional return predictability, primarily in firm-level settings. For example, Hou et al. (2020), Green et al. (2017), Jacobs and Müller (2020), and Hollstein (2022) show that many stock-level anomalies weaken under alternative specifications or after publication. In a similar vein, Bali and Cakici (2008) scrutinize the robustness of the idiosyncratic risk anomaly. In our work, we treat this methodological uncertainty as the central object of analysis. We explicitly address research design uncertainty by varying key decision nodes and systematically decomposing how these choices shape the evidence on country-level return predictability.

Second, we contribute to the discussion on cross-country variation in equity returns. Early research attempted to extend the Capital Asset Pricing Model (CAPM) to international markets, achieving relatively strong results—particularly in developed countries (Harvey, 1991). Subsequent studies, often considering partial market segmentation, documented that local factors can also influence the cross-section of country equity returns (Harvey, 1995; Bekaert and Harvey, 1995; Harvey, 2001). These factors include economic, liquidity, credit, political, and idiosyncratic risks (Erb et al., 1995, 1996; Lee, 2011; Avramov et al., 2012; Bali and Cakici, 2010; Lehkonen and Heimonen, 2015; Chen and Demirer, 2022; Cakici and Zaremba, 2023; Gala et al., 2023). Closely related, Hou et al. (2011) study global return predictability using firm-level data and show that a small set of characteristics, such as momentum and cash-flow-to-price, captures much of the common variation in returns. While both studies rely to some extent on firm-level data, their approaches differ from ours. Hou et al. (2011) estimate stock-level premia and use the resulting factors to explain returns, whereas we focus directly on country-level premia constructed through portfolio sorts.

Another strand of studies has explored country-level equivalents of firm-level anomalies, such as momentum, value, size, reversal, seasonality, low risk, and analysts' recommendations (Balvers et al., 2000; Chan et al., 2000; Bhojraj and Swaminathan, 2006; Asness et al., 2013; Frazzini and Pedersen, 2014; Keloharju et al., 2016, 2021; Fisher et al., 2017; Berkman and Yang, 2019; Baltussen et al., 2021). While most research focuses on individual predictors, some studies examine broader sets of factors and combine them using alternative methods (Calice and Lin, 2021; Fieberg et al., 2025b; Umar et al., 2024). These studies provide the closest benchmark for our analysis because they use country-level portfolio sorts to identify cross-sectional return patterns. Against this backdrop, we revisit the robustness of documented premia and examine how methodological choices shape the conclusions.

The remainder of this paper proceeds as follows: Section 2 summarizes data and methods. Section 3 presents the findings and discusses several aspects: the impact of individual design decisions, aggregate significance based on block-bootstrap methods, asset-pricing tests, the calculation of nonstandard errors, and ways to reduce them. Section 4 concludes the study.

2. Data and methods

Our study investigates how different methodological choices affect cross-sectional variation in country equity returns. To this end, we reproduce an array of return-predictive signals from equity markets and translate them into long-short country-allocation strategies using various methodological approaches. The following section describes the predictive signals and methodological designs considered, as well as the methods used to evaluate portfolio performance.

2.1. Country characteristics

Our approach provides a comprehensive overview of prominent predictors of country equity returns. Rather than reproducing an extensive set of firm-level characteristics, we focus on well-established signals already examined in the literature. However, we impose no restrictions on their type or underlying economic intuition. Accordingly, we include both classical country risk factors and practitioner signals, such as the moving average and the 52-week high effect.

Despite this broad framework, the characteristics are constrained by several key criteria. First, the relevant strategies must be implementable via sorts—i.e., by grouping countries into international portfolios. As a result, we exclude time-series predictors used solely to forecast time variation in equity premia for a single equity market (e.g., Welch and Goyal, 2008; Goyal et al., 2024). Similarly, we discard popular calendar anomalies (e.g., Białkowski et al., 2012; Zhang and Jacobsen, 2021). Second, the anomalies must be observable through monthly returns. Therefore, patterns derived from daily or intraday data are not considered. Third, the selected characteristics should be easily obtainable from established and accessible databases. Consequently, we exclude variables such as music sentiment derived from Spotify data (Edmans et al., 2022) or transformations of analysts' forecasts (Berkman and Yang, 2019), as these are not readily available in standard datasets.

Fourth, for accounting and market characteristics, we limit our selection to variables that can be sourced consistently across all databases. This ensures uniformity and comparability across the different decision nodes in our framework. For instance, while practitioner literature highlights a variety of country-level valuation ratios (e.g., Umar et al., 2024), some—such as enterprise multiples or cash flow yields—have proven particularly effective (Hou et al., 2011; Zaremba and Szczygielski, 2019). However, such ratios are not extractable from specific data sources, like Global Financial Data, and their inclusion would compromise the consistency of implementation across decision

nodes.

Lastly, we only consider signals that have been specifically investigated in studies on the cross-section of country equity returns. As a result, we focus on established characteristics examined in the finance literature rather than attempting to replicate every possible accounting or market characteristic from the firm-level universe.

Our final selection comprises 15 market characteristics, listed in Table 1 with calculation details. For clarity, we group them into four categories based on common economic intuition: country risk, momentum, value, and others, as discussed below.

2.1.1. Risk

A large empirical literature shows that various dimensions of country-specific risk are priced in returns, particularly in less integrated markets. These include economic risk (*EconRisk*) and political risk (*PolRisk*) (e.g., Erb et al., 1995, 1996; Bekaert et al., 1996; Diamonte et al., 1996; Dimic et al., 2015; Lehkonen and Heimonen, 2015; Cakici and Zaremba, 2023; Gala et al., 2023). While different metrics have been used, we rely on composite scores from the International Country Risk Guide (ICRG), which are widely employed in asset pricing studies and provide extensive historical and cross-country coverage.

Another category is financial and sovereign risk, the country-level analogue of credit risk. Erb et al. (1995) show that countries with higher credit risk earn higher returns, a finding supported by Avramov et al. (2012), who document a credit premium in global stocks. Following Avramov et al. (2012), we proxy sovereign risk using a quantified version of sovereign ratings from major credit rating agencies (*SovRat*).

Finally, we include risk measures based on factor models and the CAPM, namely systematic and idiosyncratic risk. Systematic risk is captured by the world market beta (*Beta*). While early evidence (e.g., Harvey, 1991) indicates a positive relation between beta and future country returns, Frazzini and Pedersen (2014) argue that the low-beta anomaly extends to country indices, with low-beta markets outperforming high-beta markets on a risk-adjusted basis. In contrast, the relation between idiosyncratic risk (*IVol*) and future performance is ambiguous at the firm level (e.g., Ang et al., 2006, 2009; Bali et al., 2005; Bali and Cakici, 2008; Fu et al., 2009), but appears positively priced for country indices (Bali and Cakici, 2010), depending also on the measurement method (Mercik, 2023).

2.1.2. Momentum

The second group of characteristics relates to the momentum effect. Jegadeesh and Titman (1993) demonstrate that the best-performing stocks continue to outperform, while the worst-performing stocks persistently underperform. Similar momentum patterns have been observed in country equity indices (e.g., Chan et al., 2000; Durham, 2001; Bhojraj and Swaminathan, 2006; Asness et al., 2013; Pitkäjärvi et al., 2020; Baltussen et al., 2021). Our baseline momentum signal is the 12-month trailing return, excluding the most recent month (LtMom). Additionally, we consider several related and well-established signals.

Du (2008) and Malin and Bornholt (2010) provide evidence for a market-level equivalent of the 52-week high effect (*H52*), which was initially documented by George and Hwang (2004).[1] Moreover, Blitz and Van Vliet (2008), Zaremba et al. (2019), and Baltussen et al. (2019) reveal that country portfolios, unlike individual stocks, exhibit

[1] Notably, for H52, we exclude the most recent price observation ($t-1$) from the lookback window when computing the maximum price. In a small cross-section of country indices, including $t-1$ often assigns a value of one to multiple markets in upward-trending periods, creating ties that complicate portfolio formation. Defining the maximum over $t-12$ to $t-2$ avoids this mechanical bunching by breaking ties among markets that reach new highs, while leaving the ranking unchanged otherwise. For consistency, we apply the same convention to the moving-average signal (MA).

Table 1
Country characteristics.

Feature	Symbol	Calculation	Portfolio	Key references
Country risk				
Economic risk	EconRisk	Economic risk score derived from ICRG.	H-L	Erb et al. (1996), Cakici and Zaremba (2023)
Political risk	PolRisk	Political risk score derived from ICRG.	H-L	Diamonte et al. (1996) Dimic et al. (2015)
Sovereign risk	SovRat	Average sovereign rating from S&P500, Moody's, and Fitch, converted to a numerical scale.	H-L	Avramov et al. (2012)
Idiosyncratic risk	IVol	Idiosyncratic volatility estimated using the world CAPM model over a 60-month period (min. 12 months required).	H-L	Bali and Cakici (2010), Umutlu (2015)
World market beta	Beta	Slope coefficient from the regression of world market portfolio returns, calculated over 60 months (minimum 12 months required).	H-L	Bali and Cakici (2010), Frazzini and Pedersen (2014)
Momentum				
Short-term momentum	StMom	Total return in the previous month (t-1).	H-L	Zaremba et al. (2019)
Long-term momentum	LtMom	Total return over the months t-12 to t-2.	H-L	Asness et al. (2013)
52-week high	H52	Ratio of the price at t-1 to the maximum price observed during the period t-12 to t-2.	H-L	Du (2008), Malin and Bornholt (2010)
Moving average	MA	Ratio of the price at t-1 to the average price over the period t-12 to t-2.	H-L	Clare et al. (2016), Baltussen et al. (2021)
Value				
Dividend yield	DY	Trailing 12-month dividend yield.	H-L	Keppler (1991), Baltussen et al. (2021)
Earnings-to-price ratio	EP	Ratio of trailing 12-month earnings to current market value	H-L	Altigan et al. (2025)
Long-run reversal	LtRev	Total return over the months t-60 to t-13.	L-H	Richards (1997)
Other				
Seasonality	Seas	Average return in the same calendar month over the past 20 years (t-20 to t-1), with a minimum requirement of five years of data.	H-L	Keloharju et al. (2021)
Size	Size	Total market capitalization at t-1.	L-H	Kepler and Traub (1993), Fisher et al. (2017)
Issuance	Iss	Difference between the 36-month log-change in market value and the 36-month log-change in price.	L-H	Long et al. (2024)

The table presents this study's country characteristics and their calculation details. *Symbol* denotes the running symbol used in the manuscript. The last column, labeled Portfolio, indicates whether spread strategies take a long position in the top or the bottom quantile (high-minus-low, H-L) or the bottom quantile

(low-minus-high, L-H). The last column contains key references for calculating the return-predicting signals.

short-term momentum (StMom) rather than reversal. Finally, studies by Clare et al. (2016), Hurst et al. (2017), and Baltussen et al. (2021) explore trading strategies based on moving averages (*MA*), which also contribute to the broader understanding of momentum effects in aggregate equity markets.

2.1.3. Value

The third category encompasses proxies for the value effect. Initially documented for individual stocks (Basu, 1983; Rosenberg et al., 1985), this phenomenon also holds at the aggregate level: markets with low valuation ratios tend to outperform those with high valuation ratios (Keppler, 1991; Asness et al., 1997, 2013; Ferson and Harvey, 1997; Durham, 2001; Desrosiers et al., 2004; Hjalmarsson, 2010; Angelidis and Tessaromatis, 2017; Lawrenz and Zorn, 2017; Baltussen et al., 2021; Atilgan et al., 2025). While various value indicators are acknowledged in the literature, we focus on those consistently available across all our databases: the earnings-to-price ratio (*EP*) and the dividend yield (*DY*). Additionally, we include the long-term reversal (*LtRev*) effect in this category. This closely related phenomenon, first identified by De Bondt and Thaler (1985), has also been observed in equity indices (Richards, 1997; Balvers et al., 2000; Balvers and Wu, 2006; Spierdijk et al., 2012; Zaremba et al., 2020).

2.1.4. Other

The final class of predictors includes signals not explicitly tied to the above categories. The cross-sectional seasonality effect (*Seas*) originates from Heston and Sadka (2008), who demonstrate that stocks with high same-month returns tend to outperform their low same-month-return counterparts. Recent studies (Keloharju et al., 2016, 2021; Baltussen et al., 2021; Song and Balvers, 2022) confirm similar patterns for country indices. The *Size* effect, analogous to Banz's (1981) small-firm anomaly, suggests that smaller markets outperform larger ones (Keppler and Traub, 1993; Bekaert et al., 1996; Fisher et al., 2017; Zaremba et al., 2020; Calice and Lin, 2021). Lastly, we account for the country-level *Iss*uance effect. Pontiff and Woodgate (2008) observe that firms with high *Iss*uance activity underperform those with low *Iss*uance levels. Similarly, aggregate *Iss*uance activity has been shown to predict market returns (Baker and Wurgler, 2000; Boudoukh et al., 2007), while Long et al. (2024) demonstrate its predictive power for the cross-section of country returns.

2.1.5. From characteristics to factors

We evaluate predictor performance by constructing long–short portfolios (factors). Each month, we take a long position in the group of markets (quantile) with the highest expected returns and a short position in those with the lowest. Rankings are determined by the sorting variable, following prior literature. For example, for the market capitalization (Size) anomaly, the long portfolio comprises markets with the lowest values—i.e., countries with low market capitalization—whereas for momentum anomalies, it includes assets with the highest values, such as those with strong past returns.

After formation, portfolios are held for one month and then rebalanced. While the overall structure is consistent, the devil is in the details. As discussed in the next section, seemingly minor methodological choices can yield a wide range of alternative implementations.

2.2. Decision nodes

Forming country-level factor portfolios involves a series of research design choices, many of which may appear arbitrary. These include the selection of country indices, sample cleaning and filtering, and portfolio construction methods. To account for a multiverse of possible implementations, we adopt the approach of Walter et al. (2024), drawing on concepts from the operations research literature (e.g., Kamiński et al., 2018). This framework treats each choice as a decision node: at each step, a researcher selects among alternatives, creating branches that lead to subsequent nodes. Combinations of these choices yield terminal nodes representing distinct research design specifications.

Fig. 2 presents a flow chart of our decision nodes. We consider nine common methodological choices in country-level cross-sectional studies: one relates to the data source, three to sample preparation, and five to portfolio construction and implementation. Below, we describe each node in detail.

2.3. Data sources and country representation

The first key decision concerns the representation of aggregate country-level returns, which is inherently linked to the choice of data source. Empirical research has developed several prominent approaches. The most common approach relies on reputable total-return equity indices, offering at least three advantages. First, it largely eliminates arbitrary decisions in country portfolio construction, as these are effectively "outsourced" to the index provider. Second, indices from a single provider ensure consistent design across markets. Third, through universe selection and construction, such indices typically emphasize larger firms, enhancing investability.

Three providers are particularly prominent. MSCI indices are widely used in empirical studies (e.g., Richards, 1997; Balvers et al., 2000; Asness et al., 2013; Keloharju et al., 2016; Calice and Lin, 2021) due to their broad coverage and transparency. They represent value-weighted portfolios covering about 85% of the largest and most liquid firms in each country. Importantly, MSCI does not apply retroactive changes to reported returns, reducing potential biases. These indices can be obtained directly from MSCI or via databases such as Bloomberg, which also provide additional features like valuation ratios.

Datastream Global Equity Indices are another popular choice (e.g., Bali and Cakici, 2010; Umutlu, 2015; Atilgan et al., 2019; Zaremba et al., 2020; Umar et al., 2024). They offer comparable coverage and history to MSCI and similarly represent value-weighted portfolios of large firms, covering roughly 75%–80% of total market capitalization. However, their main drawback is limited accessibility and transparency: data are available only through LSEG (formerly Refinitiv, Thomson Reuters), and detailed construction documentation is scarce.

Studies using very long samples often rely on indices from Global Financial Data (GFD) (e.g., Hjalmarsson, 2010; Zhang and Jacobsen, 2013; Albuquerque et al., 2015; Bekaert and Mehl, 2019; Baltussen et al., 2021; Cakici and Zaremba, 2023). These series can be traced back to the 19th century in many developed and emerging markets. While offering a unique historical perspective, they may suffer from biases or omissions, and their portfolios often include relatively few securities, limiting representativeness.

Beyond established index providers, the literature proposes at least two additional approaches. One involves directly aggregating individual stocks into country portfolios using CRSP and Compustat data, as in Fieberg et al. (2025b). This method has certain advantages: it captures the entire cross-section of local stock markets, including small and microcap companies, and offers substantial country coverage with data that can be structured efficiently. For example, Jensen et al. (2023) provide data for up to 93 countries.[2] However, this method has a significant caveat: the research sample may include assets that are difficult to invest in, such as microcap stocks with low liquidity or assets in segmented frontier markets. Consequently, return patterns detected using this approach may be challenging to translate into actionable trading strategies.

Lastly, some approaches take a more practical perspective by focusing on investable instruments that provide exposure to

[2] The data is available from https://jkpfactors.com/.

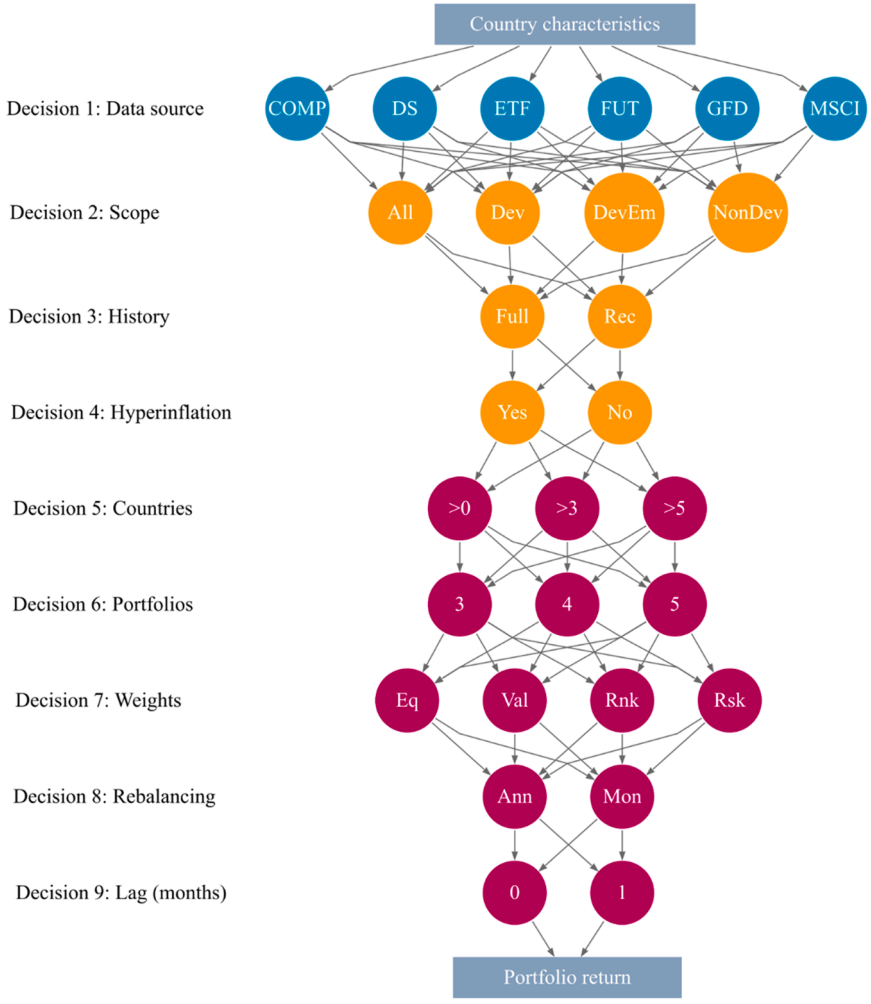

Fig. 2. Decision process for portfolio tests.
The figure illustrates the decision nodes involved in portfolio tests. Using country characteristics as inputs, we evaluate paths through nine distinct decision nodes to arrive at the final portfolio return. The first decision node involves selecting the data source for representing country portfolios: COMP, DS, ETF, FUT, GFD, or MSCI. The next three decision nodes pertain to sample preparation: the scope of markets (all markets, developed markets only, or both developed and emerging markets, non-developed only), the historical perspective (full history or recent data), and hyperinflation episodes (included or excluded). The final five decision nodes relate to portfolio construction: the minimum number of assets per portfolio (3, 5, or no restriction), the number of portfolios in the cross-section (3, 4, or 5), the weighting scheme (equal-, value-, rank-, or risk-weighted), the rebalancing frequency (monthly or annually), and the implementation lag (1 month or no lag).

international markets, such as single-country exchange-traded funds (ETFs) (e.g., Andreu et al., 2013; Breloer et al., 2014; Smith and Pantilei, 2015) and index futures (e.g., Moskowitz et al., 2012; Hurst et al., 2017; Baltussen et al., 2021; Fieberg et al., 2025b). While these frameworks align closely with investors' concerns and largely mitigate exchange rate issues, they may limit dataset size. Moreover, the most liquid futures are typically based on local equity indices, which differ in design and coverage, creating inconsistencies across markets.

In this paper, we consider six distinct representations of country equity returns: (i) value-weighted country portfolios aggregated from stock-level CRSP and Compustat data (denoted for brevity COMP); (ii) Datastream Global Equity Indices (DS) accessed via LSEG; (iii) MSCI Indices (MSCI) obtained directly from MSCI; (iv) GFD Global Equity Indices accessed from Global Financial Data (GFD); (v) iShares single-country ETF returns available through Datastream; and (vi) index futures returns sourced from Bloomberg.

Fig. 3 illustrates the global representation of these datasets over time. CRSP/Compustat provides the most comprehensive coverage of countries throughout the sample period, particularly in recent years. In contrast, during the early years of the sample, GFD takes the lead, encompassing 25–30 markets. Overall, MSCI and Datastream also include a broad set of countries, whereas datasets based on ETFs and index futures represent a relatively smaller number. Notably, the number of countries covered has increased across all data sources over time, except for GFD, which has remained nearly constant over the past two decades.

All calculations in this study are based on total returns, with prices and returns expressed in US dollars. To ensure full consistency, all market characteristics are obtained from the same sources as the indices themselves. For example, the valuation ratios for Datastream (DS) indices are sourced from Datastream. The only exceptions are exchange-traded funds (ETFs) and futures. Most ETFs in our sample predominantly track MSCI indices, so their valuation ratios (e.g., dividend yield (DY), earnings-to-price ratio (EP)), as well as *Size* (*Size*) and *Iss*uance (*Iss*) features, are calculated based on MSCI index data. A similar approach is applied to equity futures, as their underlying assets are typically local indices, for which comprehensive and consistent historical composition data are often unavailable in major databases such as Bloomberg or Datastream. Additionally, concepts like issuance or size do not directly apply to futures (derivatives) but to underlying equity markets.

It is important to note that we focus on the most common approaches and do not cover less frequent alternatives used in some studies. For

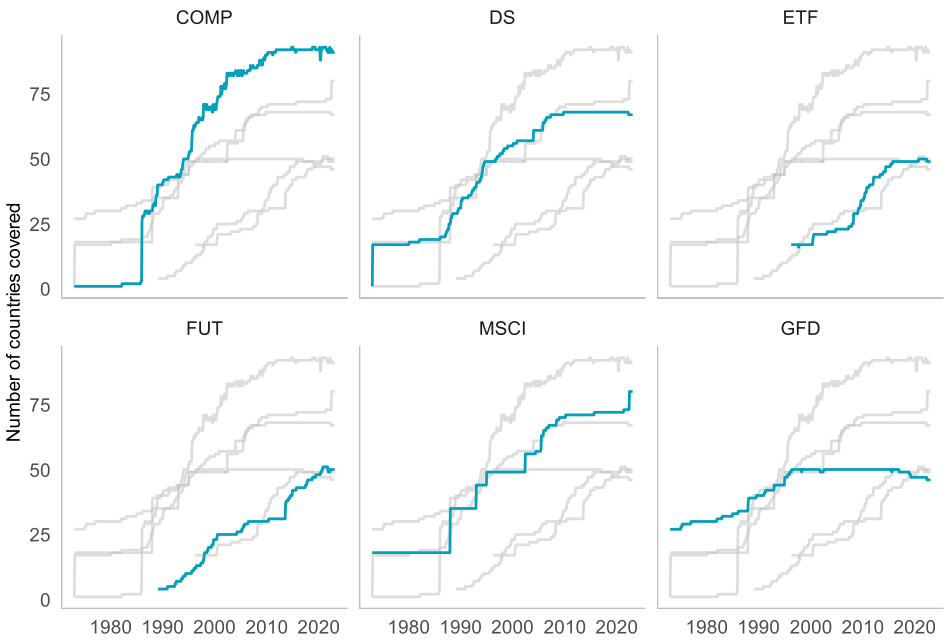

Fig. 3. Country coverage across data sources.
The figure depicts the number of countries covered by various data sources representing country-level investments: CRSP/Compustat portfolios (COMP), Datastream indices (DS), exchange-traded funds (ETF), index futures (FUT), MSCI indices (MSCI), and GFD indices (GFD). The study period spans from January 1973 to December 2022. The blue line represents the primary data source indicated in each panel's heading, while the grey lines show the other data sources included for comparison.

example, Spierdijk et al. (2012) rely on long-run returns from Dimson et al. (2002). Some studies combine multiple data sources, such as MSCI with the International Finance Corporation (Erb et al., 1995), MSCI with Datastream (Avramov et al., 2012), or Bloomberg with Global Financial Data (GFD) (Baltussen et al., 2019, 2021; Cakici and Zaremba, 2024). While merging sources can expand coverage, it may introduce design inconsistencies across index providers.

Additionally, some studies rely on local indices. For instance, Chan et al. (2000) use national indices compiled by local providers such as DAX, S&P, and NIKKEI. This approach offers two advantages. First, local indices often have longer histories than MSCI or Datastream, increasing sample length. Second, they align more closely with investment practice, as the most liquid equity index futures are typically tied to local benchmarks, such as Japan's NIKKEI or Poland's WIG20. However, a key drawback is limited computational consistency: differences in index construction and weighting across providers may affect results and introduce biases. For example, apparent outperformance may reflect greater exposure to small-cap stocks rather than the factor of interest.

In sum, while we focus on the most widely used data sources and representations of country returns, the range of methodological designs employed in practice is considerably broader.

2.4. Sample preparation

With the market data in hand, researchers proceed to sample preparation. This stage typically involves selecting the relevant markets, defining the study period, and cleaning the data. We review three potential design choices in this context.[3]

2.4.1. Scope

The cross-section of markets examined in country-level studies can vary widely, from a handful to nearly 100. The most conservative approach focuses on a small subset of the most developed markets, ensuring tradability and relative liquidity. For instance, Pitkäjärvi et al. (2020) examine 20 markets, while Balvers et al. (2000), Desrosiers et al. (2004), and Asness et al. (2013) analyze 18 countries. Keloharju et al. (2016, 2021) further reduce this number, limiting their scope to just 15 countries.

Less conservative approaches include approximately 35–50 countries, covering both developed and emerging markets (e.g., Bhojraj and Swaminathan, 2006; Bali and Cakici, 2010; Clare et al., 2016). Some studies adopt an even broader scope, sometimes analyzing more than 70 countries, including less-tradable frontier markets. For example, Avramov et al. (2012) explore credit risk premia across 75 markets. Finally, certain studies deliberately focus on non-developed markets to capture their distinctive characteristics (e.g., Harvey, 1995; Bekaert et al., 1996; Kiviaho et al., 2014; Lehkonen and Heimonen, 2015; Li and Pritamani, 2015; Zaremba et al., 2022).

The choice of market scope can have substantial implications, as less developed markets are often considered reservoirs of anomalies and

[3] Importantly, the stock-level literature acknowledges a range of sample-cleaning procedures, including filtering out microcaps, excluding extreme returns, or applying return winsorization. However, these techniques are generally less relevant at the country-level. For instance, removing return outliers is often intended to correct for data issues such as misrecorded price splits—errors that are far less common when working with portfolio-level data. Similarly, winsorizing returns or variables, even at relatively high thresholds (e.g., 1%), has a negligible impact in a cross-section comprising only a few dozen country-level observations.

mispricing. With their unique economic and institutional structures, these markets may provide insights that are unavailable in more efficient, developed markets. To address the diversity of design choices, we consider four distinct samples in this study: i) all markets (All): includes all available markets; ii) developed markets only (Dev): focuses exclusively on developed markets; iii) developed and emerging markets (DevEme): combines developed and emerging markets; iv) non-developed markets (NonDev): Excludes developed markets. To classify markets, we follow the MSCI market classification framework and dynamically account for all market reclassifications over time.[4]

2.4.2. History

Our baseline sample spans January 1973 to December 2022, which broadly aligns with the coverage used in much of the country-level literature. Nevertheless, some studies rely on shorter samples due to data availability or a focus on more recent market environments—for example, Avramov et al. (2012) begin their time series in 1989, Fisher et al. (2017) in 1990, and Gala et al. (2023) in 1992. To reflect this variation in empirical design, we report results for both the full available sample (Full) and a more recent subsample starting in January 1990 (Rec). While longer samples enhance statistical power, earlier periods may be more prone to errors and less representative of current market conditions, given structural changes in financial markets—including shifts in accounting standards and disclosure practices, decimalization, trading automation, and evolving international market integration.

Importantly, the coverage of certain characteristics begins later than the coverage of the country equity indices themselves. For example, the long-term reversal (LtRev) and idiosyncratic volatility (IVol) signals mechanically discard the first five years of observations due to their estimation windows. Likewise, the ICRG risk indicators and a sufficiently broad sample of sovereign ratings became available only in the 1980s. To avoid introducing arbitrary truncation, we use all available observations for each characteristic, so the effective sample period varies depending on data availability.

2.4.3. Hyperinflation

Hyperinflationary episodes can significantly distort stock performance, often resulting in extreme market returns, particularly when measured in local currency. Additionally, these events heighten the risk of return miscalculations due to missing or erroneous data or complications with currency conversions. For these reasons, certain studies exclude hyperinflationary periods from their datasets. While this filter is not widespread, it appears in certain studies, particularly those involving long-run historical data or broad samples (e.g., Baltussen et al., 2021; Cakici and Zaremba, 2024). We incorporate this consideration as well, testing samples both with and without the exclusion of hyperinflationary periods. To identify such periods while avoiding look-ahead bias in sample construction, we replicate the screening methodology of Baltussen et al. (2021), which follows the definition by Cagan (1956). Specifically, if the monthly inflation rate exceeds 50%, we exclude all observations from the subsequent 12 months. In practice, this filter removes a small number of observations from a limited set of countries.

2.5. Portfolio construction

After finalizing the country sample, the next step is the formation of factor portfolios. Variations in breakpoints or weighting schemes can impact the outcomes. In this process, we identify five critical methodological choices: the number of portfolios in the cross-section, the minimum number of countries required in a portfolio, the portfolio weighting scheme, the rebalancing frequency, and the implementation lag.

[4] https://www.msci.com/our-solutions/indices/market-classification.

2.5.1. Number of portfolios

Grouping countries into portfolios serves as a robust test for monotonic relationships in the cross-section of global returns. The return differential between the top and bottom quantiles highlights systematic return patterns. However, the choice of the number of portfolios is somewhat arbitrary. While stock-level studies often use deciles (Walter et al., 2024), the smaller cross-section of countries necessitates broader groupings to avoid portfolios with only a few—or even a single—markets, which would make the results more susceptible to noise.

The most common approaches include tertiles (e.g., Daniel and Moskowitz, 2016; Asness et al., 2013; Atilgan et al., 2019; Umar et al., 2024), quartiles (e.g., Erb et al., 1995; Richards, 1997; Blitz and van Vliet, 2008; Malin and Bornholt, 2010; Gala et al., 2023), and quintiles (e.g., Berkman and Yang, 2019; Clare et al., 2016; Chen and Demirer, 2022; Cakici and Zaremba, 2023). Some studies adopt a simpler binary approach, creating just two portfolios—long and short—based on extreme quantiles (e.g., Moskowitz et al., 2012; Pitkäjärvi et al., 2020). Bali and Cakici (2010) use more flexible portfolio groupings, splitting the sample into 30%, 40%, and 30%.

A less common methodology avoids quantile-based cutoffs altogether, instead focusing on a fixed number of countries per portfolio. For example, Kortas et al. (2005) use portfolios containing the 11 most extreme countries, while Keloharju et al. (2016) examine cross-sectional seasonality using portfolios with the three indices with the highest or lowest average returns in the past.

In this study, we consider three portfolio groupings: 3, 4, and 5. These correspond to tertile, quartile, and quintile sorts, respectively.

2.5.2. Minimum number of countries

The choice of breakpoints in the previous stage is closely linked to the breadth of the cross-section, as researchers need a sufficient number of markets to justify finer sorts. This issue can be addressed in two main ways: by requiring a minimum sample size or by imposing a minimum number of countries per portfolio. Both approaches are qualitatively similar and may influence the choice of the study period, as early years with fewer available markets might need to be excluded from the analysis.

To scrutinize this aspect, we consider three possible minimum portfolio size requirements: i) one, ii) three, or iii) five countries.

2.5.3. Weighting scheme

Once portfolio countries are selected, the next key step is choosing the weighting scheme. Firm-level studies typically employ value-weighted portfolios, reflecting economic importance and investability. This approach is occasionally applied at the country-level (e.g., Chan et al., 2000; Chiah et al., 2023). However, many studies use equal-weighted portfolios (e.g., Balvers and Wu, 2006; Avramov et al., 2012; Hurst et al., 2017; Berkman and Yang, 2019; Gala et al., 2023).

Equal-weighting is favored for at least two reasons. First, value-weighted portfolios at the country level tend to be dominated by the largest markets, whereas equal-weighting mitigates this concentration. Second, for some country proxies, such as index futures, aggregate market values may be unavailable or lack a clear counterpart. However, equal-weighting can tilt portfolios toward small and illiquid frontier markets, where large positions may be infeasible. Moreover, performance may be affected by “rebalancing” or “diversification” effects arising from the periodic portfolio reconstruction (see Willenbrock, 2011; Hallerback, 2014; Swade et al., 2023).

Beyond classical value- and equal-weighted schemes, some studies employ alternative—though less common—frameworks. Clare et al. (2016) and Moskowitz et al. (2012) use a risk-parity approach, weighting components inversely to their volatility. Others, such as Ilmanen et al. (2021) and Asness et al. (2013), scale weights with the rank or value of underlying characteristics, assigning larger weights to more extreme signals. However, rank-based weighting can overweight

microcap stocks (Novy-Marx and Velikov, 2022). This concern is less pronounced at the country level, where portfolios are built from country indices that already emphasize large firms, although risk-based schemes may still tilt portfolios toward smaller or more volatile markets.

In this study, we explore four different weighting schemes: i) equal-weighted portfolios; ii) value-weighted portfolios; iii) risk-weighted portfolios, where a country's weight is proportional to the inverse of its 24-month trailing return standard deviation; iv) rank-weighted portfolios, where a country's weight is tied to the absolute value of its de-meaned cross-sectional rank, following the approach of Asness et al. (2013).

2.5.4. Rebalancing

Portfolio weights change as market values evolve and country characteristics update. Markets may also enter or exit the long and short legs. Empirical studies therefore rebalance portfolios periodically to maintain intended exposures. Most studies rebalance monthly (e.g., Asness et al., 1997, 2013; Avramov et al., 2012; Bali and Cakici, 2010), though some adopt lower frequencies, including quarterly or annual rebalancing (e.g., Angelidis and Tessaromatis, 2017; Balvers et al., 2000; Balvers and Wu, 2006; Bhojraj and Swaminathan, 2006; Chen and Demirer, 2022). Monthly rebalancing captures changes in portfolio characteristics more closely, but it also increases turnover. This concern may be especially pertinent at the country level, where portfolios often include only a few markets, so even a single replacement can materially change portfolio composition.

To evaluate this methodological decision, we consider two rebalancing frequencies: i) monthly and ii) annual. These choices span the main trade-off between timely signal updating and lower turnover.

2.5.5. Implementation lag

Our final decision node concerns the implementation lag between the calculation of a country characteristic and the subsequent portfolio formation. Many studies assume that portfolios are formed immediately after the signal is calculated, implying that investors use information available at the end of month t-1 to predict returns in month t. For example, price-based signals such as momentum typically assume that the closing price used to compute the signal coincides with the trade execution. While this assumption is not implausible, it may be challenging in practice—particularly when investors must reallocate capital across countries.

Some works, therefore, introduce an explicit weekly (e.g., Chan et al., 2000) or monthly (e.g., Balvers and Wu, 2006) delay between signal calculation and return measurement. In settings where signals rely on persistent characteristics, such as country risk or sovereign credit ratings, the lag may be longer—extending even to six months (Erb et al., 1995). To account for this variation in the literature, we consider two implementations: (i) no implementation lag and (ii) a one-month skip period between signal calculation and portfolio formation.

The nine decision nodes presented above generate 13,824 possible specifications, illustrating the potential magnitude of nonstandard errors in empirical tests. A limitation of this framework, however, is that the decision nodes are not fully independent. Some methodological choices are naturally related and would rarely be combined in practice. For example, the number of portfolios and the minimum number of countries per portfolio are closely linked design decisions. When the cross-section of available markets is small, researchers are unlikely to employ finer portfolio sorts, and requiring a high minimum number of countries per portfolio becomes difficult to reconcile with narrow samples. In some cases, particular specifications may even become infeasible—for instance, when a restricted sample (such as developed markets only) does not provide enough countries in a given month to satisfy both diversification and portfolio-granularity requirements.

Notably, the last two decision nodes—rebalancing frequency and implementation lag—primarily concern practical implementation rather than the identification of cross-sectional return predictability itself. Nevertheless, their treatment varies across studies and may materially affect the estimated magnitude and statistical significance of factor premia. Moreover, both choices are routinely incorporated in studies examining nonstandard errors in asset pricing (e.g., Soebhag et al., 2024; Walter et al., 2024; Fieberg et al., 2024, 2025b). For these reasons, we include them in the research design space considered in our analysis.

To ensure robustness and feasibility, we impose a minimum threshold of 60 months of return data. This constraint reduces the number of finally analyzed strategies to between 12,736 and 13,824, depending on the strategy.[5]

2.6. Methodological choices in the literature

As the preceding sections illustrate, researchers studying cross-country return predictability face a wide range of methodological choices. Which of these designs prevail in practice? To answer this question, we survey 52 representative studies on the cross-section of country equity returns and systematically document their empirical design decisions in several key aspects: the number of markets in the sample, the length of the study period, market coverage, treatment of hyperinflation, portfolio sorting procedures, rebalancing frequency, and index representation. The detailed evidence appears in Table A1 in the Online Appendix, while Fig. 4 summarizes the distribution of the most common implementations.

Several clear patterns emerge. First, with respect to country representation, researchers predominantly rely on MSCI indices, which appear in 61% of studies. Datastream indices follow at a distance (24%), while futures, Global Financial Data, Compustat-based aggregates, and ETFs are used more sporadically. Second, most works employ relatively broad cross-sections, with the modal sample covering between 16 and 45 markets. Smaller samples typically coincide with earlier studies, when time series from emerging markets were short or unavailable. In contrast, more recent studies expand coverage to emerging markets (35% of papers), although frontier countries remain comparatively underexplored (19%). Notably, nearly all studies (94%) include developed markets.

Sample length reflects both data availability and evolving research objectives. Most studies rely on long time spans—typically between 16 and 45 years—though some extend considerably further, occasionally exceeding 75 years. Importantly, the length of our "recent" sample, starting in 1990 (see Section 2.4.2), coincides with the median of the empirical distribution. Exactly half of the surveyed studies use shorter samples, while the other half rely on longer ones. This places our choice squarely within the range of commonly used sample lengths in the literature. At the same time, the exclusion of hyperinflationary episodes is rare. Only about 6% of studies apply such filters, typically in the context of very long historical samples. We therefore treat this choice as a useful but narrow fork in research design rather than as a central source of methodological variation. Fig. 4 describes common practice, not a ranking of methodological importance or economic suitability.

Portfolio construction choices exhibit both concentration and variation. The most common approach sorts countries into three portfolios (tertiles), followed by quartile and quintile splits, while finer partitions prove rare. Equal-weighting clearly dominates, appearing in 77% of studies, whereas value-weighting and alternative schemes—such as rank- or risk-based weighting—are used less frequently. Finally, most studies rebalance portfolios monthly, with some opting for lower frequencies, such as quarterly or annual rebalancing.

Overall, the evidence highlights several recurring patterns in the literature rather than a single dominant design. Typical studies rely on

[5] The reduced numbers are 13,152 for *Beta, H52, IVol, LtMom, MA*, and *Seas*; 13,120 for *EP, DY, EconRisk, ISS, PolRisk*, and *SovRat*; and 12,736 for *LtRev*. *Size* and *StMom* implementations remain unaffected.

Fig. 4. Design choice popularity in the finance literature.
The figure displays the popularity of several design choices in the studies of the cross-section of country equity returns. Panel A illustrates the country representation: country representation: Compustat (COMP) portfolios, Datastream indices (DS), MSCI indices, Global Financial Data indices (GFD), exchange-traded funds (ETF), or index futures (FUT); Panel B: number of markets in the sample; Panel C: market types in the sample: developed (Dev), emerging (Eme), or frontier (Fro): Panel D: study period; Panel E: hyperinflation exclusion (Yes or No); Panel F: type of portfolio sorts; Panel G: weighting scheme, including equal-weighted (EW), value-weighted (VW), rank-weighted (RnkW), liquidity-weighted (LW), risk-weighted (RskW), return-weighted (RetW); Panel H: rebalancing frequency (per year). The list of source studies is available from Table A1 in the Online Appendix.

MSCI (or similar) indices, employ relatively long sample periods, focus on developed markets, often augmented with emerging markets when data permit, form tertile portfolios, use equalweighting, rebalance at relatively high frequency, and avoid extensive data filtering, such as hyperinflation exclusions. Notably, these choices describe common practice, but they need not define the most appropriate benchmark. When data constraints dictate a design, the evidence should be interpreted as conditional on that feasible sample. When choices reflect researcher discretion, the stronger test is whether the result survives alternative specifications, especially those emphasizing larger, more liquid, and more investable markets.

2.7. Performance evaluation

To evaluate the performance of the portfolio strategies, we use several straightforward statistics, such as mean monthly returns and annualized Sharpe ratios. Furthermore, we also calculate alphas from the World CAPM, represented by Eq. (1)

$$R_t = \alpha_{MKT} + \beta_{MKT} MKT_t + \varepsilon_t, \tag{1}$$

where R_t denotes the excess return on the tested portfolio, MKT_t is the value-weighted return of all markets in the sample, α_{MKT} represents the abnormal return, β_{MKT} is the slope coefficient, and ε_t denotes the residual term. The risk-free rate is captured by the U.S. Treasury bill rates from Kenneth French's website.[6] Notably, to ensure apple-to-apple comparisons, the market portfolio in (1) is always constructed from the same country sample as the tested strategies. In other words, we always use the same equity universe and apply the same filters as for the examined long-short portfolios.

3. Results

We begin by presenting the baseline results, followed by a closer examination of variation across decision nodes. Next, we analyze nonstandard errors in greater detail and consider ways to mitigate them. Finally, we assess the aggregate significance of anomaly signals across the full multiverse.

3.1. Baseline findings

Fig. 5 illustrates the distribution of mean monthly returns and CAPM alphas across different research designs. Results for both performance metrics are quite similar and reveal significant dispersion in portfolio performance across implementations. Most strategies may generate both profits and losses, depending on a particular research design. The mean returns typically vary by more than one percentage point between the most and least favorable setups. However, certain factors—such as H52 and MA—demonstrate notably higher uncertainty than others.

Interestingly, the outcomes in Fig. 5 fail to provide a clear answer on the overall profitability of different country characteristics. Most strategies exhibit potential for both sizable profits and losses, with no clear dominance of one over the other. However, some signals seem more effective than others, with Size and SovRat standing out as the most profitable in terms of average means of returns across specifications.

Table 2 provides additional insights into mean monthly returns and alphas. The results point to broad fragility in country-level premia, much as in firm-level anomaly research. Across all implementations, the average long-short return is only 17 basis points per month. About 75% of implementations generate positive returns, but only 14.5% produce t-statistics above 1.96, the conventional two-sided significance threshold. Alphas display a similar pattern: 71% are positive, but only 13.7% exceed this threshold. Although this share is low, it is still five times larger than one might expect by chance.[7]

The most successful signals include market size and dividend yield, followed by economic risk, sovereign rating, and issuance. For instance, the size portfolios generate an average monthly return of 0.41% at a Sharpe ratio of 0.25 and prove profitable in 96.3% of specifications. Similarly, the dividend yield portfolios deliver a mean return of 0.23%, a Sharpe ratio of 0.25, and demonstrate profitability in 87.5% of implementations. The other measures exhibit less consistency across different measures, but political and sovereign risks, as well as issuance, display the highest average Sharpe ratios, equaling 0.2. On the other hand, the beta and seasonality prove largely irrelevant, and the entire momentum category exhibits very low mean returns and Sharpe ratios. This last finding contrasts with the substantial body of evidence emphasizing the robustness and pervasiveness of the cross-country momentum effect (e.g., Asness et al., 2013; Baltussen et al., 2021), yet aligns, for example, with Andreu et al. (2013), who find a flat momentum pattern after 2000.[8]

Notably—consistent with the earlier insights from Fig. 5—Table 2 further confirms the substantial variability in results across different implementations. The interquartile ranges of returns and alphas typically fluctuate around 0.3–0.4% per month. Only in a few cases (e.g., political risk and issuance) does this dispersion narrow to approximately 0.2%.

3.2. Performance under most common design choices

The results in the previous subsection treat all research designs symmetrically. How do they compare with common practices in the literature? To address this question, we zoom in on a single setting that reflects the methodological choices most commonly adopted in prior studies, as documented in Section 2.6. Specifically, we construct country portfolios using MSCI indices, focus on developed and emerging markets, employ equal-weighted tertile sorts, rebalance monthly, use the full available sample, and avoid additional filters such as hyperinflation exclusions or implementation lags. Table 3 reports the corresponding results.

Restricting the analysis to conventional research designs produces stronger results and moves the evidence closer to prior findings. This does not imply that conventional choices provide the preferred benchmark; rather, it shows that these choices help explain why earlier studies often find stronger country-level premia. Several well-known predictors generate economically meaningful and statistically significant returns. In particular, economic risk, various momentum measures, long-term reversal, size, and issuance all yield positive payoffs that are statistically significant at the 10% level in the full sample (Panel A). In other words, when implemented using conventional research designs, many prominent country-level signals retain their predictive power.

Admittedly, not all characteristics perform uniformly across specifications. Importantly, a lack of statistical significance in this table does not imply that prior findings fail to replicate. Many signals were originally documented under specific sample restrictions or methodological settings. For example, the seasonality effect—primarily established in developed markets—weakens in Table 3's specification but regains significance when we impose the same restriction.[9] Similarly, valuation effects such as DY and EP may be stronger in data sources that place

[6] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html.

[7] Notably, as we aim to evaluate the distribution of significant results across possible studies, we do not apply any corrections for multiple hypothesis testing (see, e.g., Harvey et al., 2016; Harvey, 2017). However, if one were to account for this effect—even after considering the lack of independence between our hypotheses—the proportion of significant outcomes would naturally be lower.

[8] Also Richards (1997) argue that cross-momentum lacks significance, emphasizing its dependence on research design.

[9] In an unreported analysis, we confirm the return predictability by Seas when the sample is limited to developed markets.

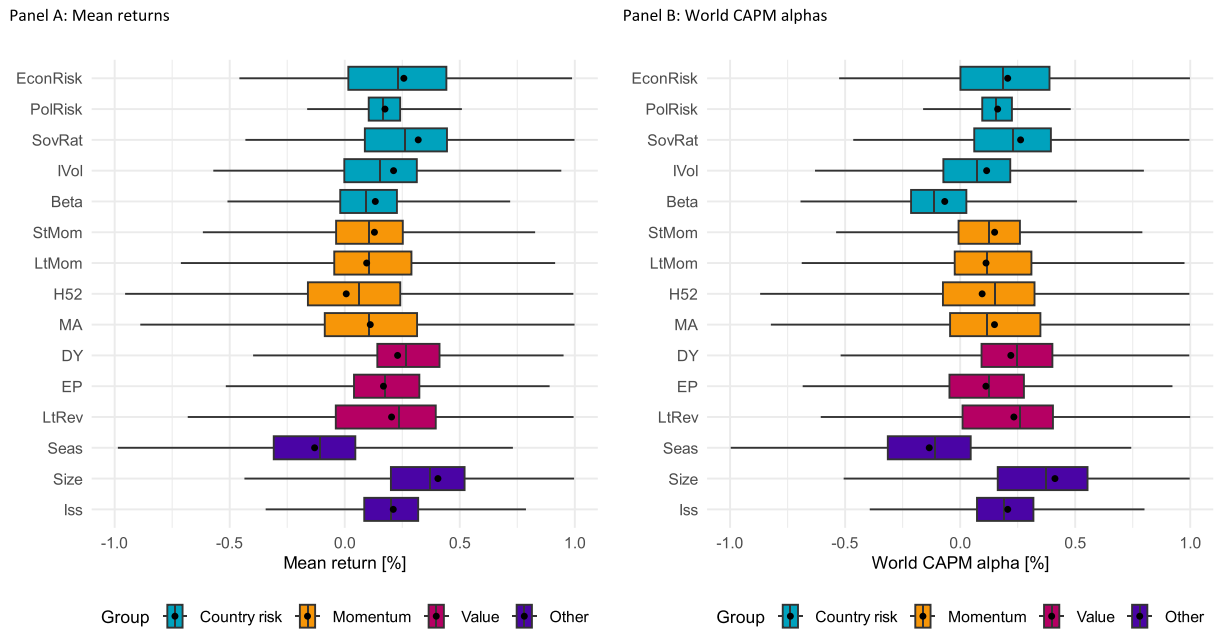

Fig. 5. Mean returns and alphas on long-short country portfolios.
The boxplots illustrate distributions of mean returns (Panel A) and world CAPM alphas (Panel B) for up to 13,824 research designs of one-way sorted long-short portfolios constructed based on 15 country characteristics (detailed in Table 1) shown on the vertical axis. The boxes represent the interquartile ranges with the median marked inside; the whiskers indicate the maximum and minimum values, while the dots denote the average value across all observations. The study period runs from January 1973 to December 2022.

Table 2
Returns on characteristic-sorted portfolios.

	Panel A: Mean returns							Panel B: World CAPM alphas					
	Average	Median	SD	IQR	% > 0	% *t*-stat > 1.96	SR	Average	Median	SD	IQR	% > 0	% *t*-stat > 1.96
Country risk													
EconRisk	0.26	0.24	0.32	0.44	77.0	25.2	0.18	0.21	0.19	0.29	0.39	75.8	16.2
PolRisk	0.18	0.17	0.17	0.14	91.4	6.8	0.20	0.16	0.15	0.17	0.13	91.0	7.5
SovRat	0.32	0.28	0.35	0.37	87.6	14.3	0.20	0.26	0.23	0.29	0.34	84.6	9.1
IVol	0.21	0.16	0.39	0.33	75.3	7.4	0.11	0.12	0.08	0.35	0.30	62.4	4.4
Beta	0.13	0.10	0.27	0.25	71.0	2.5	0.08	-0.07	-0.11	0.26	0.24	29.0	0.3
Momentum													
StMom	0.13	0.11	0.31	0.30	68.9	9.1	0.10	0.15	0.13	0.29	0.28	74.2	9.8
LtMom	0.10	0.10	0.36	0.34	67.4	13.4	0.10	0.11	0.12	0.34	0.34	70.5	13.5
H52	0.01	0.06	0.36	0.41	57.1	5.9	0.04	0.10	0.14	0.36	0.40	64.7	13.2
MA	0.11	0.10	0.35	0.41	64.6	14.9	0.10	0.15	0.12	0.35	0.40	69.7	18.9
Value													
DY	0.23	0.26	0.37	0.28	87.5	32.3	0.25	0.22	0.25	0.36	0.31	85.5	30.5
EP	0.17	0.17	0.29	0.28	80.5	14.2	0.15	0.11	0.13	0.29	0.33	70.2	8.6
LtRev	0.20	0.24	0.37	0.44	72.8	18.1	0.13	0.23	0.27	0.33	0.40	76.0	15.7
Other													
Seas	-0.13	-0.11	0.30	0.36	33.1	2.1	-0.11	-0.13	-0.11	0.28	0.36	32.7	1.9
Size	0.41	0.38	0.29	0.33	96.3	37.9	0.33	0.41	0.39	0.32	0.41	93.2	41.9
Iss	0.21	0.20	0.20	0.24	88.7	13.9	0.20	0.21	0.19	0.20	0.25	87.6	14.6
Full sample													
All	0.17	0.16	0.34	0.34	74.6	14.5	0.14	0.15	0.14	0.33	0.36	71.1	13.7

The table presents statistical properties of the performance distributions for a multiverse of up to 13,824 portfolio implementations based on 15 country characteristics, as detailed in Table 1. Panel A focuses on mean returns, and the reported statistics include the mean (*Average*), median (*Median*), standard deviation (*SD*), interquartile range (IQR), the proportion of positive mean returns (*% Mean > 0*), the proportion of mean returns with *t*-statistics exceeding 1.96 (*% t-stat > 1.96*), and the Sharpe ratio (*SR*) across all research designs. Panel B presents similar statistics for the World CAPM alphas. All values, except for the SR, are expressed in percentage terms. The study period spans January 1973 to December 2022. The bottom row summarizes the results for the pooled sample of all characteristic-sorted portfolios.

greater weight on smaller stocks, such as Compustat, which is examined, for example, by Hou et al. (2011).

The time dimension also plays a crucial role. To illustrate this, we split the sample into periods before and after 2009, which corresponds to the average end date of the original sample period in the studies surveyed in Table A1 in the Online Appendix. For the years up to 2009, return predictability is considerably stronger. For example, sovereign risk—highlighted by Avramov et al. (2012)—yields strong, significant returns in this earlier period, consistent with the original evidence.

In contrast, the post-2009 period exhibits a clear decline in return predictability. Many signals weaken substantially, both in terms of economic magnitude and statistical significance. Mean returns decrease, and Sharpe ratios fall. This pattern is consistent with growing evidence that anomaly returns have attenuated in more recent

Table 3
Characteristic-sorted portfolios under the most common design choices.

	Panel A: Full study period			Panel B: Until 2009			Panel C: After 2009		
	R	*t*-stat	SR	R	*t*-stat	SR	R	*t*-stat	SR
Country risk									
EconRisk	0.62	*(3.53)*	0.62	0.99	*(4.30)*	0.89	-0.08	*(-0.44)*	-0.12
PolRisk	0.18	*(1.50)*	0.22	0.18	*(0.88)*	0.19	0.18	*(1.71)*	0.37
SovRat	0.35	*(1.68)*	0.31	0.69	*(2.32)*	0.55	-0.26	*(-1.28)*	-0.35
IVol	0.11	*(0.41)*	0.09	0.38	*(0.86)*	0.28	-0.23	*(-1.08)*	-0.28
Beta	0.08	*(0.40)*	0.08	0.20	*(0.59)*	0.16	-0.06	*(-0.29)*	-0.07
Momentum									
StMom	0.29	*(2.02)*	0.27	0.51	*(3.30)*	0.46	-0.35	*(-1.80)*	-0.44
LtMom	0.64	*(3.19)*	0.50	0.80	*(3.04)*	0.58	0.19	*(0.89)*	0.21
H52	0.34	*(1.73)*	0.28	0.45	*(1.76)*	0.34	0.05	*(0.21)*	0.05
MA	0.57	*(2.88)*	0.45	0.76	*(3.05)*	0.56	0.04	*(0.15)*	0.04
Value									
DY	0.18	*(1.28)*	0.19	0.28	*(1.57)*	0.28	-0.10	*(-0.66)*	-0.17
EP	0.10	*(0.66)*	0.10	0.21	*(1.12)*	0.20	-0.21	*(-1.13)*	-0.28
LtRev	0.40	*(2.38)*	0.37	0.52	*(2.45)*	0.45	0.10	*(0.39)*	0.12
Country risk									
Seas	0.14	*(0.92)*	0.15	0.16	*(0.79)*	0.15	0.09	*(0.63)*	0.16
Size	0.33	*(1.85)*	0.38	0.69	*(2.56)*	0.71	-0.14	*(-0.75)*	-0.21
Iss	0.35	*(2.37)*	0.47	0.30	*(1.20)*	0.35	0.39	*(2.48)*	0.70

This table reports the returns on country equity long–short strategies constructed using design choices commonly adopted in the literature. Portfolios are formed as equal-weighted terciles using developed and emerging markets represented by MSCI indices. The implementation uses the full available sample period, applies no additional implementation lag beyond the signal calculation, and imposes no hyperinflation filters or minimum country requirements. Each month, countries are sorted on the indicated characteristic, and the long–short return is computed as the difference between the highest and lowest portfolios. The table presents results for the full period, the subsample ending in 2009, and the post-2009 period, where 2009 corresponds to the average end of the study period in Table A1 in the Online Appendix. Reported statistics include mean monthly returns (*R*), Newey and West (1987) adjusted *t*-statistics (*t*-stat), and annualized Sharpe ratios (*SR*). The returns are expressed in percentage terms.

decades—including at the country level (Zaremba et al., 2020)—potentially due to increased market integration, improved information dissemination, and stronger arbitrage activity.

To sum up, these findings help reconcile our multiverse results with the existing literature. While the full design space reveals substantial fragility, imposing the specific restrictions used in prior studies often recovers their findings, particularly in earlier sample periods, when these effects were originally documented. This exercise does not validate common choices as preferred benchmarks; it shows that many earlier results are conditional on the designs under which they were obtained.

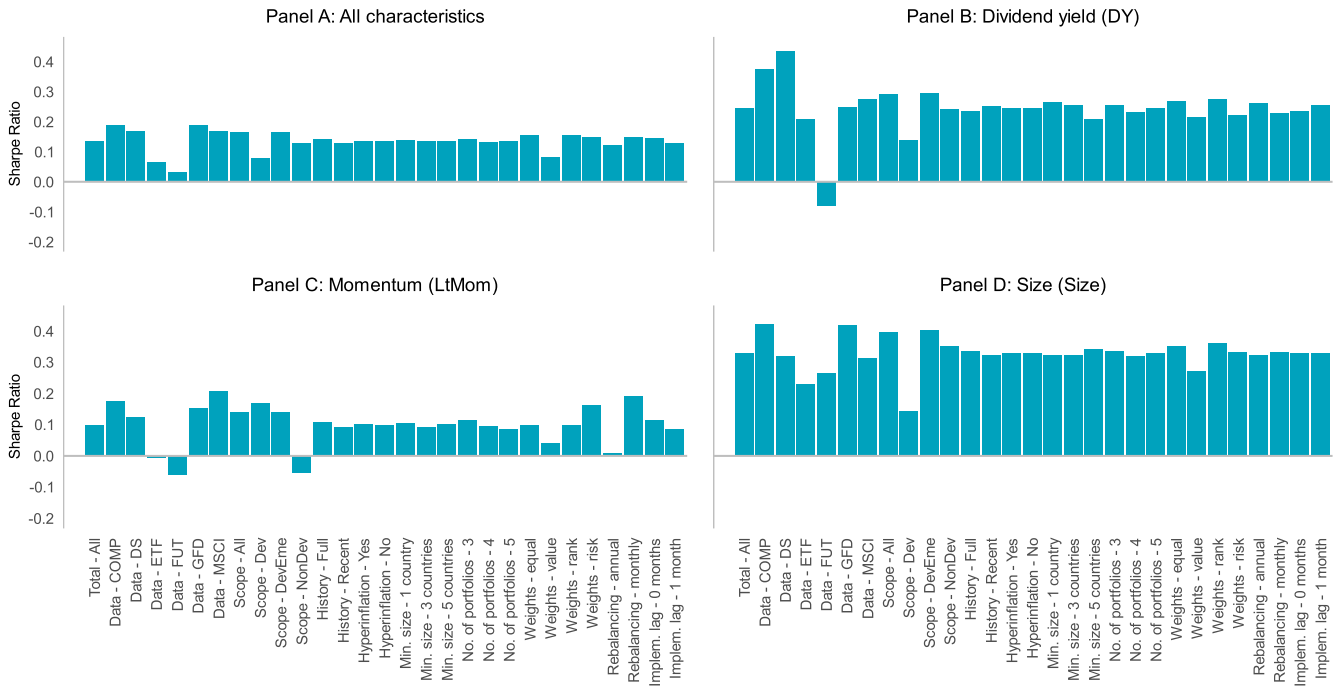

Fig. 6. Average Sharpe ratios across decision nodes.
The figure presents the average annualized Sharpe ratios of the long-short country portfolios sorted on characteristics described in Table 1. The research designs are filtered by the decision nodes indicated on the horizontal axis. Panel A reports the aggregated statistics for the pooled sample of all portfolios based on 15 different characteristics, while Panels B through D zoom in on dividend yield (*DY*), long-term momentum (*LtMom*), and size (*Size*) portfolios, respectively. The study period runs from January 1973 to December 2022.

3.3. Which research design choices matter?

Section 3.1 documents substantial variation in portfolio performance across implementations. In this section, we examine which research design choices drive these differences in returns, Sharpe ratios, and alphas. To do so, we conduct three complementary analyses. First, we analyze average portfolio performance across subsets of the multiverse, holding individual decision nodes fixed. Second, we apply the Anderson–Darling test to assess sensitivity to specific methodological choices. Third, we use regression analysis to quantify the marginal impact of each design choice while controlling for others.

3.3.1. Performance across decision nodes

Our first exercise examines subsets of the 13,824 specifications. At each decision node in Fig. 2, we hold one branch constant while varying all other design choices. This approach helps isolate the effect of individual decisions on portfolio performance and provides insight into how research design shapes outcomes. Fig. 6 presents an initial snapshot, showing the impact of methodological choices on average annualized Sharpe ratios for country portfolios sorted by 15 characteristics. We report both aggregate results (Panel A) and results for selected signals—dividend yield, momentum, and size (Panels B–D).

Several choices emerge as particularly influential. First, the data source is critical: using COMP data substantially improves risk-adjusted returns, whereas ETF and index futures data lag behind. Second, geographic scope matters, with performance differing between samples limited to developed markets and those including emerging and frontier markets. Third, the weighting scheme plays a key role, as value-weighted portfolios exhibit markedly lower performance than equal- or rank-weighted portfolios. In contrast, other decisions—such as the treatment of hyperinflation, minimum country counts, or breakpoints—appear largely inconsequential, with Sharpe ratios remaining nearly unchanged.

Notably, the most influential choices alter the economic exposure of the test. Broader datasets (e.g., COMP), wider geographic scope, and equal-weighting give more weight to small, less liquid, and often less integrated markets. By contrast, value-weighting, developed-market samples, and ETF-covered countries tilt portfolios toward larger and more accessible markets. Thus, the dispersion reflects differences across implementations in the markets being emphasized, not only alternative statistical treatments of the same return patterns.

The impact of methodological decisions also varies across characteristics. Size and dividend yield strategies are strongest in emerging and frontier markets, whereas momentum is driven entirely by developed markets. This helps explain why studies such as Asness et al. (2013) or Baltussen et al. (2021), which focus on a small set of large countries with long histories, report strong results. Rebalancing frequency also matters: for momentum, which depends on timely information, less frequent rebalancing (e.g., annual) is detrimental, while for more persistent signals such as dividend yield and size, it has little effect.

Table 4
Returns on characteristic-sorted portfolios across decision nodes.

	Average	Median	SD	IQR	% Mean > 0	% *t*-stat > 1.96	SR
Data source							
COMP	0.22	0.19	0.28	0.29	85.0	18.4	0.19
DS	0.19	0.19	0.26	0.30	81.6	17.5	0.17
ETF	0.06	0.07	0.24	0.30	60.8	5.5	0.06
FUT	0.08	0.06	0.54	0.42	58.6	0.8	0.03
GFD	0.25	0.22	0.33	0.41	80.6	25.5	0.19
MSCI	0.20	0.20	0.27	0.32	80.6	19.2	0.17
Scope							
All	0.19	0.20	0.26	0.31	79.0	17.5	0.17
Developed	0.07	0.08	0.18	0.20	69.6	6.5	0.08
Dev. & emer.	0.19	0.20	0.25	0.31	78.9	17.6	0.17
Non-developed	0.22	0.24	0.55	0.60	69.8	15.7	0.13
History							
Full	0.18	0.17	0.35	0.35	75.6	16.9	0.14
Recent	0.16	0.15	0.33	0.34	73.6	12.2	0.13
Hyperinflation							
Yes	0.17	0.16	0.34	0.34	74.7	14.5	0.14
No	0.17	0.16	0.34	0.34	74.5	14.5	0.14
Minimum size							
No requirement	0.17	0.16	0.31	0.34	75.6	14.8	0.14
3 countries	0.17	0.16	0.35	0.34	75.0	14.6	0.14
5 countries	0.16	0.16	0.36	0.35	72.9	14.2	0.14
Number of portfolios							
3	0.16	0.15	0.30	0.31	75.5	15.5	0.14
4	0.17	0.16	0.35	0.34	74.2	13.9	0.13
5	0.19	0.18	0.38	0.39	74.0	14.1	0.14
Weighting scheme							
Equal-weighted	0.19	0.18	0.34	0.34	77.8	17.2	0.16
Value-weighted	0.12	0.12	0.33	0.36	66.7	8.7	0.08
Rank-weighted	0.20	0.20	0.40	0.37	77.1	17.1	0.16
Risk-weighted	0.16	0.15	0.28	0.30	76.8	15.1	0.15
Rebalancing							
Annaul	0.15	0.14	0.33	0.33	72.4	11.8	0.12
Monthly	0.19	0.18	0.35	0.35	76.8	17.2	0.15
Implementation lag							
No lag	0.18	0.17	0.34	0.34	75.7	15.5	0.15
1 month	0.16	0.15	0.34	0.34	73.5	13.6	0.13

The table presents the statistical properties of the distributions of mean monthly returns for various portfolio implementations based on sorts on 15 country characteristics detailed in Table 1. The multiverse of 13,824 possible research designs is filtered by the decision nodes listed in the first column. The reported statistics include the mean (*Average*), median (*Median*), standard deviation (*SD*), interquartile range (*IQR*), the proportion of positive mean returns (*% Mean > 0*), the proportion of mean returns with t-statistics exceeding 1.96 (*% t-stat > 1.96*), and the Sharpe ratio (*SR*) for all research designs. All values, except for the SR, are expressed in percentage terms. The study period spans January 1973 to December 2022.

Table 4 provides more formal evidence that the most influential methodological choices are also those that change the test's economic exposure. Mean returns for COMP portfolios reach 0.22%, whereas index futures and ETF strategies generate two to three times lower returns, likely reflecting differences in index coverage. While futures and ETFs focus on large, liquid stocks, CRSP and Compustat cover the full market, including small and micro-cap stocks—the primary source of return predictability (e.g., Hou et al., 2020).

The geographical scope also plays a crucial role. Restricting the sample to developed markets yields near-zero average returns (0.07%) and a Sharpe ratio of 0.08, whereas including emerging markets raises returns to 0.19% and the Sharpe ratio to 0.17. Focusing solely on non-developed markets further increases mean returns to 0.22%. The weighting scheme likewise matters, with substantially lower profits for value-weighted strategies, and lower rebalancing frequency partly reducing gains. Portfolio granularity has a modest effect: mean returns rise from 0.16% for terciles to 0.19% for quintiles, consistent with greater dispersion in extreme portfolios. However, Sharpe ratios remain stable, as finer sorts reduce diversification and increase volatility. Other design choices appear to have limited impact.

3.3.2. Fork sensitivity

In the first exercise, we build on the approach outlined by Menkveld et al. (2024) and use the Anderson-Darling (AD) test to evaluate fork sensitivity across the 13.824 setups. This framework enables a comparison of multiple samples to determine whether they share a common underlying distribution. Specifically, the test quantifies differences between the empirical cumulative distribution functions (ECDFs) of the samples. By applying this method to subsets of our research designs, categorized by specific decisions at each fork, we systematically examine the significance of these decisions for performance outcomes. The AD *k*-sample test, with its transformation to an asymptotic standard normal distribution, provides a framework for directly comparing sensitivity across diverse research designs. The magnitude of the test statistics indicates the relative influence of particular decision nodes on performance measures.

Fig. 7 displays the results of this analysis. To provide a broader perspective, we present the outcomes for three performance parameters: mean returns, Sharpe ratios, and World CAPM alphas. The AD statistics are substantial and significant, indicating the critical role played by most decision nodes. The reported scores correspond to *p*-values close to zero for all design choices except portfolio size and hyperinflation. Nevertheless, the magnitude of the *t*-statistics highlights which decisions are most critical. These are primarily the data source and sample scope—i.e., the decision whether to focus on broader or narrower market sets. These two forks clearly stand out as playing the biggest role. They are subsequently followed by the weighting scheme and rebalancing frequency.

The importance ranking of specific decisions is relatively consistent across all three performance measures. However, the number of portfolios matters comparably more for mean returns and alphas than for Sharpe ratios. This phenomenon may be inherently linked to the specificity of country portfolios. While a broader characteristic spread may lead to higher mean returns and alphas for the short portfolios, its effect on Sharpe ratios is less clear. More extreme breakpoints may result in the top and bottom portfolios including only a handful of markets. Such low diversification substantially increases portfolio volatility, thereby dampening the Sharpe ratio. Consequently, the AD tests for this measure are less definitive than for alphas.

3.3.3. Regression tests

While earlier analyses highlight performance differences across decision nodes, they have limitations. First, although the tests in Section 3.2.2 identify which nodes matter, they do not quantify the impact of specific design choices. Second, the analysis in Section 3.2.1 shows that certain methodological decisions are associated with stronger (or weaker) factor performance, but these choices may overlap, producing similar sample effects. For example, selecting particular databases may effectively reduce exposure to frontier markets. As a result, the marginal impact of individual design choices, as well as their statistical significance, remains unclear.

To shed further light on this issue, we regress the measures of factor performance on dummy variables representing different research design choices:

$$Yi = \beta 0 + j = 1N\beta jDj + \epsilon i, t, \tag{2}$$

where Yi represents a given performance measure of the factor implementation i, Dj are dummy variables (j=1, 2,…, N) representing particular research design choices at different nodes, β0 and βj are estimated regression coefficients and ϵi is the error term.

As in Section 3.2.2, we use three performance measures for robustness: mean returns, Sharpe ratios, and CAPM alphas. Furthermore, we conduct regressions in three different settings: (i) univariate regressions considering one decision at a time, (ii) multiple regressions accounting for all decisions simultaneously, and (iii) multivariate regressions incorporating all decisions that proved significant in the α_1 specification

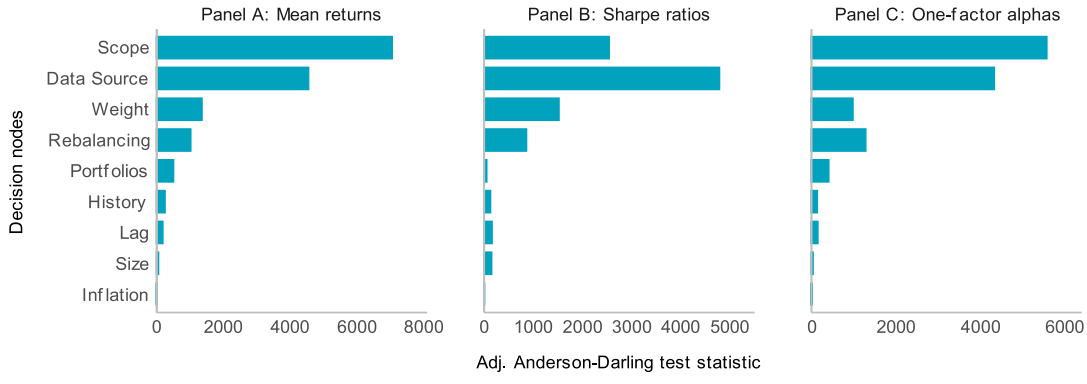

Fig. 7. Anderson-Darling tests of fork sensitivity.
This figure illustrates how sensitive portfolio performance is to the forks in the multiverse analysis associated with different research design choices across up to 13,824 possible implementation variants. The sensitivity is measured with the adjusted standardized Anderson-Darling test statistic. Higher values of the statistic indicate that distributions are more dissimilar across alternative design choices on the fork. We report the results for three different measures of risk-adjusted performance: mean returns (Panel A), Sharpe ratios (Panel B), and the World CAPM alphas (Panel C). The considered decision nodes are indicated on the vertical axis. The statistics are calculated for the pooled sample of all portfolios based on 15 different characteristics. The study period runs from January 1973 to December 2022.

(ii). In specifications with several branches from the same decision node, we omit one branch and interpret the coefficients relative to that reference branch. This normalization avoids perfect multicollinearity. This setup allows us to identify how specific design choices relate to portfolio performance.

Table 5 reports the regression results. Many choices are neither immaterial nor irrelevant, as they significantly affect returns. Since the impact on different performance measures is qualitatively similar, we use mean returns to illustrate which design choices systematically shift performance, emphasizing in particular the composite specification (iii).

Within the first category, data source (Panel A), choosing ETFs or futures as country market representations reduces mean returns by 0.135–0.155 percentage points. Regarding country selection (Panel B), the data reveal a negative impact from limiting the sample to developed markets, with mean returns lower by 0.125 percentage points. On the other hand, including a broader set of markets improves profitability. However, this effect appears only in univariate regressions and vanishes in multivariate tests. Apparently, other key design choices—such as using databases with broader coverage or longer sample periods—already capture the effect of reducing sample size.

Turning to sample length (Panel C), trimming the time series to more recent periods also negatively affects return predictability, though the impact is relatively modest. Mean payoffs decline by 0.025 percentage points. While this may seem at odds with the subperiod results reported in Table 3, the difference likely reflects an offsetting effect. Shorter samples reduce time-series variation, but recent decades also expand coverage of frontier markets. These markets—often less liquid, less efficient, and partially segmented—may act as a reservoir of return predictability, compensating for the loss of historical data.

The effect of portfolio breakpoints (Panel F) depends on the measurement period. In line with our previous findings, mean returns and alphas generally benefit from larger characteristic spreads; their magnitude increases with finer sorts. In practice, mean returns can be up to 0.025 percentage points higher when using quintile sorting rather than terciles or quartiles. However, for Sharpe ratios this effect is not observed. Broader groupings benefit from greater diversification, which helps reduce portfolio volatility.

The role of the weighting scheme (Panel G) is more clear-cut. Methods that emphasize smaller markets within the portfolio tend to enhance performance. Equally weighted and rank-weighted portfolios typically yield higher profits, with incremental effects on mean return amounting to 0.048 and 0.061 percentage points, respectively.

Finally, although decisions regarding hyperinflation treatment, minimum portfolio size, or rebalancing frequency (Panels D, E, H) do not

Table 5
Performance drivers across research designs: a regression analysis.

	Dependent variable: R			Dependent variable: α_1			Dependent variable: SR		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Panel A: Data source									
COMP	0.065**	0.019		0.082**	0.063**		0.065**	0.020	
DS	0.029	-0.009		0.049*	0.037		0.040*	-0.001	
ETF	-0.129**	-0.141**	-0.155**	-0.116**	-0.100**	-0.145**	-0.086**	-0.106**	-0.115**
FUT	-0.104	-0.120*	-0.135**	-0.120**	-0.104*	-0.149**	-0.124**	-0.138**	-0.147**
GFD	0.096**	0.047		0.097**	0.077**		0.062**	0.018	
MSCI	0.040			0.005			0.041**		
Panel B: Scope									
All	0.025**	-0.034		0.020*	-0.040		0.039**	0.039	
Developed	-0.125**	-0.148		-0.117**	-0.144		-0.076*	-0.050	
Dev. & eme.	0.024**	-0.036		0.019*	-0.041		0.040**	0.040	
Non-dev.	0.071			0.074			-0.010		
Panel C: History									
Full	0.024**			0.018**			0.013*		
Recent	-0.024**	-0.024**	-0.024**	-0.018**	-0.018**	-0.018**	-0.013*	-0.013*	-0.013*
Panel D: Hyperinflation									
Yes	0.001	0.001		0.000	0.000		0.001*		
No	-0.001			0.000			-0.001*	0.001*	
Panel E: Minimum size									
No requirement	0.006			0.006			0.002		
3 countries	0.000	-0.004		0.000	-0.004		0.000	-0.002	
5 countries	-0.007	-0.013		-0.006	-0.012	0.023**	-0.002	-0.011	
Panel F: Number of portfolios									
3	-0.019*	-0.009		-0.016**	-0.006		0.009	0.012**	
4	-0.005			-0.006			-0.008**		
5	0.026**	0.015**	0.025**	0.024**	0.015**	0.023**	-0.001	-0.003	-0.002
Panel G: Weighting scheme									
Equal-weighted	0.027**	0.027*	0.048**	0.024**	0.021*	0.042**	0.027**	0.007	0.041**
Value-weighted	-0.064*	-0.042		-0.060*	-0.042		-0.072**	-0.068**	
Rank-weighted	0.045**	0.041*	0.061**	0.040**	0.033*		0.027**	0.007	0.041**
Risk-weighted	-0.009			-0.004			0.018		
Panel H: Rebalancing									
Annaul	-0.046	-0.046		-0.052	-0.052		-0.027	-0.027	
Monthly	0.046			0.052			0.027		
Panel I: Implementation lag									
No lag	0.027*			0.025*			0.017		
One month	-0.027*	-0.027*	-0.027*	-0.025*	-0.025*	-0.025*	-0.017	-0.017	-0.017

The table presents the regression analysis of performance drivers of country premiums across different research designs. The study sample encompasses a pooled set of up to 13,824 implementations of 15 characteristic-sorted portfolios (detailed in Table 1). The dependent variables are the strategy mean monthly return (R), World CAPM alphas (α_1), and Sharpe ratio (SR), as indicated in the top row. The regressors encompass one-hot encoded dummy variables representing the research design choices indicated in the first column. The study period runs from January 1973 to December 2022. Specification (1) presents the slope coefficients from univariate regressions. Specifications (2) and (3) represent multivariate regressions, incorporating variables included in the respective columns. All standard errors are clustered by market characteristic to account for cross-sectional dependence in performance outcomes, and the asterisks * and ** denote significance at the 10% and 5% levels, respectively.

stand out as significant in the regression tests, implementation lags do matter (Panel I). Imposing a one-month skip period reduces mean returns by an average of 0.017 percentage points.

Overall, the most impactful research design choices reveal a clear pattern. Methodological decisions that emphasize the role of hard-to-invest assets or markets—either through sample construction or portfolio selection and formation procedures—tend to boost factor performance. Conversely, reducing their influence and focusing on more liquid and developed markets, or even on recent periods when markets became more liquid and accessible, leads to lower performance. This observation aligns with the argument that the cross-sectional predictability of country equity returns may be inherently tied to the integration (or segmentation) of global capital markets.

Overall, the largest coefficients in Table 5 come from choices that change which markets receive weight. Designs that place more weight on hard-to-invest assets or markets, through sample construction or portfolio formation, are associated with higher average returns. Designs that place more weight on liquid and developed markets, or on more recent periods, are associated with lower returns. This pattern aligns with the view that cross-country predictability varies with the segmentation and integration of global capital markets.

3.4. Nonstandard errors

While earlier sections focused on how research design choices affect performance levels, we now examine their impact on variability, i.e., the extent to which methodological uncertainty drives differences in results. To estimate standard and nonstandard errors, we rely on two complementary approaches from the literature, ensuring both robustness and comparability with recent findings across asset classes.

We first follow Soebhag et al. (2024), who define the standard error (SE) as the cross-sectional average of Sharpe ratio standard deviations estimated via block bootstrapping (SVV method). For each design, we resample its factor return series 1,000 times using block bootstraps, compute the Sharpe ratio for each draw, and then take the standard deviation across these draws. Averaging these standard deviations across all designs yields the SE. The nonstandard error (NSE), in turn, captures methodological uncertainty and is defined as the cross-sectional standard deviation of Sharpe ratios across all designs.

To complement this Sharpe-ratio-based decomposition, we also estimate SE and NSE using the return-based framework of Coqueret and Pérignon (2025) (CP method). This approach decomposes the total variability of average portfolio returns into two orthogonal components: SE, reflecting sampling uncertainty over time, and NSE, capturing design-induced variation across specifications. Both are expressed in return units (percentage points).

We implement the CP method by applying each research design $p = 1, 2,..., P$ to each of the N subsamples D_n, producing a matrix of return estimates $\widehat{b_p}(\mathbb{D}_n)$. These subsamples are generated via block bootstrapping, analogous to those used in the Sharpe-ratio-based method. The standard error in this method is defined as:

$$\mathrm{SE} = \sqrt{\frac{1}{2}\left(\frac{1}{P}\sum_{p=1}^{P}\widehat{\sigma}_p^2 + \frac{1}{N}\sum_{n=1}^{N}\sum\left(\widehat{\mu}_n - \widehat{\mu}\right)^2\right)}, \tag{3}$$

and the nonstandard error is:

$$\mathrm{NSE} = \sqrt{\frac{1}{2}\left(\frac{1}{N}\sum_{n=1}^{N}\widehat{\sigma}_n^2 + \frac{1}{P}\sum_{p=1}^{P}\sum\left(\widehat{\mu}_p - \widehat{\mu}\right)^2\right)}. \tag{4}$$

In these expressions $\widehat{\mu}_p$ is the average return of design p across all subsamples, $\widehat{\mu}_n$ is the average return in subsample n across all designs, $\widehat{\mu}$ denotes the overall average return across all designs and subsamples, $\widehat{\sigma}_p^2$ is the variability of design p across subsamples, and $\widehat{\sigma}_n^2$ is the variability within subsample n across designs. This structure ensures that both sampling error and specification error are fully accounted for in the estimation process.

Finally, as in Menkveld et al. (2024) and Soebhag et al. (2024), we also compute the NSE-to-SE ratio (N/S) as a summary measure of the relative magnitude of design-induced variability compared to sampling uncertainty.

3.4.1. Nonstandard errors across factors

Table 6 presents nonstandard errors (NSEs), standard errors (SEs), and the NSE/SE ratios for portfolios sorted by 15 characteristics and implemented across all research designs outlined in Fig. 2. Regardless of the methodology used—SVV or CP—the SE values remain relatively consistent across specifications. For example, under the SVV approach, they range from 0.186 for *Size* to 0.196 for *Iss*. In contrast, NSE values show greater variation, spanning from 0.142 for *Beta* to 0.247 for *LtRev*.

Importantly, because the SVV and CP methods are based on different economic units (Sharpe ratios and returns, respectively), their NSE values are not directly comparable. Nevertheless, both frameworks lead to similar conclusions: NSEs generally exceed SEs. Specifically, the average NSE/SE ratios equal 119% and 159% for the SVV and CP approaches, respectively. There are also noticeable differences across anomalies. For example, under the SVV method, the NSE/SE ratio ranges from 0.747 (*Beta*) to 1.280 (*LtRev*). While the two methodologies do not always provide consistent signals, some general patterns are shared. For instance, *Beta* consistently displays relatively low methodological uncertainty, whereas *EconRisk, DY*, and *LtRev* exhibit higher values.

The literature on nonstandard errors in other asset classes reports somewhat different ratios, depending on the context. For example, Fieberg et al. (2024) find that the corresponding NSE/SE ratio in cryptocurrency markets is 155%. Soebhag et al. (2024) report ratios ranging from 43% for the size (SMB) factor to 197% for the post-earnings announcement drift (PEAD) factor, with an average of 106%.

Table 6
Nonstandard and standard errors on country portfolios.

	SVV method			CP method		
	SE	NSE	NSE/SE	SE	NSE	NSE/SE
Country risk						
EconRisk	0.188	0.229	1.218	0.198	0.353	1.782
PolRisk	0.189	0.150	0.794	0.258	0.425	1.645
SovRat	0.189	0.161	0.852	0.252	0.424	1.682
IVol	0.187	0.192	1.027	0.239	0.409	1.713
Beta	0.190	0.142	0.747	0.257	0.332	1.293
Momentum						
StMom	0.186	0.204	1.097	0.210	0.363	1.730
LtMom	0.189	0.213	1.127	0.225	0.355	1.581
H52	0.186	0.208	1.118	0.170	0.220	1.299
MA	0.188	0.226	1.202	0.231	0.289	1.253
Value						
DY	0.185	0.229	1.238	0.278	0.455	1.635
EP	0.185	0.182	0.984	0.255	0.424	1.661
LtRev	0.193	0.247	1.280	0.223	0.421	1.893
Other						
Seas	0.187	0.232	1.241	0.225	0.366	1.627
Size	0.186	0.187	1.005	0.288	0.425	1.479
Iss	0.196	0.168	0.857	0.224	0.351	1.564
Full sample						
All	0.188	0.223	1.186	0.235	0.374	1.589

This table reports the standard errors (SE) and nonstandard errors (NSE) associated with long-short portfolios formed from sorts on the 15 country-level predictors described in Table 1. Each portfolio is implemented using the full set of up to 13,824 research designs defined in Fig. 2. The left panel follows the method of Soebhag et al. (2024), which computes errors based on Sharpe ratios (SVV method). The right panel uses the methodology of Coqueret and Pérignon (2025), which computes SE and NSE based directly on portfolio returns and expresses them in percentage terms. For both panels, the final column shows the ratio of nonstandard to standard error (NSE/SE). The study period spans from January 1973 to December 2022.

Menkveld et al. (2021) find an NSE/SE ratio of 160% in an experimental study using index futures, while Walter et al. (2024) report a ratio of 110% based on methodological uncertainty in portfolio sorting. Against this background, the uncertainty observed in country-level sorts appears qualitatively similar to the upper range of these earlier estimates.

3.4.2. Reducing nonstandard errors

Variation in design choices enables researchers to tailor their analyses to specific questions, such as identifying patterns within particular market types. For instance, researchers may narrow their scope to developed markets to focus on practical implications or, alternatively, to emerging markets to explore unique behaviors. While some design flexibility can be beneficial for uncovering patterns, excessive freedom in choices can complicate result interpretation and encourage over-reporting of significant findings.

In this section, we examine whether a limited set of methodological constraints can meaningfully reduce nonstandard errors and improve the robustness of observed return patterns. Streamlining design choices can improve interpretability and reliability while limiting spurious findings. Table 7 reports NSE, SE, and the NSE/SE ratio for the full sample, categorized by decision nodes. Estimates are based on the SVV methodology, which derives errors from Sharpe ratios.

Both SE and NSE vary across methodological choices. Unlike earlier evidence from Table 6, NSEs are not systematically more volatile than SEs; both display similar dispersion.

A modest pattern appears for data sources: NSEs are lower when characteristics are drawn from COMP (0.208) or Datastream (0.201), and higher for ETF (0.228) and GFD (0.224) data. Geographic scope has a stronger effect. Restricting the sample to developed markets reduces both SE and NSE (NSE = 0.196), whereas non-developed markets exhibit higher errors (NSE = 0.252). Portfolio breakpoints and weighting schemes have a limited impact on uncertainty. In contrast, minimum portfolio size matters: portfolios with more countries show higher NSEs. Rebalancing also plays a role, with less frequent rebalancing producing more consistent results across specifications. Most other choices, including sample length and the treatment of hyperinflation, have little effect.

Overall, both SE and NSE are shaped by specific design decisions, with geographic scope, portfolio size, and rebalancing frequency playing the largest roles in stabilizing results. These findings underscore the importance of transparent design choices in country-level research.

3.5. Aggregate significance

Sections 3.1 and 3.2 show substantial variation across individual implementations. We now ask whether each signal remains significant after aggregating information across the full multiverse. This test moves from implementation-level evidence to signal-level inference: does a predictor generate a reliable premium once methodological uncertainty is taken into account?

To address this challenge and evaluate factor robustness across various designs, we run three analyses. To begin with, we develop a bootstrap test to verify whether the average premium across different implementations is positive. Second, we run an out-of-sample examination, checking the performance of designs that performed well in the past. The two approaches help mitigate the risk of data snooping in the results by simultaneously accounting for the methodological diversity in portfolio tests.

3.5.1. Bootstrap tests

In the bootstrap test, rather than relying on a single research design, we assess whether the observed factor is statistically significant and consistent across the entire multiverse of specifications. We begin with data from N design choices, each containing T observations. First, we calculate the mean return for each design choice. We then average these N means to obtain the "mean of the means." Our goal is to determine whether this mean of the means is significantly different from zero.

In our block bootstrap procedure, we treat all design choices together as a single dataset. We draw a random sample of T observations with replacement for each design choice, simultaneously preserving cross-sectional dependencies across design choices. In each bootstrap iteration, we calculate the mean return and World CAPM alpha for each design choice based on the resampled observations and then compute the mean of these performance metrics. We repeat this process B times to generate a distribution of the means of the means and average alphas. Finally, to obtain the p-value, we count the proportion of bootstrap iterations where a given performance measure is less than zero and divide this count by B.

We apply this block bootstrap method to both the full sample of all 13,824 implementations and its subsets funneled over different decision nodes, as in Section 2.2. This facilitates the identification of the specific research decisions that strengthen the results, as well as additional verification of the reliability of the observed results. Finally, we use

Table 7
Nonstandard and standard errors across decision nodes.

	SE	NSE	NSE/SE		SE	NSE	NSE/SE
Data source				*Minimum size*			
COMP	0.177	0.208	1.175	No requirement	0.177	0.209	1.181
DS	0.168	0.201	1.196	3 countries	0.187	0.219	1.171
ETF	0.223	0.228	1.022	5 countries	0.203	0.243	1.197
FUT	0.213	0.206	0.967	*Number of portfolios*			
GFD	0.168	0.224	1.333	3	0.183	0.218	1.191
MSCI	0.181	0.214	1.182	4	0.192	0.227	1.182
Scope		5	0.190	0.223	1.180		
All	0.179	0.214	1.196	*Weighting scheme*			
Developed	0.185	0.196	1.059	Equal-weighted	0.186	0.222	1.194
Dev. & eme.	0.179	0.215	1.201	Value-weighted	0.192	0.219	1.141
Non-dev.	0.211	0.252	1.194	Rank-weighted	0.186	0.223	1.199
History		Risk-weighted	0.189	0.219	1.160		
Full	0.184	0.226	1.228	*Rebalancing*			
Recent	0.193	0.220	1.140	Annaul	0.189	0.210	1.111
Hyperinflation		Monthly	0.187	0.234	1.251		
Yes	0.188	0.223	1.186	*Implementation lag*			
No	0.188	0.223	1.186	No lag	0.188	0.219	1.165
				1 month	0.188	0.227	1.207

The table reports the nonstandard (*NSE*) and standard (*SE*) errors for the pooled set of long-short portfolios from sorts on the 15 characteristics described in Table 1 and implemented using all research designs presented in Fig. 2. The *SE* and *NSE* calculation follows the method of Soebhag et al. (2024), which computes errors based on Sharpe ratios. The tests are conducted on the subsets of all possible up to 13,824 research designs filtered by specific decision nodes indicated in the first column. The last column reports the ratio of nonstandard to standard errors (*NSE/SE*). The study period runs from January 1973 to December 2022.

B=1000 bootstraps to calculate the p-value.

Fig. 8 presents the results of our analysis, with Panel A focusing on mean returns. First, observe the left-most column, which shows the findings for the complete multiverse. Only six country characteristics emerge as significant at the 5% level: *EconRisk, SovRat, DY*, yield, *EP, Size*, and *Iss*. Among these, *Size* turns out to be the strongest predictor, the only one to pass a 1% significance threshold. Several subsequent ones are significant at the 10% level, including *PolRisk, IVol, StMom*), and *LtRev. Beta, Seas*, and different momentum and technical analysis measures play no major role. Although frequently examined in the literature, their effects appear to hold only under specific conditions and fail to generalize to alternative implementations.

The subsequent columns focus on aggregate significance within subsets of the full sample. This again underscores the sensitivity of cross-country return patterns to specific choices. Most variables identified as significant at the 5% level prove largely robust; however, performance weakens in implementations with ETFs, index futures, and developed markets. Clearly, focusing on the most liquid and largest markets weakens return predictability.

Panel B of Fig. 8 focuses on World CAPM alphas. After accounting for the market risk exposure, only five characteristics remain robust and significant at the 5% level. These include *SovRisk, StMom, DY*, and *Size*, and *Iss*. Their abnormal returns remain also robust across multiple specifications, yet, some of them—such as data source choice or sample scope—still undermine their robustness.

To sum up, the main conclusion is that cross-country return predictability is fragile. Only a few predictors survive the broad multiverse, and even these signals remain sensitive to specific design choices.

3.5.2. *Out-of-sample design selection*

As an alternative approach, we implement a simple out-of-sample design selection procedure to assess whether past performance can identify genuinely profitable implementations. For each of the 15 country-level signals, we select the top-performing implementations each month based on average returns over a historical ranking window and track their returns in the subsequent month. For robustness, we use two ranking schemes: (i) a rolling 60-month window and (ii) an expanding window (cumulative to date). We also form portfolios from the top 20, 50, or 100 implementations. Statistical significance is evaluated using Newey and West (1987) adjusted standard errors.

Table 8 reports average returns for the selected anomaly sets. The results broadly confirm earlier findings. Several strategies consistently deliver significant and economically meaningful out-of-sample returns, particularly *SovRat, DY, StMom*, and *Size*, with robustness across ranking windows and selection sizes. For example, *SovRat* generates monthly returns of 0.85%–1.30% in Panel B, all significant at the 1% level. The Size anomaly also performs strongly, with returns exceeding 0.9% in most specifications.

In contrast, strategies such as *PolRisk* and *Seas* show little to no out-of-sample predictability and occasionally yield negative returns. *LtRev* exhibits moderate profitability, especially under the rolling-window approach.

Overall, the design selection test aligns with previous results: a subset of country-level factors shows persistent out-of-sample performance, while others are more sensitive to design choices or sample-specific effects. Only a few signals retain genuine predictive power once design uncertainty is partially mitigated through a simple performance-based selection procedure.

4. Conclusion and practical recommendations

This paper examines whether research design choices play a significant role in studies of cross-sectional variation in country equity returns. Specifically, we focus on the impact of common methodological decisions in the studies of country stock portfolios. Our analysis covers up to 13,824 unique specifications based on 15 return-predictive characteristics and nine key implementation choices related to data sources, sample preparation procedures, and portfolio implementations.

Our findings unveil the critical role of several common implementation decisions on the final conclusions. To begin with, the performance of characteristic-sorted country portfolios varies substantially across research designs. Some strategies, such as momentum and dividend yield, show Sharpe ratios ranging from profoundly negative to strongly positive depending on the implementation, raising fundamental questions about the stability and validity of these return patterns. Moreover, we formally assess robustness using three complementary tests. The bootstrap-based "mean of means" test and an out-of-sample design-selection approach, both point to the limited reliability of most cross-country strategies. Only a small subset of characteristics—sovereign rating, short-term momentum, dividend yield, and market size—demonstrate consistent performance across multiple specifications, with market size standing out as the most robust. In addition to the above, among the various research decisions, the data source proves most influential. Return premia prove notably stronger for portfolios constructed from Compustat or GFD indices than for those based on more investable assets, such as ETFs or futures. Other crucial choices include geographic coverage and the portfolio weighting scheme, with stronger outcomes generally linked to greater exposure to small, segmented markets.

Finally, we analyze the role of nonstandard errors (NSEs), which capture the sensitivity of results to methodological uncertainty. On average, NSEs exceed standard errors (SEs) by 19% to 59%, and this level of uncertainty is comparable to other asset classes. The magnitude of NSEs varies across signals and research choices, being relatively high for characteristics such as long-term reversal and dividend yield, but lower for factors such as beta. Certain methodological decisions may help reduce NSEs. In particular, drawing on Compustat or Datastream data, focusing on developed markets, and reducing rebalancing frequency tend to stabilize results. Conversely, designs that emphasize emerging markets or require broader country coverage in portfolios tend to increase variability. However, the overall differences in NSEs across the decision nodes are moderate.

Our findings point to two practical recommendations for future studies on the cross-section of country equity returns. First, researchers should prioritize transparency and robustness by explicitly assessing the sensitivity of their results to alternative research design choices. While conducting a full multiverse analysis may be infeasible in many settings—particularly when access to multiple data sources is limited—we encourage the use of more parsimonious approaches to quantify nonstandard errors. For example, researchers can report the distribution of key performance metrics, such as mean returns, Sharpe ratios, or t-statistics, across a set of plausible specifications. Compact tools such as specification curve analysis (Simonsohn et al., 2020) offer a particularly useful way to summarize this variation. Even partial analyses of this kind can provide valuable insight into the robustness—or fragility—of empirical findings (Steegen et al., 2016).

Second, the baseline empirical designs should combine economic sensibility with common practice, which helps ensure comparability with existing literature. A natural starting point should rely on well-established index providers (e.g., MSCI), relatively long sample periods, and standard portfolio sorts (e.g., three to five portfolios in the cross-section). In contrast to stock level studies, value-weighted portfolios need not serve as the default approach in this setting, as they may overweight a small number of large markets and obscure cross-sectional patterns. Equal-weighting remains useful for comparability with prior studies, but it should not serve as an automatic benchmark for economic relevance. Because it gives greater influence to smaller and less liquid markets, researchers should pair it with value-weighted and investable-asset specifications.

Third, our results shed light on which methodological dimensions matter most for empirical outcomes and thus should be included in robustness checks. Country coverage (e.g., developed vs. emerging vs.

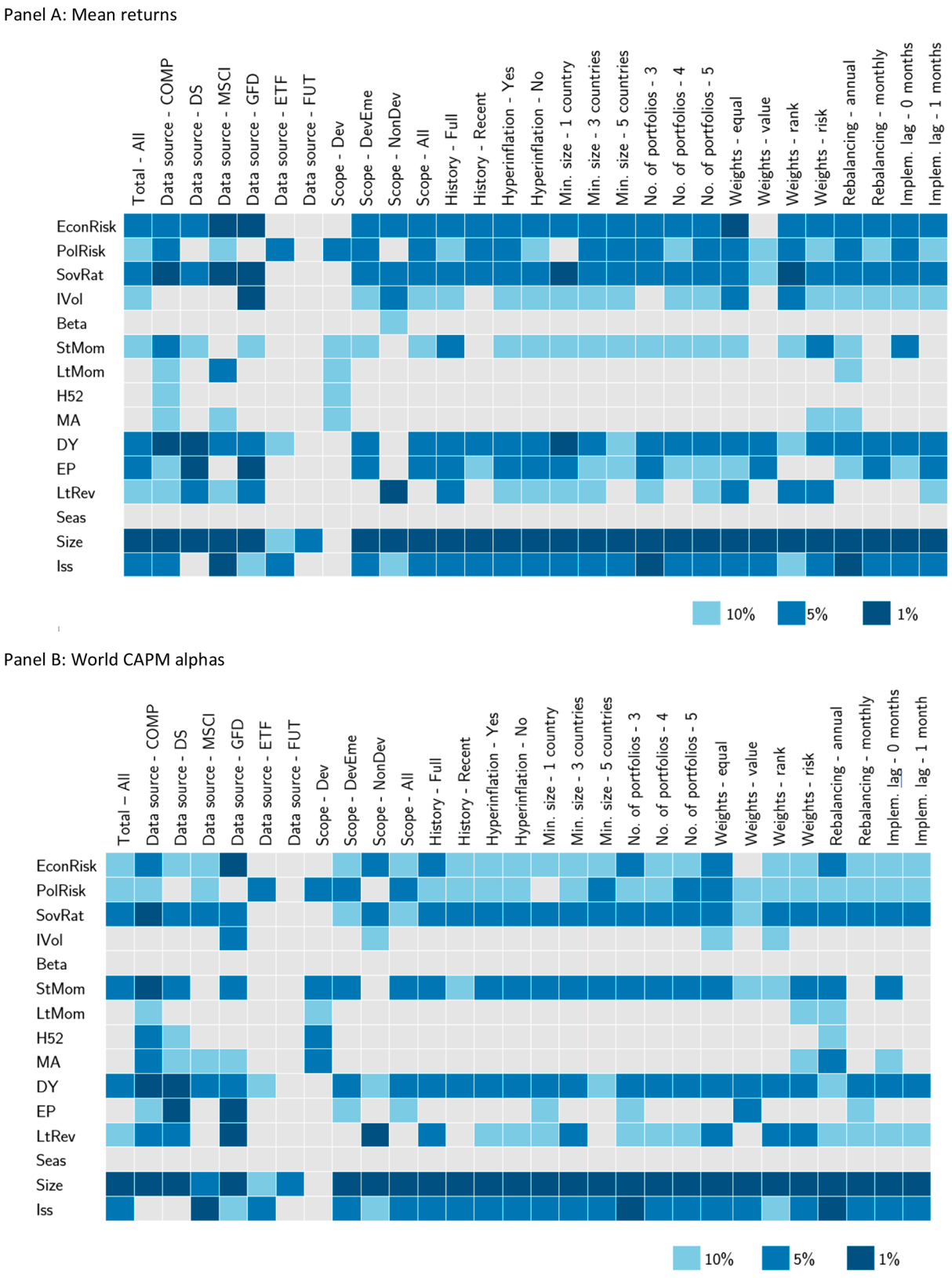

Fig. 8. Statistical significance of abnormal returns across design choices.
The figure illustrates the statistical significance of mean returns (Panel A) and the World CAPM alphas on 15 long-short portfolios based on the sorts of country characteristics described in Table 1. The *p*-values used to test whether the average returns across various specifications differ from zero are calculated using block bootstrap. These tests are conducted on the complete set of up to 13,824 research designs included in the study, as well as on subsets created by filtering by specific decision nodes. The study period spans from January 1973 to December 2022. Different shades of blue denote significance levels of 10%, 5%, and 1%, with darker shades representing stronger significance. The grey color indicates a lack of statistical significance.

Table 8
Out-of-sample design selection based on past performance.

	Panel A: 60-month sorting period			Panel B: Extending sorting period		
	N = 20	N = 50	N = 100	N = 20	N = 50	N = 100
EconRisk	0.70	0.78*	0.77**	0.69	0.61*	0.62**
PolRisk	-0.28	-0.01	-0.01	-0.26	0.04	0.01
SovRat	1.30***	1.29***	1.04***	1.05***	1.01***	0.85***
IVol	0.56	0.60	0.71*	0.55	0.38	0.63*
Beta	0.20	0.25	0.35	0.16	0.37	0.26
StMom	0.67*	0.59*	0.76***	0.82**	0.91***	0.97***
LtMom	0.09	-0.16	-0.13	0.31	0.24	-0.05
H52	-0.13	-0.04	-0.01	0.44	0.13	0.20
MA	0.18	0.24	0.18	0.45	0.21	0.31
DY	0.77**	0.55**	0.46**	1.00***	0.70***	0.55***
EP	0.16	0.10	0.09	0.08	0.05	0.15
LtRev	0.81	0.75*	0.71*	0.99**	0.59*	0.49**
Seas	-1.09	-0.59	-0.26	-1.14	-0.65	-0.33
Size	1.19***	0.92***	0.88***	1.07**	0.96***	0.62**
Iss	0.54	0.59	0.59	0.42	0.44	0.47

This table reports mean monthly returns (%) on equally weighted long–short portfolios formed from the best-performing implementations of each of the 15 country-level signals. Each month, the top N implementations are selected based on their average return over the sorting period and their average return is evaluated out-of-sample in the following month. The ranking period may be either rolling 60-month-long (Panel A) or extending (Panel B). *N* indicates the number of implementations included each month. The asterisks *, **, and *** denote the values significantly different from zero based on Newey and West (1987) adjusted standard errors at the 10%, 5%, and 1%, level, respectively.

frontier markets), the length of the sample period (particularly the inclusion of recent decades), the number of portfolios, and the choice of weighting scheme are among the most influential design decisions. While the choice of data source is also critical, it may not always be feasible to vary in practice. Nevertheless, researchers should, where possible, scrutinize implementations based on tradable instruments, such as ETFs or index futures, which are both economically meaningful and, as our results show, can materially affect the magnitude of return predictability. On the other hand, considering certain filters, such as excluding hyperinflationary episodes or imposing minimum country counts, may be of secondary importance, as these aspects appear to play a limited role in most applications.

CRediT authorship contribution statement

Nusret Cakici: Writing – review & editing, Methodology, Conceptualization. **Christian Fieberg:** Writing – review & editing, Resources, Methodology, Data curation, Conceptualization. **Gabor Neszveda:** Writing – review & editing, Resources, Methodology, Data curation, Conceptualization. **Vanja Piljak:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Adam Zaremba:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Acknowledgment

We thank Geert Bekaert, Campbell R. Harvey, Matthias Hanauer, Ivan Indriawan, Vaibhav Lalwani, Albert J. Menkveld, Amar Soebhag, Laurens Swinkels, and Pim van Vliet for valuable comments and suggestions, as well as participants of the 2025 FMA European Conference, GLOBAFA 2025 Conference, CINSC-INFINITI-2025, and the research seminar at Zhejiang University of Finance and Economics. The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Magyar Nemzeti Bank. All errors are our own. Adam Zaremba acknowledges the support of the National Science Center of Poland [Grant No. 2022/45/B/HS4/00451]. We have no conflicts of interest to disclose.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jbankfin.2026.107775.

Data availability

The authors do not have permission to share data.

References

Albuquerque, R., Eichenbaum, M., Papanikolaou, D., Rebelo, S., 2015. Long-run bulls and bears. J. Monet. Econ. 76, 21–36.
Andreu, L., Swinkels, L., Tjong-A-Tjoe, L., 2013. Can exchange-traded funds be used to exploit industry and country momentum? Financ. Mark. Portf. Manag. 27, 127–148.
Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. J. Finance 61 (1), 259–299.
Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2009. High idiosyncratic volatility and low returns: international and further US evidence. J. Financ. Econ. 91 (1), 1–23.
Angelidis, T., Tessaromatis, N., 2017. Global equity country allocation: an application of factor investing. Financ. Anal. J. 73 (4), 55–73.
Asness, C.S., Liew, J.M., Stevens, R.L., 1997. Parallels between the cross-sectional predictability of stock and country returns. J. Portf. Manag. 23 (3), 79–87.
Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. J. Finance 68 (3), 929–985.
Atilgan, Y., Bali, T.G., Demirtas, K.O., Gunaydin, A.D., 2019. Global downside risk and equity returns. J. Int. Money Finance 98, 102065.
Atilgan, Y., Demirtas, K.O., Gunaydin, A.D., Tosun, A.D., Zirek, D., 2025. Aggregate earnings and global equity returns. J. Int. Financ. Mark. Inst. Money 100, 102125.
Avramov, D., Chordia, T., Jostova, G., Philipov, A., 2012. The world price of credit risk. Rev. Asset Pricing Stud. 2 (2), 112–152.
Baker, M., Wurgler, J., 2000. The equity share in new issues and aggregate stock returns. J. Finance 55 (5), 2219–2257.
Bali, T.G., Cakici, N., 2008. Idiosyncratic volatility and the cross section of expected returns. J. Financ. Quant. Anal. 43 (1), 29–58.
Bali, T.G., Cakici, N., 2010. World market risk, country-specific risk and expected returns in international stock markets. J. Bank. Finance 34 (6), 1152–1165.
Bali, T.G., Cakici, N., Yan, X., Zhang, Z., 2005. Does idiosyncratic risk really matter? J. Finance 60 (2), 905–929.
Baltussen, G., Swinkels, L., Van Vliet, P., 2021. Global factor premiums. J. Financ. Econ. 142 (3), 1128–1154.
Baltussen, G., van Bekkum, S., Da, Z., 2019. Indexing and stock market serial dependence around the world. J. Financ. Econ. 132 (1), 26–48.
Balvers, R.J., Wu, Y., 2006. Momentum and mean reversion across national equity markets. J. Empir. Finance 13 (1), 24–48.
Balvers, R., Wu, Y., Gilliland, E., 2000. Mean reversion across national stock markets and parametric contrarian investment strategies. J. Finance 55 (2), 745–772.
Banz, R.W., 1981. The relationship between return and market value of common stocks. J. Financ. Econ. 9 (1), 3–18.
Basu, S., 1983. The relationship between earnings yield, market value and return for NYSE common stocks: further evidence. J. Financ. Econ. 12 (1), 129–156.
Bekaert, G., Harvey, C.R., 1995. Time-varying world market integration. J. Finance 50 (2), 403–444.
Bekaert, G., Mehl, A., 2019. On the global financial market integration, "swoosh" and the trilemma. J. Int. Money Finance 94, 227–245.
Bekaert, G., Erb, C.B., Harvey, C.R., Viskanta, T.E., 1996. In: Carman, P. (Ed.), The Cross-Sectional Determinants of Emerging Equity Market Returns. Quantitative Investing for Global Markets, New York.
Berkman, H., Yang, W., 2019. Country-level analyst recommendations and international stock market returns. J. Bank. Finance 103, 1–17.
Bauckloh, T., Beyer, V., 2026. Non-standard errors in carbon premia. J. Bank. Finance, 107727.
Bhojraj, S., Swaminathan, B., 2006. Macromomentum: returns predictability in international equity indices. J. Bus. 79 (1), 429–451.
Białkowski, J., Etebari, A., Wisniewski, T.P., 2012. Fast profits: investor sentiment and stock returns during Ramadan. J. Bank. Finance 36 (3), 835–845.
Blitz, D., Van Vliet, P., 2008. Global tactical cross-asset allocation: applying value and momentum across asset classes. J. Portf. Manag. 23–28.
Boudoukh, J., Michaely, R., Richardson, M., Roberts, M.R., 2007. On the importance of measuring payout yield: implications for empirical asset pricing. J. Finance 62 (2), 877–915.
Breloer, B., Scholz, H., Wilkens, M., 2014. Performance of international and global equity mutual funds: do country momentum and sector momentum matter? J. Bank. Finance 43, 58–77.
Cagan, P., 1956. The monetary dynamics of hyperinflation. In: Friedman, M. (Ed.), Studies in the Quantity Theory of Money. University of Chicago Press, Chicago, IL, pp. 25–117, 1956.
Cakici, N., Zaremba, A., 2023. Misery on Main Street, victory on Wall Street: economic discomfort and the cross-section of global stock returns. J. Bank. Finance 149, 106760.
Cakici, N., Zaremba, A., 2024. What drives stock returns across countries? Insights from machine learning models. Int. Rev. Financ. Anal. 96, 103569.

Cakici, N., Fieberg, C., Metko, D., Zaremba, A., 2024. Do anomalies really predict market returns? New data and new evidence. Rev. Finance 28 (1), 1–44.
Calice, G., Lin, M.T., 2021. Exploring risk premium factors for country equity returns. J. Empir. Finance 63, 294–322.
Chan, K., Hameed, A., Tong, W., 2000. Profitability of momentum strategies in the international equity markets. J. Financ. Quant. Anal. 35 (2), 153–172.
Chen, C.-D., Demirer, R., 2022. Oil beta uncertainty and global stock returns. Energy Econ. 112, 106150.
Chen, M., Hanauer, M.X., Kalsbach, T., 2024. Design Choices, Machine Learning, and the Cross-Section of Stock Returns. SSRN. Available at https://ssrn.com/abstract=5031755.
Chiah, M., Long, H., Zaremba, A., Umar, Z., 2023. Trade competitiveness and the aggregate returns in global stock markets. J. Econ. Dyn. Control 148, 104618.
Cirulli, A., De Nard, G., Traut, J., Walker, P., 2025. Low Risk, High Variability: Practical Guide for Portfolio Construction. SSRN. Available at https://ssrn.com/abstract=5105457.
Clare, A., Seaton, J., Smith, P.N., Thomas, S., 2016. The trend is our friend: risk parity, momentum, and trend following in global asset allocation. J. Behav. Exp. Finance 9, 63–80.
Coqueret, G. (2023). Forking paths in financial economics. arXiv preprint arXiv: 2401.08606.
Coqueret, G., Pérignon, C., 2025. Anomaly Persistence and Nonstandard Errors. SSRN, 5276723. Available at.
Daniel, K., Moskowitz, T.J., 2016. Momentum crashes. J. Financ. Econ. 122 (2), 221–247.
De Bondt, W.F., Thaler, R., 1985. Does the stock market overreact? J. Finance 40 (3), 793–805.
Desrosiers, S., L'Her, J.F., Plante, J.F., 2004. Style management in equity country allocation. Financ. Anal. J. 60 (6), 40–54.
Diamonte, R.L., Liew, J.M., Stevens, R.L., 1996. Political risk in emerging and developed markets. Financ. Anal. J. 52 (3), 71–76.
Dimic, N., Orlov, V., Piljak, V., 2015. The political risk factor in emerging, frontier, and developed stock markets. Finance Res. Lett. 15, 239–245.
Dimson, E., Marsh, P., Staunton, M., 2002. Triumph of the Optimists: 101 Years of Global Investment Returns. Princeton University Press.
Du, D., 2008. The 52-week high and momentum investing in international stock indices. Q. Rev. Econ. Finance 48 (1), 61–77.
Durham, J.B., 2001. Sensitivity analyses of anomalies in developed stock markets. J. Bank. Finance 25 (8), 1503–1541.
Edmans, A., Fernandez-Perez, A., Garel, A., Indriawan, I., 2022. Music sentiment and stock returns around the world. J. Financ. Econ. 145 (2), 234–254.
Erb, C.B., Harvey, C.R., Viskanta, T.E., 1995. Country risk and global equity selection. J. Portf. Manag. 21 (2), 74–83.
Erb, C.B., Harvey, C.R., Viskanta, T.E., 1996. Political risk, economic risk, and financial risk. Financ. Anal. J. 52 (6), 29–46.
Ferson, W.E., Harvey, C.R., 1997. Fundamental determinants of national equity market returns: a perspective on conditional asset pricing. J. Bank. Finance 21 (11-12), 1625–1665.
Fieberg, C., Günther, S., Poddig, T., Zaremba, A., 2024. Nonstandard errors in the cryptocurrency world. Int. Rev. Financ. Anal. 92, 103106.
Fieberg, C., Liedtke, G., Poddig, T., Walker, T., Zaremba, A., 2025a. A trend factor for the cross section of cryptocurrency returns. J. Financ. Quant. Anal. 60 (7), 3116–3153.
Fieberg, C., Liedtke, G., Zaremba, A., Cakici, N., 2025b. A factor model for the cross-section of country equity risk premia. J. Bank. Finance 171, 107373.
Fisher, G.S., Shah, R., Titman, S., 2017. Should you tilt your equity portfolio to smaller countries? J. Portf. Manag. 44 (1), 127–141.
Frazzini, A., Pedersen, L.H., 2014. Betting against beta. J. Financ. Econ. 111 (1), 1–25.
Fu, F., 2009. Idiosyncratic risk and the cross-section of expected stock returns. J. Financ. Econ. 91 (1), 24–37.
Gala, V.D., Pagliardi, G., Zenios, S.A., 2023. Global political risk and international stock returns. J. Empir. Finance 72, 78–102.
George, T.J., Hwang, C.Y., 2004. The 52-week high and momentum investing. J. Finance 59 (5), 2145–2176.
Goyal, A., Welch, I., Zafirov, A., 2024. A comprehensive 2022 look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 37 (11), 3490–3557.
Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average US monthly stock returns. Rev. Financ. Stud. 30 (12), 4389–4436.
Hallerbach, W.G., 2014. Disentangling rebalancing return. J. Asset Manag. 15 (5), 301–316.
Harvey, C.R., 1991. The world price of covariance risk. J. Finance 46 (1), 111–157.
Harvey, C.R., 1995. Predictable risk and returns in emerging markets. Rev. Financ. Stud. 8 (3), 773–816.
Harvey, C.R., 2001. Asset pricing in emerging markets. Int. Encycl. Soc. Behav. Sci. 840–845.
Harvey, C.R., 2017. Presidential address: the scientific outlook in financial economics. J. Finance 72 (4), 1399–1440.
Harvey, C.R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. Rev. Financ. Stud. 29 (1), 5–68.
Heston, S.L., Sadka, R., 2008. Seasonality in the cross-section of stock returns. J. Financ. Econ. 87 (2), 418–445.
Hjalmarsson, E., 2010. Predicting global stock returns. J. Financ. Quant. Anal. 45 (1), 49–80.
Hollstein, F., 2022. The world of anomalies: Smaller than we think? J. Int. Money Finance 129, 102741.
Hou, K., Karolyi, G.A., Kho, B.C., 2011. What factors drive global stock returns? Rev. Financ. Stud. 24 (8), 2527–2574.
Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Rev. Financ. Stud. 33 (5), 2019–2133.
Hurst, B., Ooi, Y.H., Pedersen, L.H., 2017. A century of evidence on trend-following investing. J. Portf. Manag. 44 (1), 15–29.
Ilmanen, A., Israel, R., Lee, R., Moskowitz, T.J., Thapar, A., 2021. How do factor premia vary over time? A century of evidence. J. Invest. Manag. 19 (4), 15–57.
Jacobs, H., Müller, S., 2020. Anomalies across the globe: once public, no longer existent? J. Financ. Econ. 135 (1), 213–230.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. J. Finance 48 (1), 65–91.
Jensen, T.I., Kelly, B., Pedersen, L.H., 2023. Is there a replication crisis in finance? J. Finance 78 (5), 2465–2518.
Kamiński, B., Jakubczyk, M., Szufel, P., 2018. A framework for sensitivity analysis of decision trees. Cent. Eur. J. Oper. Res. 26, 135–159.
Keloharju, M., Linnainmaa, J.T., Nyberg, P., 2016. Return seasonalities. J. Finance 71 (4), 1557–1590.
Keloharju, M., Linnainmaa, J.T., Nyberg, P., 2021. Are return seasonalities due to risk or mispricing? J. Financ. Econ. 139 (1), 138–161.
Keppler, A.M., 1991. The importance of dividend yields in country selection. J. Portf. Manag. 17 (2), 24–29.
Keppler, A.M., Traub, H.D., 1993. The small-country effect. J. Invest. 2 (3), 17–24.
Kiviaho, J., Nikkinen, J., Piljak, V., Rothovius, T., 2014. The co-movement dynamics of European frontier stock markets. Eur. Financ. Manag. 20, 574–595.
Kortas, M., L'Her, J.F., Roberge, M., 2005. Country selection of emerging equity markets: benefits from country attribute diversification. Emerg. Mark. Rev. 6 (1), 1–19.
Lalwani, V., Meshram, V., Jindal, V., 2025. Empirical asset pricing via machine learning: the role of research design choices. Eur. Financ. Manag.
Lawrenz, J., Zorn, J., 2017. Predicting international stock returns with conditional price-to-fundamental ratios. J. Empir. Finance 43, 159–184.
Lee, K.H., 2011. The world price of liquidity risk. J. Financ. Econ. 99 (1), 136–161.
Lehkonen, H., Heimonen, K., 2015. Democracy, political risks, and stock market performance. J. Int. Money Finance 59, 77–99.
Li, T., Pritamani, M., 2015. Country size and country momentum affect emerging and frontier markets. J. Invest. 24 (1), 102–108.
Long, H., Chiah, M., Zaremba, A., Umar, Z., 2024. Changes in shares outstanding and country stock returns around the world. J. Int. Financ. Mark. Inst. Money 90, 101883.
Malin, M., Bornholt, G., 2010. Predictability of future index returns based on the 52-week high strategy. Q. Rev. Econ. Finance 50 (4), 501–508.
Menkveld, A.J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Khomyn, M.K., 2024. Nonstandard errors. J. Finance 79 (3), 2339–2390.
Mercik, A.R., 2023. Is tail risk priced in the cross-section of international stock index returns? Mod. Finance 1 (1), 17–29.
Mitton, T., 2022. Methodological variation in empirical corporate finance. Rev. Financ. Stud. 35 (2), 527–575.
Moskowitz, T.J., Ooi, Y.H., Pedersen, L.H., 2012. Time series momentum. J. Financ. Econ. 104 (2), 228–250.
Newey, W.K., West, K.D., 1987. A simple, positive, semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–708.
Novy-Marx, R., Velikov, M., 2022. Betting against betting against beta. J. Financ. Econ. 143 (1), 80–106.
Pitkäjärvi, A., Suominen, M., Vaittinen, L., 2020. Cross-asset signals and time series momentum. J. Financ. Econ. 136 (1), 63–85.
Pontiff, J., Woodgate, A., 2008. Share issuance and cross-sectional returns. J. Finance 63 (2), 921–945.
Richards, A.J., 1997. Winner-loser reversals in national stock market indices: can they be explained? J. Finance 52 (5), 2129–2144.
Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. J. Portf. Manag. 11 (3), 9–16.
Simonsohn, U., Simmons, J.P., Nelson, L.D., 2020. Specification curve analysis. Nat. Hum. Behav. 4 (11), 1208–1214.
Smith, D.M., Pantilei, V.S., 2015. Do 'Dogs of the World' bark or bite? Evidence from single-country ETFs. J. Invest. 24 (1), 7–15.
Soebhag, A., Van Vliet, B., Verwijmeren, P., 2024. Nonstandard errors in asset pricing: mind your sorts. J. Empir. Finance, 101517.
Song, J., Balvers, R.J., 2022. Seasonality and momentum across national equity markets. N. Am. J. Econ. Finance 61, 101706.
Spierdijk, L., Bikker, J.A., Van den Hoek, P., 2012. Mean reversion in international stock markets: an empirical analysis of the 20th century. J. Int. Money Finance 31 (2), 228–249.
Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. 11 (5), 702–712.
Swade, A., Nolte, S., Shackleton, M., Lohre, H., 2023. Why do equally weighted portfolios beat value-weighted ones? J. Portf. Manag. 49 (5).
Umar, Z., Zaremba, A., Umutlu, M., Mercik, A., 2024. Interaction effects in the cross-section of country and industry returns. J. Bank. Finance 165, 107200.
Umutlu, M., 2015. Idiosyncratic volatility and expected returns at the global level. Financ. Anal. J. 71 (6), 58–71.
Walter, D., Weber, R., Weiss, P., 2024. Methodological Uncertainty in Portfolio Sorts. SSRN, 4164117. Available at.
Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.
Willenbrock, S., 2011. Diversification return, portfolio rebalancing, and the commodity return puzzle. Financ. Anal. J. 67 (4), 42–49.

Zaremba, A., Szczygielski, J.J., 2019. And the winner is… a comparison of valuation measures for equity country allocation. J. Portf. Manag. 45 (5), 84–98.
Zaremba, A., Cakici, N., Demir, E., Long, H., 2022. When bad news is good news: geopolitical risk and the cross-section of emerging market stock returns. J. Financ. Stab. 58, 100964.
Zaremba, A., Long, H., Karathanasopoulos, A., 2019. Short-term momentum (almost) everywhere. J. Int. Financ. Mark. Inst. Money 63, 101140.
Zaremba, A., Umutlu, M., Maydybura, A., 2020. Where have the profits gone? Market efficiency and the disappearing equity anomalies in country and industry returns. J. Bank. Finance 121, 105966.
Zhang, C.Y., Jacobsen, B., 2013. Are monthly seasonals real? A three-century perspective. Rev. Finance 17 (5), 1743–1785.
Zhang, C.Y., Jacobsen, B., 2021. The Halloween indicator, "Sell in May and Go Away," is everywhere and all the time. J. Int. Money Finance 110, 102268.