Internet Appendix

War Discourse and the Cross Section of Expected Stock Returns

Internet Appendix **2**

A Method, Estimation, and Data **2**

A.1 Seeded Latent Dirichlet Distribution 2

A.2 Estimation 5

B Additional Results with Our Test Portfolios **9**

C Protocol for Factor Identification **14**

C.1 First Criterion: Correlation of factors with the Systematic Risk of Returns . 14

C.2 Second Criterion: Risk Premium Estimation using WMP 18

D Details on Portfolio Construction **21**

D.1 128 Own Constructed Anomalies 21

D.2 2190 Own Constructed Nonlinear Portfolios 26

D.3 360 ML-based Nonlinear Portfolios 30

A Method, Estimation, and Data

This appendix discusses the semisupervised topic model and our estimation scheme in more detail. Table A.1 reports the full list of seed words as input into our topic model and Figure A.1 plots the monthly count and length of *NYT* articles in our sample.

A.1 Seeded Latent Dirichlet Distribution

This appendix provides more details on the seeded latent Dirichlet distribution model. This paper uses a stochastic topic model to extract latent topic weights from news articles. Topic models are developed based on the core idea that documents are mixtures of topics, where each topic has a probability distribution over words (Blei 2012; Steyvers and Griffiths 2007). Under topic models, we assume that text documents derive from a stochastic generative process. The creation of a new document starts with a document-specific distribution over topics (the document-topic distribution). Each word in the document is chosen first by picking a topic randomly from the document-topic distribution and then drawing a word from the topic-word distribution for that topic. To model this, every possible word must be assigned to a topic.

In this setup, the document-topic distribution for each document and topic-word distribution for each topic (the same across documents) are unobserved parameters that are estimated from the observable word frequencies in the document collection. In other words, we can use standard statistical techniques to estimate the generative process, inferring the topics responsible for generating a collection of documents (Steyvers and Griffiths 2007).

The most widely used topic model is latent Dirichlet allocation (LDA) as introduced by Blei, Ng, and Jordan (2003) and further developed by Griffiths and Steyvers (2004). Under

LDA, a document d is generated under the following hierarchical process:

- The word weight vector ω_k of topic k is the vector of probabilities of each word value for the topic k. The prior for these weights is assumed to have a Dirichlet distribution governed by parameter β: $\omega_k \sim \text{Dirichlet}(\beta)$.[1]
- The topic weight in a document d, denoted τ_d, is a vector of topic probabilities, i.e., probabilities that any given word location in the document is about any given topic. The topic weight vector of document d follows a prior Dirichlet distribution governed by parameter α, the same for all documents: $\tau_d \sim \text{Dirichlet}(\alpha)$, the same for all documents.[2]
- We use v to indicate a word location in a given document and w to indicate a word value (such as "the" or "cat"). For each word location v in document d, we
  - randomly select a topic from the document-topic distribution:
    $z_{dv} \sim \text{Multinomial}(\tau_d)$ (a distribution which does not depend on v), and then
  - randomly select a word from that topic:
    $w \sim \text{Multinomial}(\omega_{z_{dv}})$.

In other words, it is the multinomial distribution of word values for the realized topic z_{dv}.

In this setup, the topic-word distribution ω_k and document-topic distribution τ_d are latent parameters that we want to estimate. Estimating these involves a backward inference based on observed word frequencies across documents. The parameters α and β are hyperparameters of the prior distribution whose values are taken from the Latent Dirichlet Distribution topic modeling literature.

[1] To illustrate, suppose that topic k has three words: $word_1$, $word_2$, and $word_3$ with weights $\omega_k = [w_1, w_2, w_3]$ with $w_1 + w_2 + w_3 = 1$. The model assumes that this ω_k vector follows a Dirichlet distribution.

[2] Similarly, assume document d has four topics $topic_1$, $topic_2$, $topic_3$, $topic_4$ with the weights given to these topics captured by $\tau_d = [\theta_1, \theta_2, \theta_3, \theta_4]$ with $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$. The model assumes that this τ_d vector follows a Dirichlet distribution.

The document-topic distribution τ_d is of utmost interest because it summarizes the attention allocated to each topic in each news article. To estimate these parameters using a Bayesian method, Griffiths and Steyvers (2004) specifies that ω_k and τ_d follow two Dirichlet distributions (these two are referred to as the "prior" distribution in Bayesian statistic). From these specifications, we can derive the distribution of the topic assignment z_{dv} conditioned on observed word frequencies (this conditional distribution is referred to as the "posterior" distribution). We then use Gibbs sampling to simulate this posterior distribution and estimate the two hidden model parameters.[3]

Users of the traditional unsupervised LDA developed by Blei, Ng, and Jordan (2003) and Griffiths and Steyvers (2004) only need to prespecify the number of topics K and let the model cluster words into these topics based on word frequencies in a completely unsupervised manner. Specifically, the LDA model is more likely to assign a word w to a topic k in a document d if w has been assigned to k across many different documents and k has been used multiple times in d (Steyvers and Griffiths 2007). The model automatically extracts underlying topics, so users of LDA have no control over topic assignments.

Since we are interested in uncovering specific topics, we employ a recent extension of LDA called seeded LDA (sLDA) developed by Lu et al. (2011). sLDA allows users to regulate topic contents using domain knowledge by injecting seed words (prior knowledge) into the model. Precisely, under sLDA, we specify the topic-word distribution as follows:

$$\omega_k \sim \text{Dirichlet}\left(\beta + C_w\right)_{w \in V}, \tag{A.1}$$

where V is the corpus or text collection, $C_w > 0$ when w is a seed word in topic k and $C_w = 0$ when w is not a seed word. The higher is C_w, the stronger the tilt toward word w

[3] Gibbs sampling is a sampling technique to simulate a high-dimensional distribution by sampling from lower-dimensional subsets of variables where each subset is conditioned on the value of all others. See Griffiths and Steyvers (2004) for details on the implementation of Gibbs sampling in LDA.

appearing in any given topic. Intuitively, sLDA gives preference to seed words w in topic k in the form of pseudo count C_w and clusters words into topics based on their co-occurrences with the seed words. When a seed word is not present in a text collection, it does not enter the sLDA model and has no impact on the estimation process.

Estimation is implemented by the seededlda package in R and run on a high-performance computing (HPC) cluster. Full estimation of the model parallelized on 80 computational nodes requires about one week to complete. Following standard practice, we set $\alpha = 50/K$ where K is the number of topics, $\beta = 0.1$, and $C_w = 0.01$ times the number of terms in the corpus.

A.2 Estimation

We use Gibbs sampling to estimate the model's parameters during each monthly estimation. We draw 200 drawings from the posterior distribution of z_{dv}, the realized topic for word location v in document d in the sLDA model, where we are conditioning on observed word frequencies.[4] In each drawing, we condition on the estimated values of the parameters of the model derived from previous drawing (where in the first draw, the initial estimate comes from a random number generator). In the last draw, we estimate our final value of the document-topic weights τ_d; that is, we estimate one 14×1 vector $\tau_d = [\tau_d^1, \tau_d^2, \ldots, \tau_d^{14}]$ for each news article, d, in the estimation window.

We then provide estimates of model parameters that condition on month t within the dataset. We compute the global monthly weights of each topic k ($k = 1, 2, \ldots, 14$) as the average weight of each topic across all articles in month t, weighted by the length $L(d)$ of

[4] In addition to the number of topics and articles, the number of samples drawn from the posterior distribution is a computational cost consideration in any topic model.

each article:

$$\tau_t^k = \frac{\sum_{d=1}^{n_t} \tau_d^k L(d)}{\sum_{d=1}^{n_t} L(d)}, \tag{A.2}$$

where τ_t^k is the weight of topic k in month t, n_t is the total number of news articles in month t, and $L(d)$ is the total number of unigrams (one-word terms), bigrams (two-word terms), and trigrams (three-word terms) in article d.[5] (Equal weighting of topic weights across articles yields similar results.)

[5] An n-gram is a sequence of n words. For instance, "San Diego" is a bigram, and "A study of topics is needed" is a 6-gram.

Figure A.1. ***NYT* Article Count and Length**

This figure plots the time series of the monthly total count and the monthly average length of articles in the *NYT*. Article length is measured as the sum of unigrams (one-word terms), bigrams (two-word terms), and trigrams (three-word terms) of each article. The sample period is from January 1871 to October 2019. Articles with limited content have been removed.

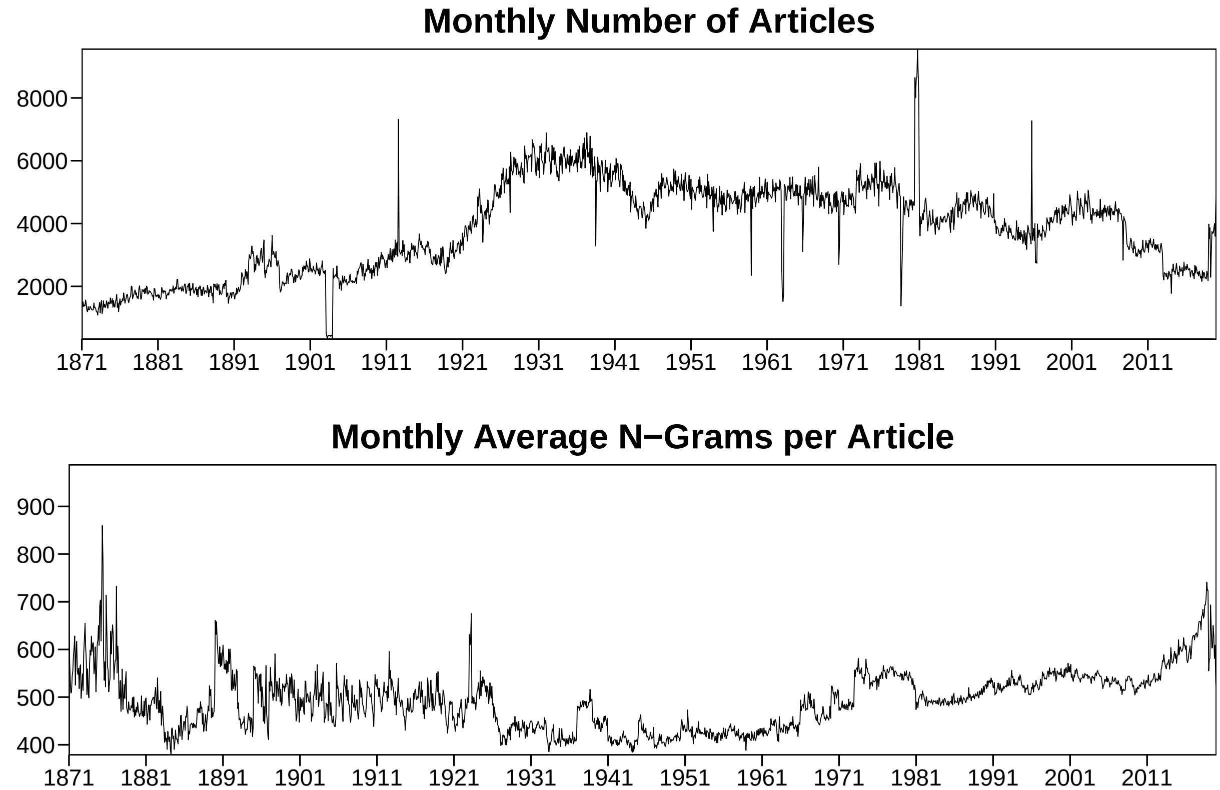

Table A.1
Seed Words

This table lists the *lemmatized* seed words for each of the 14 discourse topics. The first column presents the full name of the topic, and the second column reports the short name used in the paper.

Narrative	Short Name	Seed Words
War	War	conflict, tension, terrorism, war
Pandemic	Pandemic	epidemic, pandemic
Panic	Panic	bank failure, bank panic, bank run, crisis, depression, downturn, fear, financial panic, hard time, panic, recession
Confidence	Confidence	business confidence, consumer confidence
Frugality	Saving	compassion, family morale, frugal, frugality, modesty, moral, poverty, saving
Conspicuous Consumption	Consumption	american dream, conspicuous consumption, consumption, equal opportunity, equality, homeownership, luxury, patriotism, prosperity
Monetary Standard	Money	bimetallism, devaluation, gold, gold standard, inflation, monetary standard, money, silver
Techmology Replacing Jobs	Tech	automate, computer, digital divide, electronic brain, invention, labor save, labor save machine, machine, mechanize, network, technocracy, technological unemployment, technology, unemployment
Real Estate Booms	Real_estate_boom	boom, bubble, flip, flipper, home ownership, home purchase, house boom, house bubble, land boom, land bubble, price increase, real estate boom, real estate bubble, speculation
Real Estate Bursts	Real_estate_bust	bust, crash, house bust, house crash, land bust, land crash, price decrease, real estate bust, real estate crash
Stock Market Bubbles	Stock_bubble	advance market, boom, bubble, bull, bull market, bullish, earnings per share, inflate market, margin, margin requirement, market boom, market bubble, price earn ratio, price increase, sell short, short sell, speculation, stock market boom, stock market bubble
Stock Market Crashes	Stock_crash	bear, bear market, bearish, bust, crash, fall market, market crash, stock crash, stock market crash, stock market decline
Boycotts and Evil Business	Boycott	anger, boycott, community, evil business, excess profit, fair wage, moral, outrage, postpone purchase, profiteer, protest, strike, wage cut
Wage and Labor Unions	Wage	consumer price, cost of live, cost push, cost push inflation, high wage, increase wage, inflation, labor union, rise cost, wage, wage demand, wage lag, wage price, wage price spiral

B Additional Results with Our Test Portfolios

This subsection presents additional results with our own constructed portfolios in Table B.1 and Table B.2. We discuss these results in the main text.

Table B.1
War Factor and Risk Premium

This table presents the results from the second-pass cross-sectional regressions of average portfolio returns on factor betas:

$$\bar{R}^e_{i,t} = \lambda_0 + \beta'_{i,f}\lambda_f + e_i,$$

where $\bar{R}^e_{i,t}$ is the time-series average return of portfolio i, $\beta_{i,f}$ is the vector of factor exposures of portfolio i estimated via a multivariate time-series regression of portfolio returns onto factors, and λ_f is the vector of factor risk premiums. Test assets include 128 own constructed anomalies in Panel A and 1173 own constructed nonlinear portfolios in Panel B. "WarFac" is the scaled innovations in *NYT War* ($WarFac_t$); "MKT, SMB, HML, RMW, CMA, MOM" are Fama and French (2018) six factors; "MKT, SMB, MGMT, PERF" are Stambaugh and Yuan (2017) mispricing factors; "PEAD" and "FIN" are Daniel, Hirshleifer, and Sun (2020) behavioral factors; and "R_MKT, R_ME, R_IA, R_ROE, R_EG" are Hou et al. (2021) Q5 factors. Reported are monthly risk premium λ and t-statistic with Shanken (1992) correction. R^2 is cross-sectional R^2 in percentages, MAPE is mean absolute pricing error in percentages, and $p(\text{PE} = 0)$ is p-value of the Chi-squared test that pricing errors are jointly zero. N is the number of test portfolios, and T is the number of months. The sample is from July 1972 to December 2016. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Own Constructed Anomalies

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.09 *	0.02	0.04	-0.02	0.00	0.00	0.05	0.05	0.01	-0.01
	(1.77)	(1.08)	(1.58)	(-0.79)	(0.03)	(0.07)	(1.26)	(0.99)	(0.15)	(-0.42)
WarFac	-19.99 ***					-25.76 ***	-22.25 ***	-23.99 ***	-18.26 ***	-18.49 ***
	(-2.89)					(-5.09)	(-4.74)	(-3.98)	(-4.59)	(-4.37)
MKT		0.09	0.41	0.82 **		-0.16	-0.09	0.25		0.20
		(0.33)	(1.48)	(2.51)		(-0.39)	(-0.24)	(0.52)		(0.50)
SMB		0.24	0.34 **			0.17	0.16			0.21
		(1.57)	(2.00)			(0.86)	(0.72)			(1.04)
HML		0.33 **				0.46 **				0.47 **
		(2.16)				(2.18)				(2.39)
RMW		0.09				0.22				0.01
		(0.74)				(1.36)				(0.06)
CMA		0.26 ***				0.27 *				0.24 *
		(2.62)				(1.94)				(1.76)
MOM		0.74 ***				0.85 ***				0.53 **
		(3.50)				(3.13)				(2.09)
MGMT			0.50 ***				0.43 **			0.54 ***
			(3.26)				(2.11)			(2.75)
PERF			0.63 ***				0.67 **			0.29
			(2.86)				(2.24)			(0.98)
PEAD				0.47 ***				0.52 *		0.76 ***
				(2.91)				(1.91)		(2.58)
FIN				0.42 **				0.50 *		1.10 ***
				(2.08)				(1.90)		(3.80)
R_MKT					0.50				0.13	
					(1.59)				(0.36)	
R_ME					0.51 ***				0.58 **	0.71 ***
					(2.59)				(2.47)	(3.37)
R_IA					0.33 ***				0.28 **	0.31 **
					(2.70)				(1.98)	(2.14)
R_ROE					0.15				0.25	0.40 **
					(0.97)				(1.50)	(2.40)
R_EG					1.26 ***				0.85 ***	1.08 ***
					(7.43)				(4.40)	(5.96)
R^2	31	23	30	14	47	47	47	41	56	58
MAPE	0.31	0.29	0.28	0.33	0.25	0.27	0.26	0.29	0.24	0.24
$p(PE = 0)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	128	128	128	128	128	128	128	128	128	128
T	532	532	532	532	532	532	532	532	532	532

Table B.1
***War* Factor and Risk Premium (Cont.)**

Panel B: Own Constructed Nonlinear Portfolios

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.99 ***	0.96 ***	1.03 ***	1.37 ***	1.02 ***	0.78 ***	0.90 ***	0.77 ***	0.84 ***	0.54 **
	(3.14)	(4.72)	(5.07)	(5.28)	(4.71)	(3.27)	(3.67)	(2.66)	(3.19)	(2.39)
WarFac	-11.13 ***					-11.05 ***	-10.67 ***	-13.97 ***	-11.79 ***	-9.51 ***
	(-3.55)					(-6.14)	(-4.88)	(-4.48)	(-6.53)	(-6.47)
MKT		-0.29	-0.32	-0.29		-0.02	-0.03	0.28		0.24
		(-0.96)	(-1.04)	(-0.94)		(-0.05)	(-0.08)	(0.82)		(0.79)
SMB		0.32 *	0.32			0.10	-0.00			0.03
		(1.84)	(1.61)			(0.56)	(-0.02)			(0.16)
HML		0.15				0.16				0.26
		(0.91)				(0.85)				(1.44)
RMW		0.23				0.42 **				0.37 **
		(1.46)				(2.53)				(2.52)
CMA		0.57 ***				0.50 ***				0.51 ***
		(4.98)				(3.99)				(4.22)
MOM		0.53 **				0.76 ***				0.80 ***
		(2.33)				(3.28)				(3.43)
MGMT			0.48 ***				0.51 ***			0.40 **
			(2.90)				(2.74)			(2.14)
PERF			0.58 **				0.81 ***			1.39 ***
			(2.36)				(2.96)			(5.08)
PEAD				0.29 *				0.55 **		0.21
				(1.70)				(2.47)		(1.18)
FIN				0.38				0.79 ***		1.25 ***
				(1.52)				(2.87)		(5.22)
R_MKT					-0.26				0.02	
					(-0.87)				(0.07)	
R_ME					0.29 *				0.17	0.04
					(1.72)				(0.97)	(0.26)
R_IA					0.33 **				0.31 **	0.33 **
					(2.51)				(2.01)	(2.57)
R_ROE					0.24				0.52 **	0.53 ***
					(1.27)				(2.54)	(2.95)
R_EG					0.88 ***				0.92 ***	0.88 ***
					(4.48)				(3.86)	(5.52)
R^2	16	48	46	30	48	55	54	49	56	63
MAPE	0.10	0.07	0.07	0.09	0.07	0.07	0.07	0.08	0.07	0.06
$p(PE = 0)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N	2190	2190	2190	2190	2190	2190	2190	2190	2190	2190
T	532	532	532	532	532	532	532	532	532	532

Table B.2
War Mimicking Portfolio and Risk Premium

This table presents the results from the second-pass cross-sectional regressions of average portfolio returns on factor betas:

$$\bar{R}^e_{i,t} = \lambda_0 + \beta'_{i,f}\lambda_f + e_i,$$

where $\bar{R}^e_{i,t}$ is the time-series average return of portfolio i, $\beta_{i,f}$ is the vector of factor exposures of portfolio i estimated via a multivariate time-series regression of portfolio returns onto factors, and λ_f is the vector of factor risk premiums. Test assets include constructed anomaly portfolios in Panel A and constructed nonlinear portfolios in Panel B. "WMP" is the *War* mimicking portfolio; "MKT, SMB, HML, RMW, CMA, MOM" are Fama and French (2018) six factors; "MKT, SMB, MGMT, PERF" are Stambaugh and Yuan (2017) mispricing factors; "PEAD" and "FIN" are Daniel, Hirshleifer, and Sun (2020) behavioral factors; and "R_MKT, R_ME, R_IA, R_ROE, R_EG" are Hou et al. (2021) Q5 factors. Reported are monthly risk premium λ and t-statistic with Shanken (1992) correction. R^2 is cross-sectional R^2 in percentages, MAPE is mean absolute pricing error in percentages, and $p(\text{PE} = 0)$ is the p-value of the Chi-squared test that pricing errors are jointly zero. N is the number of test portfolios, and T is the number of months. The sample is from July 1972 to December 2016. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Own Constructed Anomalies

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.07 ***	0.02	0.03	-0.02	-0.00	0.06 **	0.02	0.03	-0.02	0.03
	(2.65)	(1.05)	(1.43)	(-0.84)	(-0.03)	(2.20)	(0.67)	(1.24)	(-0.60)	(1.01)
WMP	-1.18 ***					-3.44 ***	-2.06 ***	-1.36 ***	-1.92 ***	-3.32 ***
	(-3.87)					(-8.31)	(-6.26)	(-3.93)	(-4.63)	(-6.56)
MKT		0.01	0.33	0.77 **		0.51	0.60 **	0.22		0.91 **
		(0.03)	(1.19)	(2.25)		(1.58)	(2.07)	(0.76)		(2.37)
SMB		0.21	0.31 *			-0.32 *	-0.03			-0.16
		(1.37)	(1.84)			(-1.79)	(-0.18)			(-0.79)
HML		0.32 **				0.12				0.27
		(2.12)				(0.70)				(1.39)
RMW		0.08				-0.02				-0.06
		(0.69)				(-0.16)				(-0.41)
CMA		0.26 ***				0.31 ***				0.18
		(2.63)				(2.69)				(1.33)
MOM		0.71 ***				0.62 ***				0.38
		(3.36)				(2.64)				(1.57)
MGMT			0.48 ***				0.09			0.67 ***
			(3.19)				(0.59)			(3.45)
PERF			0.61 ***				0.68 ***			-0.28
			(2.78)				(2.95)			(-0.92)
PEAD				0.46 ***				0.56 ***		0.83 ***
				(2.88)				(3.44)		(2.92)
FIN				0.45 **				0.41 **		0.58 **
				(2.24)				(1.97)		(2.01)
R_MKT					0.39				0.83 ***	
					(1.25)				(2.66)	
R_ME					0.45 **				0.18	0.04
					(2.32)				(0.81)	(0.17)
R_IA					0.31 ***				0.14	0.45 ***
					(2.61)				(1.25)	(3.21)
R_ROE					0.13				0.07	0.29 *
					(0.84)				(0.48)	(1.77)
R_EG					1.22 ***				1.05 ***	1.19 ***
					(7.40)				(5.84)	(6.78)
R^2	22	22	29	15	45	39	35	35	50	59
MAPE	0.34	0.29	0.28	0.32	0.26	0.30	0.29	0.29	0.26	0.25
$p(PE = 0)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	128	128	128	128	128	128	128	128	128	128
T	534	534	534	534	534	534	534	534	534	534

Table B.2
***War* Mimicking Portfolio and Risk Premium (Cont.)**

Panel B: Own Constructed Nonlinear Portfolios

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.75 ***	0.96 ***	1.03 ***	1.31 ***	1.01 ***	0.62 ***	0.65 ***	0.39 **	0.60 ***	0.43 **
	(2.69)	(4.75)	(5.06)	(5.15)	(4.72)	(3.59)	(3.42)	(2.10)	(3.21)	(2.29)
WMP	-1.35 ***					-1.24 ***	-1.24 ***	-1.49 ***	-1.31 ***	-1.10 ***
	(-4.37)					(-3.97)	(-3.88)	(-4.54)	(-4.29)	(-3.70)
MKT		-0.32	-0.35	-0.28		0.07	0.10	0.36		0.27
		(-1.06)	(-1.14)	(-0.92)		(0.25)	(0.38)	(1.25)		(0.98)
SMB		0.29	0.30			0.10	0.02			0.13
		(1.64)	(1.49)			(0.60)	(0.11)			(0.75)
HML		0.18				0.24				0.29 *
		(1.06)				(1.43)				(1.69)
RMW		0.22				0.35 **				0.22
		(1.41)				(2.44)				(1.52)
CMA		0.56 ***				0.49 ***				0.53 ***
		(4.96)				(4.33)				(4.61)
MOM		0.49 **				0.56 **				0.53 **
		(2.18)				(2.49)				(2.30)
MGMT			0.49 ***				0.46 ***			0.46 ***
			(2.94)				(2.69)			(2.63)
PERF			0.54 **				0.59 **			1.15 ***
			(2.21)				(2.39)			(4.76)
PEAD				0.28				0.39 **		0.09
				(1.63)				(2.12)		(0.58)
FIN				0.40				0.77 ***		1.12 ***
				(1.58)				(3.26)		(5.01)
R_MKT					-0.27				0.12	
					(-0.93)				(0.46)	
R_ME					0.24				0.16	0.02
					(1.46)				(0.99)	(0.15)
R_IA					0.34 ***				0.33 **	0.31 ***
					(2.59)				(2.50)	(2.60)
R_ROE					0.22				0.33 *	0.38 **
					(1.18)				(1.87)	(2.26)
R_EG					0.87 ***				0.65 ***	0.74 ***
					(4.42)				(3.52)	(5.00)
R^2	25	48	46	32	48	52	52	51	53	59
MAPE	0.09	0.07	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07
$p(PE=0)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
N	2190	2190	2190	2190	2190	2190	2190	2190	2190	2190
T	534	534	534	534	534	534	534	534	534	534

C Protocol for Factor Identification

C.1 First Criterion: Correlation of factors with the Systematic Risk of Returns

If the factor mimicking portfolio of an observed factor represents a risk factor, it should be related to the systematic risk of returns. Following Pukthuanthong, Roll, and Subrahmanyam (2019) (hereafter, PRS), we test whether our War mimicking portfolio (WMP) is related to the cross-sectional covariance of asset returns. Specifically, we apply the asymptotic approach of Connor and Korajczyk (1988) (CK) to extract ten principal components from the equities return series. The principal components of the covariance matrix of returns represent the systematic part of the asset returns. We then compute canonical correlations between the ten CK principal components and the factor candidates and test the significance of these canonical correlations.

The implementation of the PRS approach comprises three steps. First, we collect a set of N equities for the factor candidates. The test assets should be from industries with enough heterogeneity to detect the underlying risk premium associated with factors. Second, we apply the CK approach to extract L principal components (PCs) from the return series. With T time-series units up to time t, we compute the $T \times T$ matrix $\Omega_t = \frac{1}{T}RR'$, where R is the return vector. CK demonstrate that for large N, analyzing the eigenvectors of Ω_t is asymptotically equivalent to factor analysis. The first L eigenvectors of Ω_t form the factor estimates. The cutoff point for $L < N$ is chosen so that the PCs explain at least 90% of the cumulative variance. Third, we collect a set of K factor candidates. Our study includes 14 factors, including *WMP*, five factors from FF6, three from M4, four from Q5, two from DHS, and one from the market.

Finally, from the second step above, we compute the canonical correlation between the factor candidates and the corresponding eigenvectors. First, we use the L eigenvectors from step 2 and the K factor candidates from step 3 and calculate the covariance matrix over a sample period t, V_t $(L+K \times L+K)$. We break out a submatrix from the covariance matrix V_t in each period, the cross-covariance matrix, denoted by C_t having K rows and L columns. The entry in the i^{th} row and j^{th} column is the covariance between factor candidate i and eigenvector j. We need to break out the covariance submatrix of the factor candidates, $V_{f,t}$ $(K \times K)$, and the covariance submatrix of the real eigenvectors, $V_{e,t}$ $(L \times L)$. We then can find two weighting column vectors, λ_t and κ_t, on the factor candidates and eigenvectors, respectively (λ_t has K rows, κ_t has L rows), that maximize the correlation between the two weighted vectors. The covariance between the weighted averages of factor candidates and eigenvectors is $\lambda_t' C_t \kappa_t$, and their correlation is

$$\rho = \frac{\lambda_t' C_t \kappa_t}{\sqrt{\lambda_t' V_{f,t} \lambda_t \kappa_t' V_{e,t} \kappa_t}} \tag{C.1}$$

We maximize the correlation across all choices of λ_t and κ_t. The maximum exists when the weight is $\lambda_t = V_{f,t}^{-1/2} h_t$, where h_t is the eigenvector corresponding to the maximum eigenvalue in the matrix $V_{f,t}^{-1/2} C_t V_{e,t}^{-1} C_t' V_{f,t}^{-1/2}$. κ_t is proportional to h_t.

We maximize the correlation again, subject to the constraint that the new vectors are orthogonal to the old ones, and so on. As a result, there are $min(L, K)$ pairs of orthogonal canonical variables sorted from the highest correlation to the smallest. We transform each correlation into a variable asymptotically distributed as Chi-Square under the null hypothesis that the actual correlation is zero. This provides a method of testing whether the factor candidates are conditionally related (on date t) to the covariance matrix of returns. Also, by examining the relative sizes of the weightings in λ_t, we can understand which factor candidates are more related to real return covariances. The intuition behind the canonical correlation approach is that the proper underlying drivers of returns are undoubtedly changes in perceptions about macroeconomic variables. But the factor candidates and the eigenvectors need not be isomorphic to a particular macro variable. Instead, each candidate or eigenvector is some linear combination of all the pertinent macro variables. This is the well-known "rotation" problem in principal components or factor analysis. The PRS criteria assert that some linear combinations of the factor candidates are strongly related to different linear combinations of the eigenvectors representing the actual factors. Canonical correlation is intended for this application. Any factor candidate that does not display a significant (canonical) correlation with its associated best linear combination of eigenvectors can be rejected as a viable factor. It is not significantly associated with the covariance matrix of asset returns.

We compute asymptotic PCs that represent the covariance matrix. We split the overall sample into five subsamples with ten years.[6] For each subsample, we use CK to extract PCs and retain the first ten PCs, which account for close to 90% of the cumulative eigenvalues or the total volatility in the covariance matrix, implying they capture most of the stock variations.

Next, we proceed to estimate the canonical correlations. We have several factor candidates and, thus, several pairs of canonical variates. We take the following steps to derive the significance levels of each factor candidate reported in the first row of Table C.1. First, for each of the ten canonical pairs, the eigenvector weights for the ten CK PCs are taken, and the weighted average CK PC or the canonical variate for the ten CK PCs that produced the canonical correlation for this particular pair is constructed.[7] Then a regression using each

[6]These five subsamples are 1967-1976, 1977-1986, 1987-1996, 1997-2006, and 2007-2016.

[7]There are $min(L, K)$ possible pairs. In our application, $L = 10$ and $K = 14$.

CK PC canonical variate as the dependent variable and the actual candidate factor values as independent variables are run over the sample months in each subperiod. The square root of the R-squared from the regression is the canonical correlation. After proper normalization, the coefficients for the regressions are equal to the eigenvector's weighting elements for the candidate factors. The t-statistic from the regression then gives the significance level of each candidate factor. With the ten pairs of canonical variates in each subperiod, and a canonical correlation for each one, we have 50 such regressions. The first row presents the mean t-statistic of all canonical correlations. The second row shows the mean t-statistic across cases when the canonical correlation is statistically significant. The last row shows the average number of significant canonical correlations across subperiods.

A risk factor *must* satisfy the necessary and sufficient conditions: (1) the risk factor is significantly related to any canonical variate in all decades, or it has a mean t-statistic exceeding the one-tailed 2.5% cutoff based on the Chi-squared value, and (2), in each subperiod, the risk factor has an average number of significant canonical correlations exceeding 2.50 (the bottom row of each panel). Researchers should test the augmented condition (the third condition) to ensure the robustness of the result. We leave it for other researchers to implement it. [8]

We examine this criterion for the WMP. As suggested in Pukthuanthong, Roll, and Subrahmanyam (2019), we use individual stocks to test the necessary condition. Our stock universe are stocks in the CRSP/Compustat merge file. Following standard practice, we exclude

[8]Pukthuanthong, Roll, and Subrahmanyam (2019) require an average number of significant decade t-statistics exceeding 2.5 from 10 canonical variates (one-fourth of the total canonical variates). We use the same criteria as ten canonical variates (see the previous footnote). The reason to choose this value comes from Pukthuanthong, Roll, and Subrahmanyam (2019): "This is a conservative threshold to ensure we do not miss a true factor at our necessary condition stage. We focus on the significant canonical correlations rather than all canonical correlations because insignificant CCs imply that none of the factors matter, so using them would be over-fitting."

financial and utilities companies. We then split our sample from July 1972 to December 2016 into five subperiods: 1972-1976, 1977-1986, 1987-1996, 1997-2006, and 2007-2016.[9] We keep only stocks with non-missing returns during the five subperiods. As reported in Table C.1, WMP passes this criterion with an average t-stat of the significant canonical correlations (CC) of 2.26 and the average number of the significant CCs of 2.6, above the 2.5 threshold. Besides WMP, market (MKT) and momentum (MOM) are the other factors that pass this condition. MKT represents the most substantial pass. We conclude that our candidate global risk factors are WMP, MKT, and MOM.

C.2 Second Criterion: Risk Premium Estimation using WMP

The second criterion of the PRS protocol requires that the global risk factors or the factors that pass the necessary condition command a risk premium in the cross-section of asset returns. To perform this step, we re-run the standard two-pass regressions with the factors that pass the necessary condition. We report in Table C.2 that WMP prices all six sets of test assets after controlling for the other two risk factors that pass the first criterion of the protocol. MKT does not price any set of test assets with a positive risk premium, while MOM fails to price the 360 machine learning-based portfolios. We conclude that our *War* factor is a genuine risk factor.

[9]The sample period from July 1972 to December 2016 is when data on all factors are available.

Table C.1
Protocol Step 1: Correlation of WMP with the Systematic Risk of Returns

This table presents the results from the first step (necessary condition) of the protocol in Pukthuanthong, Roll, and Subrahmanyam (2019). Test assets are individual stocks in the CRSP/Compustat merge file, excluding financial and utilities firms. "*WMP*" is the *War* mimicking portfolio; "MKT, SMB, HML, RMW, CMA, MOM" are Fama and French (2018) six factors; "MKT, SMB, MGMT, PERF" are Stambaugh and Yuan (2017) mispricing factors; "PEAD" and "FIN" are Daniel, Hirshleifer, and Sun (2020) behavioral factors; and "R_ME, R_IA, R_ROE, R_EG" are Hou et al. (2021) Q5 factors. The sample is from July 1972 to December 2016.

	WMP	MKT	SMB	HML	RMW	CMA	MOM	MGMT	PERF	PEAD	FIN	R_ME	R_IA	R_ROE	R_EG
Avg. t	1.46	5.93	1.27	1.27	1.27	1.15	1.58	1.20	1.36	1.00	1.27	1.36	1.16	1.17	1.08
Avg. t (Sig. CC)	2.26	11.32	1.71	1.59	1.57	1.56	2.25	1.65	1.91	1.17	1.49	1.86	1.62	1.58	1.45
Period 1972:1976	1	4	2	3	3	1	1	1	2	0	0	2	2	1	1
Period 1977:1986	3	3	2	1	3	2	4	2	4	1	3	3	4	3	0
Period 1987:1996	3	2	2	2	1	1	3	2	0	2	5	0	2	1	3
Period 1997:2006	2	3	2	1	2	3	4	2	2	1	2	2	2	3	3
Period 2007:2016	4	3	3	2	3	3	3	3	3	2	2	3	2	3	2
# Sign. CC	2.60	3.00	2.20	1.80	2.40	2.00	3.00	2.00	2.20	1.20	2.40	2.00	2.40	2.20	1.80

Table C.2
Protocol Step 2: Two-Pass Regressions

This table presents the results from the second step (sufficient condition) of the protocol in Pukthuanthong, Roll, and Subrahmanyam (2019), i.e., second-pass cross-sectional regressions of average portfolio returns on factor betas:

$$\bar{R}^e_{i,t} = \lambda_0 + \beta'_{i,f}\lambda_f + e_i,$$

where $\bar{R}^e_{i,t}$ is the time-series average return of portfolio i, $\beta_{i,f}$ is the vector of factor exposures of portfolio i estimated via a multivariate time-series regression of portfolio returns onto factors, and λ_f is the vector of factor risk premiums. Test assets include long-short portfolios from Hou, Xue, and Zhang (2020) ("HXZ LS"), single-sorted portfolios from Hou, Xue, and Zhang (2020) ("HXZ Single"), single-sorted portfolios from Chen and Zimmermann (2022) ("CZ Single"), ML-based nonlinear portfolios from Bryzgalova, Pelger, and Zhu (2020) ("ML-Based"), own constructed anomalies, and own constructed nonlinear portfolios. "*WMP*" is the *War* mimicking portfolio; and "MKT" and MOM" are from Fama and French (2018). Reported are monthly risk premium λ and t-statistic with Shanken (1992) correction. R^2 is cross-sectional R^2 in percentages, MAPE is mean absolute pricing error in percentages, and $p(\text{PE} = 0)$ is the p-value of the Chi-squared test that pricing errors are jointly zero. N is the number of test portfolios, and T is the number of months. The sample is from July 1972 to December 2016. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	HXZ LS	HXZ Single	CZ Single	ML-Based	Own Anomalies	Own Nonlinear
Intercept	0.16 ***	0.68 ***	0.80 ***	1.97 ***	0.03	1.30 ***
	(5.24)	(2.64)	(2.58)	(4.79)	(1.35)	(3.49)
WMP	-1.13 ***	-1.06 ***	-1.65 ***	-2.17 ***	-1.40 ***	-1.14 ***
	(-3.36)	(-3.52)	(-4.65)	(-4.01)	(-4.32)	(-3.02)
MKT	-0.36	-0.12	-0.25	-1.93 ***	-0.02	-0.52
	(-1.08)	(-0.35)	(-0.66)	(-4.28)	(-0.05)	(-1.33)
MOM	0.57 ***	0.56 ***	0.72 ***	0.29	0.73 ***	0.66 ***
	(2.72)	(2.75)	(3.29)	(1.23)	(3.42)	(2.86)
R^2	49	39	42	45	32	45
MAPE	0.26	0.09	0.15	0.47	0.29	0.08
$p(PE = 0)$	0.00	0.03	0.00	0.00	0.00	1.00
N	138	1372	904	360	128	2190
T	534	534	534	534	534	534

D Details on Portfolio Construction

D.1 128 Own Constructed Anomalies

In this subsection, we report our constructed portfolios' descriptive statistics based on Hou, Xue, and Zhang (2020).

Table D.1
128 Own Constructed Anomalies

This table presents descriptive statistics of the portfolios we apply in our cross-sectional tests. Our sample period is from 1967 to 2016. The candidate factors are constructed similarly to Hou, Xue, and Zhang (2020). We use the same screening criteria, delisting procedure, and period similar to what they do. The first column presents the identification numbers and names of the candidate factors according to their papers. The last four columns present the number of observations, the mean of candidate factors, t-stat testing the mean is statistically different from zero, and the standard deviation of candidate factors. All candidate factors are based on one-month calculation, and these portfolios are equal-weighted returns. ***, **, and * present 1%, 5%, and 10% significance level.

Candidate factors	# obs	mean	t-stat	std.dev
A. Momentum				
A.1.1 Standardized unexpected earnings	534	0.01	4.97***	0.04
A.1.2 Cumulative abnormal returns around earnings announcement dates	521	0.02	8.57***	0.04
A.1.4 Price momentum, prior 6-month returns	534	0.01	3.13***	0.08
A.1.5 Price momentum, prior 11-month returns	534	0.01	4.24***	0.08
A.1.6 Industry momentum	534	0.57	2.23**	5.90
A.1.7 Revenue surprises	534	0.00	1.37	0.04
A.1.10 The number of quarters with consecutive earnings increase	533	0.00	1.76*	0.07
A.1.11 52-week high	529	-0.00	-0.18	0.07
A.1.12 Residual momentum, prior 6-month returns	534	0.00	1.40	0.06
A.1.13 Residual momentum, prior 11-month returns	534	0.01	3.89***	0.06
B. Value versus growth				
B.2.1 Book-to-market equity	534	0.00	2.17**	0.05
B.2.2 Book-to-June-end market equity	534	0.00	2.33**	0.05
B.2.3 Quarterly book-to-market equity	534	0.02	6.88***	0.06
B.2.6 Assets-to-market	534	0.00	2.07**	0.06
B.2.8 Reversal.	534	-0.00	-1.81*	0.06
B.2.9 Earnings-to-price	534	0.00	0.93	0.06
B.2.12 Cash flow-to-price	534	0.00	0.12	0.05
B.2.14 Dividend yield	534	0.00	1.01	0.04
B.2.16 Payout yield	529	0.00	2.59**	0.05
B.2.16 Net payout yield	529	0.00	2.51**	0.05
B.2.18 5-year sales growth rank	534	-0.00	-0.72	0.04
B.2.19 Sales growth	534	-0.00	-1.13	0.04
B.2.20 Enterprise multiple	534	-0.00	-2.00**	0.06
B.2.22 Sales-to-price	534	0.01	2.68**	0.06
B.2.26 Intangible return	534	-0.01	-4.88***	0.04
B.2.30 Equity duration	534	-0.01	-3.18***	0.06
C. Investment				
C.3.1 Abnormal corporate investment	534	-0.00	-2.11**	0.03
C.3.2 Investment-to-assets	534	0.00	4.03***	0.01
C.3.3 Quarterly investment-to-assets	522	-0.00	-0.67	0.03
C.3.4 Changes in PPE and inventory-to-assets	534	-0.00	-3.01***	0.03
C.3.5 Noa and dNoa, (changes in) net operating assets	534	-0.01	-4.06***	0.03
C.3.6 Changes in long-term net operating assets.	534	-0.00	-3.00**	0.03
C.3.7 Investment growth	534	-0.00	-3.48***	0.03
C.3.8 2-year investment growth	534	-0.00	-1.93**	0.03
C.3.9 3-year investment growth	534	-0.00	-1.38	0.03
C.3.10 Net stock issues	534	-0.00	-3.55***	0.03

C.3.11 Percentage change in investment relative to industry	534	-0.00	-2.34**	0.03
C.3.12 Composite equity issuance	534	-0.00	-0.82	0.04
C.3.13 Composite debt issuance	534	-0.00	-0.42	0.04
C.3.14 Inventory growth	534	-0.00	-2.06**	0.03
C.3.15 Inventory changes	534	-0.00	-2.92***	0.03
C.3.16 Operating accruals	534	-0.00	-2.17**	0.03
C.3.17 Total accruals	534	-0.00	-1.96*	0.04
C.3.18 Changes in net noncash working capital, in current operating assets, and in current operating liabilities	534	-0.00	-1.12	0.04
C.3.19 Changes in noncurrent operating assets	534	-0.00	-3.42***	0.03
C.3.19 Changes in noncurrent operating liabilities	534	-0.00	-0.87	0.03
C.3.19 Changes in net noncurrent operating assets	534	-0.00	-3.34***	0.03
C.3.20 Changes in book equity	534	-0.00	-0.28	0.05
C.3.20 Changes in net financial assets	534	0.00	2.04**	0.03
C.3.20 Changes in financial liabilities	534	-0.00	-1.34	0.02
C.3.20 Changes in in long-term investments	534	-0.00	-1.36	0.03
C.3.20 Changes in short-term investments	534	0.00	0.39	0.02
C.3.21 Discretionary accruals computed from Nasdaq Index	516	-0.00	-1.94*	0.04
C.3.21 Discretionary accruals computed from NYSE and Amex	534	-0.00	-1.41	0.03
C.3.22 Percent operating accruals	534	-0.00	-3.06***	0.03
C.3.23 Percent total accruals	534	-0.00	-1.42	0.03
C.3.24 Percent discretionary accruals	534	-0.00	-2.31**	0.03
C.3.25 Net debt financing	528	-0.00	-1.94*	0.03
C.3.25 Net equity financing	528	-0.00	-0.80	0.05
C.3.25 Net external financing	528	-0.00	-1.83*	0.04
D. Profitability				
D.4.1 Return on equity	534	0.02	8.13***	0.05
D.4.2 4-quarter change in return on equity	528	0.00	2.71**	0.04
D.4.3 Roa1, Roa6, and Return on assets	534	0.01	7.40***	0.05
D.4.4 4-quarter change in return on assets.	522	0.00	2.81***	0.04
D.4.5 Assets turnover	534	0.00	0.54	0.04
D.4.5 Profit margin	534	0.00	0.24	0.05
D.4.5 Return on net operating assets	534	0.00	0.48	0.04
D.4.6 Capital turnover	534	0.00	0.89	0.04
D.4.7 Quarterly assets turnover	534	0.00	2.18**	0.04
D.4.7 Quarterly profit margin	534	0.00	2.43**	0.05
D.4.7 Quarterly return on net operating assets	486	0.00	2.37**	0.04
D.4.8 Quarterly capital turnover	534	0.01	3.63***	0.04
D.4.9 Gross profits-to-assets.	534	0.00	1.90*	0.03
D.4.10 Gross profits-to-lagged assets	534	0.00	0.20	0.04
D.4.11 Quarterly gross profits-to-lagged assets	486	0.00	3.14***	0.03

D.4.12 Operating profits to equity	534	0.00	1.14	0.05
D.4.13 Operating profits-to-lagged equity	534	0.00	0.40	0.04
D.4.14 Quarterly operating profits-to-lagged equity	534	0.01	3.39***	0.06
D.4.15 Operating profits-to-assets	534	0.00	2.04**	0.04
D.4.16 Operating profits-to-lagged assets	534	0.00	1.41	0.04
D.4.17 Quarterly operating profits-to-lagged assets	486	0.01	4.30***	0.04
D.4.18 Cash-based operating profitability	534	0.01	3.53***	0.04
D.4.19 Cash-based operating profits-to-lagged asset	534	0.00	2.76***	0.04
D.4.20 Quarterly cash-based operating profits-to-lagged assets	486	0.01	4.32***	0.04
D.4.21 Fundamental score.	528	0.00	1.70*	0.03
D.4.24 Ohlsons O-score	534	0.00	0.32	0.04
D.4.25 Quarterly O-score	486	-0.00	-1.26	0.03
D.4.26 Altmans Z-score	534	-0.00	-2.01**	0.05
D.4.27 Quarterly Z-score	486	-0.00	-2.07**	0.05
D.4.29 Taxable income-to-book income.	534	0.00	0.24	0.03
D.4.30 Quarterly taxable income-to-book income	534	0.00	0.58	0.04
D.4.31 Growth score	348	0.00	1.08	0.08
D.4.32 Book leverage	534	0.00	0.44	0.04
D.4.33 Quarterly book leverage	534	0.00	0.14	0.04
E. Intangibles				
E.5.1 Industry adjusted organizational capital-to-assets	534	0.00	0.35	0.04
E.5.2 Advertising expense-to-market	534	0.00	0.17	0.03
E.5.3 Growth in advertising expense.	534	0.00	3.38***	0.01
E.5.4 R&D expense-to-market	534	-0.00	-0.93	0.04
E.5.8 Operating leverage	534	0.00	0.44	0.03
E.5.9 Olq1, Olq6, and Olq12, quarterly operating leverage	522	0.00	2.55**	0.03
E.5.10 Hiring rate	534	0.00	2.94***	0.01
E.5.11 R&D capital-to-assets	534	0.00	0.25	0.04
E.5.12 Bca, brand capital-to-assets.	516	0.01	2.05**	0.07
E.5.17 Ha, industry concentration (assets)	534	-0.00	-1.25	0.05
E.5.17 He, industry concentration (book equity)	534	-0.00	-1.10	0.04
E.5.17 Hs, industry concentration (sales)	534	-0.00	-1.39	0.04
E.5.19 D1, price delay	534	0.00	0.98	0.04
E.5.19 D2, price delay	534	0.00	-0.11	0.02
E.5.19 D3, price delay	534	0.00	-0.41	0.02
E.5.20 % change in sales minus % change in inventory	534	0.00	0.44	0.00
E.5.21 % change in sales minus % change in accounts receivable	534	0.00	1.08	0.01
E.5.22 % change in gross margin minus % change in sales	534	0.00	2.28**	0.01
E.5.23 % change in sales minus % change in SG&A	534	0.00	1.43	0.00
E.5.24 Effective tax rate	534	0.00	1.43	0.00

E.5.25 Labor force efficiency	534	0.00	1.17	0.00
E.5.26 Analysts coverage	485	-0.00	-0.43	0.03
E.5.27 Tangibility	534	-0.00	-0.83	0.03
E.5.28 Quarterly tangibility.	534	0.00	0.16	0.03
E.5.29 Industry-adjusted real estate ratio	534	0.00	0.47	0.04
E.5.30 Financial constraints (the Kaplan-Zingales index)	534	0.00	1.52	0.03
E.5.32 Financial constraints (the Whited-Wu index)	534	0.00	0.12	0.03
E.5.33 Wwq1, Wwq6, and Wwq12, the quarterly Whited-Wu index	534	0.00	0.33	0.04
E.5.34 Secured debt-to-total debt	534	-0.00	-0.65	0.03
E.5.35 Convertible debt-to-total debt	534	0.00	0.83	0.04
E.5.37 Cta1, Cta6, and Cta12, cash-to-assets	534	0.00	1.08	0.04
E.5.41 Earnings persistence	534	-0.00	-0.66	0.03
E.5.41 Earnings predictability	534	-0.00	-2.16**	0.04
E.5.42 Earnings smoothness	534	-0.00	-1.01	0.03
E.5.44 Earnings conservatism	534	-0.00	-1.48	0.03
E.5.44 Earnings timeliness	534	0.00	0.10	0.03
E.5.44 Earnings conservatism	534	0.00	0.76	0.02
E.5.44 Earnings timeliness	534	0.00	1.11	0.02
E.5.45 FRM, Pension plan funding rate	534	0.00	0.98	0.02
E.5.45 FRA, Pension plan funding rate	534	-0.00	-1.70*	0.03
E.5.46 Ala, asset liquidity	486	0.00	-0.12	0.04
E.5.46 Alm, asset liquidity	486	0.00	1.69	0.05
E.5.51 Average returns Ra1	534	0.00	7.55***	0.00
E. 5.51 Average returns Ra[2,5]	534	0.00	3.60***	0.00
E.5.51 Average returns Ra[6,10]	534	0.00	3.55***	0.00
E.5.51 Average returns Rn1	534	0.00	4.95***	0.01
E. 5.51 Average returns Rn[2,5]	534	0.00	3.35***	0.01
E.5.51 Average returns Rn[6,10]	534	0.00	3.14***	0.01
E.5.51 Average returns Rn[16,20]	534	0.00	1.15	0.04
F. Trading frictions				
F.6.1 Me, market equity	534	-0.00	-0.48	0.05
F.6.2 Ivff1, Ivff6, and Ivff12, idiosyncratic volatility per the Fama and French (1993) 3-factor model	534	-0.01	-2.54**	0.09
F.6.3 Iv, idiosyncratic volatility	534	-0.01	-3.41***	0.08
F.6.5 Ivq1, Ivq6, and Ivq12, idiosyncratic volatility	534	-0.01	-3.34***	0.08
F.6.6 Tv1, Tv6, and Tv12, total volatility	534	-0.01	-3.43***	0.09
F.6.8 beta_1, beta_6, and beta_12, market beta	534	0.00	-0.12	0.08
F.6.9 beta_FP1, beta_FP6, and beta_FP12, the Frazzini-Pedersen beta	534	-0.01	-1.53	0.10
F.6.10 beta_D1, beta_D6, and beta_D12, the Dimson beta	533	-0.00	-0.51	0.06
F.6.11 Tur1, Tur6, and Tur12, share turnover	534	-0.00	-0.82	0.06
F.6.12 Cvt1, Cvt6, and Cvt12, coefficient of variation of share turnover	533	0.00	-0.11	0.03

F.6.13 Dtv1, Dtv6, and Dtv12, dollar trading volume	533	-0.00	-0.60	0.03
F.6.14 Cvd1, Cvd6, and Cvd12, coefficient of variation of dollar trading volume.	533	0.00	0.37	0.03
F.6.15 Pps1, Pps6, and Pps12, share price	534	0.00	0.13	0.08
F.6.16 Ami1, Ami6, and Ami12, absolute return-to-volume	533	-0.00	-0.40	0.05
F.6.17 Lm11, Lm16, Lm112, turnover-adjusted number of zero daily volume	533	0.00	-0.01	0.06
F.6.17. Lm121, Lm126, Lm1212, turnover-adjusted number of zero daily volume	533	0.00	0.70	0.06
F.6.17, Lm61, Lm66, Lm612, turnover-adjusted number of zero daily volume	533	0.00	0.71	0.06
F.6.18 Mdr1, Mdr6, and Mdr12, maximum daily return	534	-0.01	-2.59**	0.07
F.6.20 Isc1, Isc6, and Isc12, idiosyncratic skewness per the CAPM	534	0.00	2.25**	0.03
F.6.21 Isff1, Isff6, and Isff12, idiosyncratic skewness per the Fama and French	534	0.00	2.76***	0.03
F.6.23 Cs1, Cs6, and Cs12, coskewness	534	-0.00	-0.81	0.03
F.6.25 beta_lcc1, beta_lcc6, beta_lcc12, liquidity betas illiquidity-illiquidity	533	0.03	9.20***	0.06
F.6.25 beta_lcr1, beta_lcr6, beta_lcr12, liquidity betas (illiquidity-return)	533	0.00	0.44	0.04
F.6.25 beta_lrc1, beta_lrc6, beta_lrc12, liquidity betas return illiquidity	533	-0.00	-1.63	0.05
F.6.25 beta_net1, beta_net6, and beta_net12, liquidity betas (net)	533	0.01	1.86*	0.08
F.6.25 beta_ret1, beta_ret6, and beta_ret12, liquidity betas (return-return)	533	0.01	1.89*	0.08
F.6.26 Short-term reversal	533	0.00	1.31	0.05
F.6.27 beta_-1, beta_-6, and beta_-12, downside beta	533	-0.00	-0.63	0.07
F.6.31 beta_PS1, beta_PS6, and beta_PS12, the Pastor-Stambaugh beta	534	0.00	0.30	0.04

D.2 2190 Own Constructed Nonlinear Portfolios

We construct 2190 portfolios based on the nonlinear functions of nine characteristics that Freyberger, Neuhierl, and Weber (2020) find significantly explain cross-sectional stock returns. We call them nonlinear portfolios or factors. We apply the following procedure to construct these portfolios. As an indication, we use three characteristics (X1, X2, and X3) as an example.

[1] We generate the following characteristics up to polynomials of degree 3, including X1, X2, X3, X1X2, X1X3, X2X3, X1X2X3, X1X1X3, X1X1X2, X1X2X2, X2X2X3, X1X3X3, X2X3X3, $X1^2$, $X2^2$, $X3^2$, $X1^3$, $X2^3$, $X3^3$.

We alleviate multicollinearity concerns among these characteristics by orthogonalizing each characteristic using a residual from regressing characteristics on its linear and nonlinear components. For instance, we use the residual of regressing X1X2X2 on X1, X2, X1X2, and X2X2 instead of using X1X2X2 directly, or the residual of regressing X1X2 on X1 and X2, instead of using X1X2. Generally, we use X^3-C1*X-C2*X^2 where C1 and C2 are estimated coefficients from regressing X^3 on X and X^2, respectively to eliminate the impact of both X and X^2 from X^3, or X^2-C2*X1, which is a residual from regressing X^2 on X1.

There are two benefits of using residuals. First, the residual methods can remove all the possible correlations between X and X^2. Second, if X's have a different sign (for instance, X1=-2, X2 = 0, and X3=1), $X1 < X2 < X3$ but $X2^2 < X3^2 < X1^2$ and $X1^3 < X2^3 < X3^3$. The relation is not monotonic if X1 to Xn have different signs. Using residuals will take care of this regardless of the sign.

[2] We standardize characteristics firm by firm each time to avoid look-ahead bias and prevent the mis-ranking issue.

[3] We sort stocks into deciles based on the transformed characteristics above, calculate the average returns next period by group, and assign them to the corresponding characteristics and decile (for example, X1_1).

[4] We create long-short portfolios, i.e., ten minus one for each transformed characteristic.

[5] We use 360 ML-based nonlinear portfolios developed by Bryzgalova, Pelger, and Zhu

(2020) (see Appendix D.3 for the list and description).[10] These portfolios can capture the higher dimensional nonlinear information from characteristics.

Table D.2
2190 Own Constructed Nonlinear Portfolios

This table presents nine characteristics selected by Freyberger, Neuhierl, and Weber (2020) and 2190 nonlinear characteristic-sorted decile portfolios constructed based on 219 characteristics. See Section D.2 for the detailed construction. The nine characteristics are *agr* defined as annual percent change in total assets from Cooper and Priestley (2009), *chcsho* or annual percent change in shares outstanding from Pontiff and Woodgate (2008), *mom1m* defined as 1-month cumulative return from Jegadeesh and Titman (1993), *mom12m* defined as 11-month cumulative returns ending one month before month end from Jegadeesh (1990), *mom36m* defined as cumulative returns from months t-36 to t-13, *operprof* or revenue minus cost of goods sold, SG&A expense, and interest expense divided by lagged common shareholders' equity (Fama and French, 2015), *mve* or natural log of market capitalization at the end of month t-1 from Banz (1981), *retvol* or standard deviation of daily returns from month t-1 from Ang et al. (2006), and *turn* or the average monthly trading volume for most recent 3 months scaled by number of shares outstanding in current month from Datar, Naik, and Radcliffe (1998).

[10]We thank Marcus Pelger for generously providing us with the portfolio data.

Portfolio name	ID	Portfolio name	ID	Portfolio name	ID	Portfolio name	ID	Portfolio name	ID
agr_agr	X8	agr_mom12m_retvol	X52	mom12m_retvol_retvol	X97	chcsho_operprof_turn	X143	chcsho_mom36m_operprof	X189
agr_chcsho	X9	agr_mom12m_turn	X53	mom12m_retvol_turn	X98	mom12m_mom12m_operprof	X144	chcsho_mom36m_retvol	X190
agr_mom12m	X10	agr_mom1m_mom1m	X54	mom12m_turn_turn	X99	mom12m_mom1m_operprof	X145	chcsho_mom36m_turn	X191
agr_mom1m	X11	agr_mom1m_mve	X55	mom1m_mom1m_mom1m	X100	mom12m_mve_operprof	X146	mom2m_mom2m_mom36m	X192
agr_mve	X12	agr_mom1m_retvol	X56	mom1m_mom1m_mve	X101	mom12m_operprof_operprof	X147	mom12m_mom1m_mom36m	X193
agr_retvol	X13	agr_mom1m_turn	X57	mom1m_mom1m_retvol	X102	mom12m_operprof_retvol	X148	mom12m_mom36m_mom36m	X194
agr_turn	X14	agr_mve_mve	X58	mom1m_mom1m_turn	X103	mom12m_operprof_turn	X149	mom12m_mom36m_mve	X195
chcsho_chcsho	X15	agr_mve_retvol	X59	mom1m_mve_mve	X104	mom1m_mom1m_operprof	X150	mom12m_mom36m_operprof	X196
chcsho_mom12m	X16	agr_mve_turn	X60	mom1m_mve_retvol	X105	mom1m_mve_operprof	X151	mom12m_mom36m_retvol	X197
chcsho_mom1m	X17	agr_retvol_retvol	X61	mom1m_mve_turn	X106	mom1m_operprof_operprof	X152	mom12m_mom36m_turn	X198
chcsho_mve	X18	agr_retvol_turn	X62	mom1m_retvol_retvol	X107	mom1m_operprof_retvol	X153	mom1m_mom1m_mom36m	X199
chcsho_retvol	X19	agr_turn_turn	X63	mom1m_retvol_turn	X108	mom1m_operprof_turn	X154	mom1m_mom36m_mom36m	X200
chcsho_turn	X20	chcsho_chcsho_chcsho	X64	mom1m_turn_turn	X109	mve_mve_operprof	X155	mom1m_mom36m_mve	X201
mom12m_mom12m	X21	chcsho_chcsho_mom12m	X65	mve_mve_mve	X110	mve_operprof_operprof	X156	mom1m_mom36m_operprof	X202
mom12m_mom1m	X22	chcsho_chcsho_mom1m	X66	mve_mve_retvol	X111	mve_operprof_retvol	X157	mom1m_mom36m_retvol	X203
mom12m_mve	X23	chcsho_chcsho_mve	X67	mve_mve_turn	X112	mve_operprof_turn	X158	mom1m_mom36m_turn	X204
mom12m_retvol	X24	chcsho_chcsho_retvol	X68	mve_retvol_retvol	X113	operprof_operprof_operprof	X159	mom36m_mom36m_mom36m	X205
mom12m_turn	X25	chcsho_chcsho_turn	X69	mve_retvol_turn	X114	operprof_operprof_retvol	X160	mom36m_mom36m_mve	X206
mom1m_mom1m	X26	chcsho_mom12m_mom12m	X70	mve_turn_turn	X115	operprof_operprof_turn	X161	mom36m_mom36m_operprof	X207
mom1m_mve	X27	chcsho_mom12m_mom1m	X71	retvol_retvol_retvol	X116	operprof_retvol_retvol	X162	mom36m_mom36m_retvol	X208
mom1m_retvol	X28	chcsho_mom12m_mve	X72	retvol_retvol_turn	X117	operprof_retvol_turn	X163	mom36m_mom36m_turn	X209
mom1m_turn	X29	chcsho_mom12m_retvol	X73	retvol_turn_turn	X118	operprof_turn_turn	X164	mom36m_mve_mve	X210
mve_mve	X30	chcsho_mom12m_turn	X74	turn_turn_turn	X119	agr_mom36m	X166	mom36m_mve_operprof	X211
mve_retvol	X31	chcsho_mom1m_mom1m	X75	agr_operprof	X121	chcsho_mom36m	X167	mom36m_mve_retvol	X212
mve_turn	X32	chcsho_mom1m_mve	X76	chcsho_operprof	X122	mom12m_mom36m	X168	mom36m_mve_turn	X213
retvol_retvol	X33	chcsho_mom1m_retvol	X77	mom12m_operprof	X123	mom1m_mom36m	X169	mom36m_operprof_operprof	X214
retvol_turn	X34	chcsho_mom1m_turn	X78	mom1m_operprof	X124	mom36m_mom36m	X170	mom36m_operprof_retvol	X215
turn_turn	X35	chcsho_mve_mve	X79	mve_operprof	X125	mom36m_mve	X171	mom36m_operprof_turn	X216
agr_agr_agr	X36	chcsho_mve_retvol	X80	operprof_operprof	X126	mom36m_operprof	X172	mom36m_retvol_retvol	X217
agr_agr_chcsho	X37	chcsho_mve_turn	X81	operprof_retvol	X127	mom36m_retvol	X173	mom36m_retvol_turn	X218
agr_agr_mom12m	X38	chcsho_retvol_retvol	X82	operprof_turn	X128	mom36m_turn	X174	mom36m_turn_turn	X219
agr_agr_mom1m	X39	chcsho_retvol_turn	X83	agr_agr_operprof	X129	agr_agr_mom36m	X175		
agr_agr_mve	X40	chcsho_turn_turn	X84	agr_chcsho_operprof	X130	agr_chcsho_mom36m	X176		
agr_agr_retvol	X41	mom12m_mom12m_mom12m	X85	agr_mom12m_operprof	X131	agr_mom12m_mom36m	X177		
agr_agr_turn	X42	mom12m_mom12m_mom1m	X86	agr_mom1m_operprof	X132	agr_mom1m_mom36m	X178		
agr_chcsho_chcsho	X43	mom12m_mom12m_mve	X87	agr_mve_operprof	X133	agr_mom36m_mom36m	X179		
agr_chcsho_mom12m	X44	mom12m_mom12m_retvol	X88	agr_operprof_operprof	X134	agr_mom36m_mve	X180		
agr_chcsho_mom1m	X45	mom12m_mom12m_turn	X89	agr_operprof_retvol	X135	agr_mom36m_operprof	X181		
agr_chcsho_mve	X46	mom12m_mom1m_mom1m	X90	agr_operprof_turn	X136	agr_mom36m_retvol	X182		
agr_chcsho_retvol	X47	mom12m_mom1m_mve	X91	chcsho_chcsho_operprof	X137	agr_mom36m_turn	X183		
agr_chcsho_turn	X48	mom12m_mom1m_retvol	X92	chcsho_mom12m_operprof	X138	chcsho_chcsho_mom36m	X184		
agr_mom12m_mom12m	X49	mom12m_mom1m_turn	X93	chcsho_mom1m_operprof	X139	chcsho_mom12m_mom36m	X185		
agr_mom12m_mom1m	X50	mom12m_mve_mve	X94	chcsho_mve_operprof	X140	chcsho_mom1m_mom36m	X186		
agr_mom12m_mve	X51	mom12m_mve_retvol	X95	chcsho_operprof_operprof	X141	chcsho_mom36m_mom36m	X187		
		mom12m_mve_turn	X96	chcsho_operprof_retvol	X142	chcsho_mom36m_mve	X188		

D.3 360 ML-based Nonlinear Portfolios

This subsection shows 360 ML-based nonlinear portfolios for each characteristic group containing ten decile portfolios. See Bryzgalova, Pelger, and Zhu (2020) for the detailed construction and the following tables for variable descriptions.

Table D.3
360 ML-based Nonlinear Characteristics Groups

LME_AC_IdioVol
LME_AC_Lturnover
LME_BEME_AC
LME_BEME_IdioVol
LME_BEME_Investment
LME_BEME_LT_Rev
LME_BEME_Lturnover
LME_BEME_OP
LME_BEME_r12_2
LME_BEME_ST_Rev
LME_IdioVol_Lturnover
LME_Investment_AC
LME_Investment_Idiovol
LME_Investment_LT_Rev
LME_investment_Lturnover
LME_Investment_ST_Rev
LME_LT_Rev_AC
LME_LT_Rev_IdioVol
LME_LT_Rev_Lturnover
LME_OP_AC
LME_OP_IdioVol
LME_OP_Investment
LME_OP_LT_Rev
LME_OP_Lturnover
LME_OP_ST_Rev
LME_r12_2_AC
LME_r12_2_IdioVol
LME_r12_2_Investment
LME_r12_2_LT_Rev
LME_r12_2_Lturnover
LME_r12_2_OP
LME_r12_2_ST_Rev
LME_ST_REV_AC
LME_ST_Rev_IdioVol
LME_ST_Rev_LT_Rev

Table D.4
10 ML-Based Characteristics

Symbol	Names	Description	References
AC	Accrual	Change in operating working capital per split-adjusted share from the scal year	Sloan (1996)
BEME	Book-to-Market ratio	Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholders equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, we use the redemption (item PSTKRV), liquidating (item PSTKL), or per value (item PSTK) for PS. The market value of equity (PRC*SHROUT) is as of December t-1.	Basu (1983), Fama and French (1992)
IdioVol	Idiosyncratic volatility	Standard deviation of the residuals from a regression of excess returns on the Fama and French three-factor model	Ang et al. (2006)
Investment	Investment	Change in total assets (AT) from the fiscal year ending in year t-2 to the fiscal year ending in t-1, divided by t-2 total assets	Fama and French (2015)
LME	Size	Total market capitalization at the end of the previous month defined as price times shares outstanding	Banz (1981), Fama and French (1992)
LT_Rev	Long-term reversal	Cumulative return from 60 months before the return prediction to 13 months before	De Bondt and Thaler (1985)
LTurnover	Turnover	Last month's volume (VOL) over shares outstanding (SHROUT)	Datar, Naik, and Radcliffe (1998)
OP	Operating profitability	Annual revenues (REVT) minus cost of goods sold (COGS), interest expense (TIE), and selling, general, and administrative expenses (XSGA) divided by book equity (defined in BEME)	Fama and French (2015)
r12_2	Momentum	Return for the first 12 months except for the first month	Jegadeesh (1990)
ST_Rev	Short-term reversal	Prior month return	Jegadeesh (1990)