Journal of
Risk and Financial Management

MDPI

Article

Testing and Ranking of Asset Pricing Models Using the GRS Statistic

Mark J. Kamstra [1,*,†] **and Ruoyao Shi** [2,†]

[1] Schulich School of Business, Room N204-C, York University, 4700 Keele St., Toronto, ON M3J 1P3, Canada
[2] Department of Economics, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA; ruoyao.shi@ucr.edu
* Correspondence: mkamstra@yorku.ca; Tel.: +1-(416)-736-2100 (ext. 33302)
† These authors contributed equally to this work.

Abstract: We clear up an ambiguity in the statement of the GRS statistic by providing the correct formula of the GRS statistic and the first proof of its F-distribution in the general multiple-factor case. Casual generalization of the Sharpe-ratio-based interpretation of the single-factor GRS statistic to the multiple-portfolio case makes experts in asset pricing studies susceptible to an incorrect formula. We illustrate the consequences of using the incorrect formulas that the ambiguity in GRS leads to—over-rejecting and misranking asset pricing models. In addition, we suggest a new approach to ranking models using the GRS statistic *p*-value.

Keywords: GRS; asset pricing; CAPM; multivariate test; portfolio efficiency; Sharpe ratio; over-rejection; model ranking

JEL Classification: G12; G14

Citation: Kamstra, Mark J., and Ruoyao Shi. 2024. Testing and Ranking of Asset Pricing Models Using the GRS Statistic. *Journal of Risk and Financial Management* 17: 168. https://doi.org/10.3390/jrfm17040168

Academic Editor: Thanasis Stengos

Received: 21 March 2024
Revised: 4 April 2024
Accepted: 8 April 2024
Published: 19 April 2024

1. Introduction

In an influential paper, Gibbons et al. (1989) developed and analyzed a test of the ex ante mean-variance efficiency of portfolios. This test statistic is now widely used to evaluate asset pricing models and has also been exploited to rank competing models. For the single factor case, Gibbons et al. (1989) carefully developed the statistic in a linear regression model (hereafter referred to as the GRS statistic or test), derived its small-sample F distribution, investigated its power properties, and highlighted its significance in asset pricing theory by purveying an alternative interpretation involving the Sharpe ratio (Sharpe 1966)—the excess return to a portfolio per unit of risk (or volatility, measured by standard deviation)—which is a key measure of portfolio efficiency. For the multiple factor case, however, Gibbons et al. (1989, sec. 7), were ambiguous on how the statistic should be constructed.

The solution to the portfolio optimization problem that yields the Sharpe ratio has us estimate a variance–covariance matrix of the portfolio excess returns, but the equivalence of the GRS statistic and the F test statistic relies on this matrix arising in the projection of the test asset returns on the column space of the asset pricing factors, not as a variance–covariance matrix. Unfortunately, Gibbons et al. (1989) used equivocal language to describe this matrix, referring to it as a "variance-covariance matrix", and this has apparently caused confusion about the function of the GRS statistic, which is further exacerbated by the fact that the small-sample F distribution wrongfully conjures up a degrees-of-freedom (d.f. hereafter) adjustment that is improper in this case. This has led to the application of a very common incorrect formula that, paradoxically, is more likely to be used by financial economists, the experts in the field, than by someone who focuses only on the statistical aspects of the problem.[1] We find that using the incorrect formula, which we will refer to for conciseness as $\widehat{W}$ below, leads to (i) a test statistic that does not follow the F distribution

as prescribed and over-rejects the null hypothesis of portfolio efficiency; and (ii) smaller models often being favored over larger ones when the statistic is used to rank asset pricing models. This error comes from mixing terms that fall out of portfolio optimization with a statistical object that comes from the small-sample F test derivation.

The main contribution of our paper is to clear up the ambiguity in the calculation of the GRS statistic and highlight (both theoretically and empirically) issues that arise from the use of $\widehat{W}$ and two related and popularly used statistics, one which folds in a second degree of freedom error (we will refer to this as $\check{W}$), and the asymptotic χ^2 version often used to replace the GRS test.[2] The asymptotic χ^2 and $\check{W}$ implementations result in much higher model rejection rates than the correct GRS statistic, most notably for the asymptotic χ^2 test, even with 50 years of monthly data. The use of an incorrect implementation of the GRS statistic also results in inconsistent model rankings across $\widehat{W}$, $\check{W}$ and the correct calculation of the GRS statistic, with 40 or even 50 years of data. Finally, we propose a new methodology for the ranking of competing asset pricing models, making use of test p-values rather than the raw GRS statistic values, meant to properly internalize the model sizes. While a determination of statistically significantly different model performance is valuable, often researchers are simply attempting to rank models. Our approach is a computationally straightforward approach to answering this question.

We will adopt the notation in Gibbons et al. (1989, sec. 7), whenever possible. The proofs of the theoretical results and the details of the empirical results are in the Appendix A.

2. The GRS Test for Multiple Factors

The proofs of all the claims in this section can be found in Appendix A. The problem is to test the mean-variance efficiency of L portfolios utilizing another type of N assets (known as test assets).

We start with a linear regression model:

$$\tilde{r}_{it} = \delta_{i0} + \delta_i' \tilde{r}_{pt} + \tilde{\eta}_{it}, \quad \forall i = 1, \ldots, N, \text{ and } t = 1, \ldots, T, \tag{1}$$

where $\tilde{r}_{it}$ denotes the excess return on test asset i in period t, the L-vector of portfolio excess returns $\tilde{r}_{pt}$ serves as factors, and $\tilde{\eta}_{it}$ denotes the disturbance. Mean-variance efficiency of the L portfolios implies (Sharpe 1964)

$$H_0 : \delta_{i0} = 0, \quad \forall i = 1, \ldots, N. \tag{2}$$

Lemma 1 (Joint F test). *Let $\tilde{r}_p \equiv \left[\tilde{r}_{p1}, \ldots, \tilde{r}_{pT}\right]'$, $\bar{r}_p \equiv T^{-1} \sum_{t=1}^{T} \tilde{r}_{pt}$, and let $\hat{\delta}_0$ be the ordinary least squares (OLS) estimator of $\delta_0 \equiv (\delta_{10}, \ldots, \delta_{N0})'$; also let $\hat{\eta}_t \equiv (\hat{\eta}_{1t}, \ldots, \hat{\eta}_{Nt})'$ be the OLS residuals of model (1). We follow Gibbons et al. (1989) to assume that the disturbance $\tilde{\eta}_t \equiv (\tilde{\eta}_{1t}, \ldots, \tilde{\eta}_{Nt})'$ is independent from the factors $\tilde{r}_{pt}$ and has a joint normal distribution[3] with mean zero and nonsingular variance–covariance matrix Σ and is iid over t. Define*

$$\widetilde{\Omega}_* \equiv \frac{1}{T} \sum_{t=1}^{T} \tilde{r}_{pt} \tilde{r}_{pt}', \tag{3}$$

$$\widehat{\Sigma} \equiv \frac{1}{T - L - 1} \sum_{t=1}^{T} \hat{\eta}_t \hat{\eta}_t'. \tag{4}$$

Then, the F statistic

$$\widetilde{W}_* \equiv \frac{T(T - N - L)}{N(T - L - 1)} \left(1 - \bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p\right) \hat{\delta}_0' \widehat{\Sigma}^{-1} \hat{\delta}_0 \tag{5}$$

follows the $F_{N,T-N-L}$ distribution under H_0.

From a purely statistical perspective, Lemma 1 is all we need for testing the implication (2) of mean-variance efficiency, which is just the usual joint F test of zero intercepts in a linear regression system.[4]

The economic interpretation of the GRS test, however, is better understood via another implication of mean-variance efficiency—θ^*_{N+L}, the Sharpe ratio of the optimal portfolio consisting of the L portfolios and the N test assets, equals θ^*_p, the Sharpe ratio of the L portfolios alone (Gibbons et al. 1989).

We consider a general portfolio optimization problem that yields the Sharpe ratio. Let $\tilde{r}$ denote a vector of excess returns of K assets ($K \geq 1$), and let $\mu_{\tilde{r}}$ and $\Omega_{\tilde{r}}$ be their ex ante mean vector and variance–covariance matrix, respectively. Let m be the target mean excess return and ω be a vector of K asset weights. The optimal portfolio weights ω^* solve

$$\min_{\omega} \omega' \Omega_{\tilde{r}} \omega, \quad \text{subject to} \quad \omega' \mu_{\tilde{r}} = m.$$

The square of the Sharpe ratio of the optimal portfolio composed of these K assets, therefore, is

$$\theta^{*2} \equiv \left(\frac{m}{\sqrt{\omega^{*\prime} \Omega_{\tilde{r}} \omega^*}} \right)^2 = \mu'_{\tilde{r}} \Omega_{\tilde{r}}^{-1} \mu_{\tilde{r}},$$

in which the variance–covariance matrix $\Omega_{\tilde{r}}$ of the K assets plays a central role. Applying this general result twice, we obtain that

$$W \equiv \left(\frac{\sqrt{1 + \theta^{*2}_{N+L}}}{\sqrt{1 + \theta^{*2}_p}} \right)^2 - 1 = \left(1 + \mu'_{\tilde{r}_p} \Omega^{-1} \mu_{\tilde{r}_p} \right)^{-1} \delta'_0 \Sigma^{-1} \delta_0, \tag{6}$$

where Ω is the variance–covariance matrix of L portfolio excess returns $\tilde{r}_{pt}$ and Σ is that of the disturbances $\tilde{\eta}_t$. So, $W = 0$ if the L portfolios are efficient, and this is the basis of the GRS test in asset pricing theory.

Theorem 1 (Generalized GRS statistic)**.** *Define*

$$\widetilde{\Omega} \equiv \frac{1}{T} \sum_{t=1}^{T} \left(\tilde{r}_{pt} - \bar{r}_p \right) \left(\tilde{r}_{pt} - \bar{r}_p \right)' = \frac{1}{T} \sum_{t=1}^{T} \tilde{r}_{pt} \tilde{r}'_{pt} - \bar{r}_p \bar{r}'_p \tag{7}$$

and the generalized GRS statistic

$$\widetilde{W} \equiv \frac{T(T-N-L)}{N(T-L-1)} \left(1 + \bar{r}'_p \widetilde{\Omega}^{-1} \bar{r}_p \right)^{-1} \hat{\delta}'_0 \widehat{\Sigma}^{-1} \hat{\delta}_0. \tag{8}$$

Then, $\widetilde{W} = \widetilde{W}_*$, and therefore under the conditions of Lemma 1, $\widetilde{W}$ follows the $F_{N,T-N-L}$ distribution under the H_0.

Theorem 1 connects the statistical perspective to the economic interpretation of the GRS test, because $\widetilde{W}$ equals the F statistic $\widetilde{W}_*$ and can be regarded as a sample analog of W—replace Ω in Equation (6) with its maximum likelihood estimator (MLE) $\widetilde{\Omega}$, Σ with its unbiased estimator $\widehat{\Sigma}$, $\mu_{\tilde{r}_p}$ with $\bar{r}_p$, δ_0 with $\hat{\delta}_0$, and pre-multiply the ratio $\frac{T(T-N-L)}{N(T-L-1)}$, then one obtains $\widetilde{W}$ in Equation (8).

Common Mistakes and Consequences

$\widetilde{W}$ equals the original GRS statistic when $L = 1$. For the $L > 1$ case, however, Gibbons et al. (1989, p. 1146) gave a statistic $\widehat{W}$, almost identical to $\widetilde{W}$, but instead of $\widetilde{\Omega}$, they prescribed "sample variance-covariance matrix" $\widehat{\Omega}$ without giving its explicit formula. Since the sample variance–covariance matrix customarily entails a d.f. adjustment, i.e.,

$$\widehat{\Omega} \equiv \frac{1}{T-1} \sum_{t=1}^{T} \left(\tilde{r}_{pt} - \bar{r}_p \right) \left(\tilde{r}_{pt} - \bar{r}_p \right)' = \frac{T}{T-1} \widetilde{\Omega}, \tag{9}$$

this would cause $\widehat{W}$ to differ from $\widetilde{W}$, and therefore Theorem 1 implies that $\widehat{W}$ does not follow the $F_{N,T-N-L}$ distribution as prescribed.

This incorrect GRS statistic $\widehat{W}$ inflicts two consequences on empirical asset pricing studies. First, it over-rejects mean-variance efficiency when gauged against the $F_{N,T-N-L}$ distribution, because the ratio between $\widehat{W}$ and $\widetilde{W}$ is always larger than 1. Second, it misranks competing asset pricing models, because the ratio between $\widehat{W}$ and $\widetilde{W}$ tends to be disproportionally larger for models with more factors.

The significance of the GRS statistic in recent financial studies, as Fama and French (2015) advocate, resides in the ranking of competing asset pricing models, rather than testing them. Using even the correct GRS statistics to rank models, albeit its portfolio optimization interpretation, is subject to a familiar critique akin to the use of R^2 for linear regression model comparison. Instead, the p-values associated with $\widetilde{W}$ in respective $F_{N,T-N-L}$ distributions are a statistically sound metric for this purpose, as they internalize the difference in the second d.f.

The implementations of the GRS test found in popular user-defined software packages, such as `GRS.test` in R and `grstest` and `grstest2` in Stata, not only use $\widehat{\Omega}$ when computing the GRS statistic, but also fold in additional errors. Results for the R formula are labeled with $\check{W}$ in the rest of this paper.

The asymptotic χ^2 test is frequently recommended as an alternative of the F test. One might think that when T is large, the d.f. issue we point out here can be circumvented by using the asymptotic χ^2 test. Unfortunately, we find that the commonly used χ^2 test statistics also over-reject for any sample size, especially if the number of test assets N or the number of factors L is large. In addition, they erroneously favor smaller models to an extent worse than $\widehat{W}$.

3. Empirical Results

We use portfolios of test assets borrowed from Fama and French (2015, 2016) to show that over-rejection and misranking of $\widehat{W}$, $\check{W}$ and χ^2 relative to $\widetilde{W}$ is empirically significant, even remarkable in many cases. To summarize our empirical findings (which are detailed in Appendix B and below), the $\check{W}$ misapplication of the GRS statistic and the asymptotic χ^2 both result in much higher model rejection rates than the correct GRS statistic, most notably for the asymptotic χ^2 test, even with 50 years of monthly data, and also result in scrambled model rankings, with 40 or even 50 years of data, most notably for the $\check{W}$ version of the GRS statistic. While the F test is asymptotically equivalent to χ^2, typical sample sizes available in financial markets research are not large enough to make this approximation innocuous. The exact F test construction is also the most conservative test, resulting in less over-rejection of the null hypothesis when the null is correct, even with highly non-normal return data, by measure of the bootstrap resampling experiments we perform.

In Table 1, we highlight the over-rejection issue, with five-year windows. This span of data shows serious over-rejection of asset pricing models from the application of the alternative formulations of the GRS test. Results for the largest number of test assets we considered, 32, are displayed in the first three rows of the table, the results for cases with 25 test assets follow in rows four through thirteen, the 17 test assets of the industry portfolios follow in row fourteen, and the remaining six rows present results for sets of 10 test assets. In Table 1, we use a total of 53 five-year overlapping windows starting from 1963 and consider six different asset pricing models, meaning that we have 318 cases for which an asset pricing model might be rejected, for each of the 19 sets of test assets.

The asymptotic χ^2 statistic fares the worst relative to the correct GRS statistic among the alternatives; it over-rejects dramatically relative to the correct GRS statistic, 50% of the time on average, faring worse when we consider more test assets. $\check{W}$ and $\widehat{W}$ over-reject relative to the correct GRS statistic about 10% and 1% of the time on average. $\widehat{W}$ is unstable across test assets, however, with a few cases displaying close to 4% over-rejection, and some with no over-rejection.

Table 1. Number and proportion of cases with more rejections relative to the GRS test for five-year windows sampled during 1963–2019, across 19 sets of test assets.

Test Assets **Nb of Subsamples = 53**	**Asymptotic** χ^2			$\check{W}$			$\widehat{W}$		
	1%	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
2 × 4 × 4 MExMEBExINV	289	275	247	4	17	21	0	1	0
2 × 4 × 4 MExMEBExOP	262	246	220	4	14	18	0	0	0
2 × 4 × 4 MExOPxINV	282	238	194	10	27	28	0	3	1
5 × 5 AccrualsxME	222	218	207	7	17	31	0	0	1
5 × 5 BExME	212	191	166	8	21	20	1	0	1
5 × 5 BetaxME	229	221	204	11	16	25	0	1	0
5 × 5 MExOP	238	227	188	9	27	42	0	0	0
5 × 5 MomentumxME	210	166	124	14	29	27	0	0	0
5 × 5 NetIssuexME	216	176	147	9	29	27	0	2	2
5 × 5 RVariancexME	147	94	65	18	28	12	0	2	0
5 × 5 VariancexME	137	81	45	26	27	12	0	3	1
5 × 5 BExInv	248	259	245	4	18	16	0	2	1
5 × 5 MExInv	214	189	171	14	17	29	0	0	0
Industry	126	153	152	12	9	22	0	1	0
Book-to-Market Deciles	48	70	67	5	22	21	0	1	1
Investment Deciles	22	30	56	4	4	6	0	0	0
Momentum Deciles	56	60	66	18	14	20	1	1	1
Size Deciles	49	90	68	6	23	27	0	0	2
Operating Profitability Deciles	48	70	65	4	24	15	0	1	0
Average	171.3	160.7	142	9.8	20.2	22.1	0.11	0.95	0.58
Proportion (%)	53.9	50.5	44.6	3.1	6.3	6.9	0.0	0.3	0.2

Notes: (1) The figures are the number of all decisions at the stated significance level for which the test statistic rejects the model when the correct GRS statistic $\widetilde{W}$ does not reject, out of a total of 318 possible cases, with the exception of the last row in which the proportion is given. The sample periods of five years are sampled from July 1963 to December 2019, and the number of models tested is 6 for each window, which we sum over to obtain the total number of over-rejections. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. This means that there are 53 samples for the 5-year window. (2) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

In Table 2, we consider the model misranking among six models for each of the 53 five-year windows of Table 1. Misranking of models shows similar problems for the χ^2 statistic as we saw for test rejections, with over 40% of the cases displaying misranking of at least one asset pricing model. The $\check{W}$ statistic displays much worse performance than test rejections, with close to 60% of the cases displaying misranking, and we see worse performance even for the $\widehat{W}$, with close to 3% of the cases misranked.

While the typical Fama and French paper uses 40 or 50 years of data, it is also true that much empirical work uses far less data. Gibbons et al. (1989) noted that issues of stationarity can reasonably constrain the length of a time series used, so that "it is not uncommon to see published work where T is around 60", Affleck-Graves and McDonald (1989) limited their analysis and simulations to 60 month periods, Ferson and Foerster (1994) studied 60, 120, and 720 monthly observations in their simulation, Rouwenhorst (1999) used five years in subsample analysis, and among recent works that exploited as little as four or five years of data are Belimam et al. (2018) and Qin (2019). Leite et al. (2018) used as few as 98 months of data, Lewellen et al. (2010) used 168 observations of quarterly data, Choi et al. (2020) performed subsample stability tests using eight years of monthly data, and many studies of emerging economy markets have used 10 to 15 years of monthly data. See, for instance, Alhomaidi et al. (2019), Alshammari and Goto (2022), Merdad et al. (2015), and Sha and Gao (2019).

One takeaway from these papers is that many situations involving specialized data (like (Sha and Gao 2019) and their exploration of mutual fund returns in China) or subsample robustness checks (like Baek and Bilson 2015) are necessarily constrained to shorter samples than fifty or even twenty years, so that the bias from an incorrectly calculated GRS statistic becomes large.

Table 2. Number and proportion of cases with different ranking outcomes from the GRS statistic for five-year windows sampled during 1963–2019, across 19 sets of test assets

	Any Model Mis-Ranked		**Top Model Mis-Ranked**	
Test Assets **Nb of Subsamples = 53**	$\check{W}$	$\widehat{W}$	$\check{W}$	$\widehat{W}$
2 × 4 × 4 MExMEBExINV	26	2	5	0
2 × 4 × 4 MExMEBExOP	37	1	6	0
2 × 4 × 4 MExOPxINV	31	1	12	0
5 × 5 AccrualsxME	36	1	12	0
5 × 5 BExME	26	1	7	0
5 × 5 BetaxME	31	1	10	0
5 × 5 MExOP	33	0	10	0
5 × 5 MomentumxME	33	0	14	0
5 × 5 NetIssuexME	27	4	9	2
5 × 5 RVariancexME	32	0	14	0
5 × 5 VariancexME	34	1	10	0
5 × 5 BExInv	35	3	6	0
5 × 5 MExInv	30	3	9	0
Industry	29	2	4	0
Book-to-Market Deciles	28	0	4	0
Investment Deciles	24	2	4	0
Momentum Deciles	30	2	5	0
Size Deciles	27	3	5	0
Operating Profitability Deciles	30	0	6	0
Average	30.47	1.42	8.00	0.11
Proportion (%)	57.5	2.7	15.1	0.2

Notes: (1) The figures are the number of all misrankings by the test statistic value across models relative to the correct GRS statistic ranking, out of a total of 53 possible cases, with the exception of the last row in which the proportion is given. The sample periods of five years are sampled over July 1963 to December 2019, and the number of models ranked is 6 for each window. These windows overlap, adjusted in a rolling window, so that all but 1 year of data overlaps with the next sample window. This means that there are 53 samples for the 5-year window. (2) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

4. Concluding Remarks

The GRS statistic of Gibbons et al. (1989), developed to provide a test of the ex ante mean-variance efficiency of portfolios and more recently exploited to rank competing models, can be easily implemented incorrectly due to an ambiguity in the presentation of the multivariate form of the test in Gibbons et al. (1989). This presentation suggests a degree-of-freedom-adjusted unbiased variance–covariance matrix estimator $\widehat{\Omega}$ of the portfolio excess returns used in the small-sample GRS F test. Indeed, the portfolio optimization problem naturally has us estimate the variance–covariance matrix Ω of the portfolio excess returns, but the equivalence of the GRS statistic to the F test relies on $\widetilde{\Omega}$, which arises in the projection matrix of the test asset returns on the column space of the asset pricing factors, not as a variance–covariance matrix. Paradoxically, this error is clearly visible when turning a blind eye to the economic interpretation of the GRS statistic and taking a purely statistical approach. Although an unbiased estimator $\widehat{\Omega}$ appears intuitive in the context of portfolio optimization, it does not yield a correct small-sample exact F test.

Further complicating this ambiguity, Cochrane (2005) presented the GRS statistic omitting a degree-of-freedom adjustment in the calculation of the variance–covariance matrix of the regression residuals, Σ.[5] Perhaps an outcome of Cochrane (2005), there is an implementation of the GRS statistic in an R package, which we label $\check{W}$, that omits the d.f. adjustment when estimating Σ but fails to pre-multiply the correct ratio.[6] It has also become common in the field to ignore the F distribution completely and employ an asymptotic χ^2 approximation in place of the F test.

The main results for both the asymptotic asymptotic χ^2 and $\check{W}$ implementations is much higher model rejection rates than the correct GRS statistic, most notably for the asymptotic χ^2 test, and they also result in scrambled model rankings. Further, the F distribution is inherently pertinent to small-sample exact tests, where one should make a point of computing the d.f. correctly. For this reason, we recommend the exact F test construction with its attendant F distribution, for both testing and ranking of asset pricing models. The exact F test construction is also the most conservative test, resulting in less over-rejection of the null hypothesis when the null is correct, even with highly non-normal return data.

Another result of this research inquiry is that we provided the first proof of the F-distribution of this test for the general multi-factor case and we recommended a new ranking method, making use of the p-value rather than the raw GRS statistic value. Although ranking by the values of the GRS statistic has a desirable economic intuition attached to it, the applied researcher taking advantage of this must recognize that this ranking is statistically as unsound as favoring a regression model with the highest R^2.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jrfm17040168/s1.

Author Contributions: Conceptualization, methodology, review, editing, software, validation, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, funding acquisition were all shared equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Social Sciences and Humanities Research Council of Canada, grant number 510991.

Data Availability Statement: Data used is available through Ken French's data library.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs and Details for the Results in Section 2

The following two lemmas are used in the proof of Lemma 1.

Lemma A1. *If a random vector Y and a random matrix W satisfy: (i) $Y \sim \mathcal{N}_d(\mu, \Sigma)$, the d dimensional normal distribution; (ii) $W \sim \mathcal{W}_d(f, \Sigma)$, the $d \times d$ dimensional Wishart distribution; and (iii) $Y \perp W$. Then, given Hotelling's T-squared defined as $T^2 \equiv f(Y-\mu)'W^{-1}(Y-\mu)$, we have $F \equiv \frac{f-d+1}{fd} T^2 \sim F_{d,f-d+1}$.*

Lemma A2 (Sherman–Morrison formula)**.** *Suppose A is an invertible $L \times L$ matrix and u and v are $L \times 1$ vectors. If $A + uv'$ is invertible, then $(A+uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}$.*

Lemma A1 is a standard result in multivariate statistics (see, e.g., Anderson 2003, Theorem 5.2.2), and Lemma A2 is a standard result in linear algebra (see, e.g., Bartlett 1951, p. 107).

Proof of Lemma 1. The proof proceeds in three steps.

Step 1. In this step, we show that under the null hypothesis (2),

$$\sqrt{T(1-\bar{r}_p'\widetilde{\Omega}_*^{-1}\bar{r}_p)}\hat{\delta}_0 \sim \mathcal{N}_N(0,\Sigma), \tag{A1}$$

where $\widetilde{\Omega}_*$ is defined in Equation (3).

Let ℓ_T denote a $T\times 1$ vector with every element being one, and let I_T denote the $T\times T$ identity matrix. Define $P_{p,T}=\tilde{r}_p\left(\tilde{r}_p'\tilde{r}_p\right)^{-1}\tilde{r}_p'$ as the $T\times T$ projection matrix (onto the column space of $\tilde{r}_p$) and its $T\times T$ complement matrix $Q_{p,T}=I_T-P_{p,T}$. It is a standard result (e.g., Hayashi 2000, pp. 18–19) that the OLS estimator of δ_{i0} satisfies $\hat{\delta}_{i0}-\delta_{i0}=\left(\ell_T'Q_{p,T}\ell_T\right)^{-1}\ell_T'Q_{p,T}\tilde{\eta}_i$, where $\tilde{\eta}_i\equiv(\tilde{\eta}_{i1},\ldots,\tilde{\eta}_{iT})'$ for $\forall i=1,\ldots,N$. Since $\tilde{\eta}_{it}$ has a normal distribution, and let σ_{ii}^2 denote the (i,i) entry of Σ, then it is a standard result (e.g., Hayashi (2000, Sec. 1.3) that $\sqrt{\ell_T'Q_{p,T}\ell_T}(\hat{\delta}_{i0}-\delta_{i0})\sim\mathcal{N}_1(0,\sigma_{ii}^2)$. It then only takes some algebra to show that

$$\sqrt{\ell_T'Q_{p,T}\ell_T}(\hat{\delta}_0-\delta_0)\sim\mathcal{N}_N(0,\Sigma). \tag{A2}$$

Now, let us take a closer look at $\ell_T'Q_{p,T}\ell_T$:

$$\begin{aligned}
\ell_T'Q_{p,T}\ell_T &= \ell_T'\ell_T-\ell_T'\tilde{r}_p\left(\tilde{r}_p'\tilde{r}_p\right)^{-1}\tilde{r}_p'\ell_T\\
&= T-\left(\sum_{t=1}^{T}\tilde{r}_{pt}'\right)\left(\sum_{t=1}^{T}\tilde{r}_{pt}\tilde{r}_{pt}'\right)^{-1}\left(\sum_{t=1}^{T}\tilde{r}_{pt}\right)\\
&= T-T\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{r}_{pt}'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{r}_{pt}\tilde{r}_{pt}'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{r}_{pt}\right)\\
&= T(1-\bar{r}_p'\widetilde{\Omega}_*^{-1}\bar{r}_p).
\end{aligned} \tag{A3}$$

Recall that $\delta_0=0$ under the null hypothesis (2), so Equation (A2) and (A3) together imply (A1), the claim of Step 1.

Step 2. In this step, we will show that $\hat{\delta}_0\perp\widehat{\Sigma}$ and

$$(T-L-1)\widehat{\Sigma}\sim\mathcal{W}_N(T-L-1,\Sigma). \tag{A4}$$

Let $X=[\ell_T,\tilde{r}_p]$ denote the $T\times(L+1)$ design matrix of Equation (1). Define the $T\times T$ projection matrix $P=X(X'X)^{-1}X'$ and its complement $Q=I_T-P$. Let $\tilde{\eta}=[\tilde{\eta}_1,\ldots,\tilde{\eta}_N]$ denote the $T\times N$ matrix of all disturbances in Equation (1). Then, by the standard results of the OLS estimators with normal disturbances (e.g., Hayashi (2000, Sec. 1.3), we have $\hat{\delta}_0\perp\widehat{\Sigma}$ and $(T-L-1)\widehat{\Sigma}=\sum_{t=1}^{T}\hat{\eta}_t\hat{\eta}_t'=\tilde{\eta}'Q\tilde{\eta}=\tilde{\eta}'UDU'\tilde{\eta}$, where the last equality holds by the singular value decomposition of Q, in which U is a $T\times T$ unitary matrix, and D is a $T\times T$ diagonal matrix with $T-L-1$ diagonal entries being ones and the rest being zeros. Since we assume that the rows of $\tilde{\eta}$ are mutually independent and follow the $\mathcal{N}_N(0,\Sigma)$ distribution, the rows of $U'\tilde{\eta}$ are also mutually independent and follow the $\mathcal{N}_N(0,\Sigma)$ distribution. This further implies that $\tilde{\eta}'UDU'\tilde{\eta}$ has the same distribution as such sum $S=\sum_{j=1}^{T-L-1}\xi_j\xi_j'$, where ξ_j are mutually independent and $\xi_j\sim\mathcal{N}_N(0,\Sigma)$ $(j=1,\ldots,T-L-1)$. By construction, the distribution of S is the Wishart distribution $\mathcal{W}_N(T-L-1,\Sigma)$. This proves the claim of Step 2.

Step 3. In this step, we apply Lemma A1 to the results of Steps 1 and 2. After some simple algebra, we obtain $\widetilde{W}_*\sim F_{N,T-N-L}$ with $\widetilde{W}_*$ defined in Equation (5). This completes the proof of Lemma 1. □

Derivation of Equation (6). Gibbons et al. (1989) derives this, in Section 6, for the $L = 1$ case, and here we provide the derivation for the general $L \geq 1$ case. We start by considering a general portfolio optimization problem that yields the Sharpe ratio—mean excess return to a portfolio per unit of volatility (standard deviation)—of the optimal portfolio consisting of given assets. Let $\tilde{r}$ denote a vector of excess returns of K assets ($K \geq 1$), and let $\mu_{\tilde{r}}$ and $\Omega_{\tilde{r}}$ be their ex ante mean vector and variance–covariance matrix, respectively. Let m be the target mean excess return and ω be a vector of K asset weights. The optimal portfolio weights ω^* solve

$$\min_{\omega} \omega' \Omega_{\tilde{r}} \omega, \quad \text{subject to} \quad \omega' \mu_{\tilde{r}} = m.$$

The first order conditions for this problems are $\omega^* = \varphi \Omega_{\tilde{r}}^{-1} \mu_{\tilde{r}}$ and $\varphi = m/(\mu_{\tilde{r}}' \Omega_{\tilde{r}}^{-1} \mu_{\tilde{r}})$, where φ is the Lagrange multiplier. The squared Sharpe ratio of the optimal portfolio consisting of these K assets is, therefore,

$$\theta^{*2} \equiv \left(\frac{m}{\sqrt{\omega^{*\prime} \Omega_{\tilde{r}} \omega^*}} \right)^2 = \mu_{\tilde{r}}' \Omega_{\tilde{r}}^{-1} \mu_{\tilde{r}},$$

in which the variance–covariance matrix $\Omega_{\tilde{r}}$ plays a central role.

Applying this general result, we know that when the constituent assets are the L portfolios, the squared Sharpe ratio is

$$\theta_p^{*2} = \mu_{\tilde{r}_p}' \Omega^{-1} \mu_{\tilde{r}_p}. \tag{A5}$$

When the constituent assets include both the N test assets and the L portfolios, the squared Sharpe ratio is $\theta_{N+L}^{*2} = \mu_{\tilde{r}_{N+L}}' \Omega_{\tilde{r}_{N+L}}^{-1} \mu_{\tilde{r}_{N+L}}$, where $\mu_{\tilde{r}_{N+L}} \equiv \left(\mu_{\tilde{r}_N}', \mu_{\tilde{r}_p}' \right)'$,

$$\Omega_{\tilde{r}_{N+L}} \equiv \begin{bmatrix} \delta \Omega \delta' + \Sigma & \delta \Omega \\ \Omega \delta' & \Omega \end{bmatrix}, \tag{A6}$$

and $\delta \equiv [\delta_1, \ldots, \delta_N]'$ with δ_i being the slope coefficient in model (1). Equation (A6) holds because we can rewrite the variance–covariance matrix of the N test assets and their covariance matrix with the L portfolios using Ω, Σ and δ (in the same way as $\hat{V}$ on p. 1143 and eq. (24) in Gibbons et al. 1989). Applying the inverse formula for a block matrix and noticing the relationship between $\mu_{\tilde{r}_N}$ and $\mu_{\tilde{r}_p}$ implied by model (1), we obtain

$$\theta_{N+L}^{*2} = \theta_p^{*2} + \delta_0' \Sigma^{-1} \delta_0, \tag{A7}$$

which is essentially the same as Equations (22) and (23) in MacKinlay and Richardson (1991). This, together with Equation (A5) and simple algebra, further implies Equation (6).

Proof of Theorem 1. Based on Lemma 1, we only need to show that $\widetilde{W}_*$ defined in Equation (5) equals $\widetilde{W}$ in Equation (8). By comparing Equation (3) and (7), we see that $\widetilde{\Omega} = \widetilde{\Omega}_* - \bar{r}_p \bar{r}_p'$, so it suffices to show that

$$1 - \bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p = \left(1 + \bar{r}_p' \widetilde{\Omega}^{-1} \bar{r}_p \right)^{-1} = \left[1 + \bar{r}_p' \left(\widetilde{\Omega}_* - \bar{r}_p \bar{r}_p' \right)^{-1} \bar{r}_p \right]^{-1}. \tag{A8}$$

Applying Lemma A2 with $A = \widetilde{\Omega}_*$, $u = \bar{r}_p$ and $v = -\bar{r}_p$, we get $\left(\widetilde{\Omega}_* - \bar{r}_p \bar{r}_p' \right)^{-1} = \widetilde{\Omega}_*^{-1} + \frac{\widetilde{\Omega}_*^{-1} \bar{r}_p \bar{r}_p' \widetilde{\Omega}_*^{-1}}{1 - \bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p}$, which implies that $1 + \bar{r}_p' \left(\widetilde{\Omega}_* - \bar{r}_p \bar{r}_p' \right)^{-1} \bar{r}_p = 1 + \bar{r}_p \widetilde{\Omega}_*^{-1} \bar{r}_p + \frac{\left(\bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p \right)^2}{1 - \bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p} = \left(1 - \bar{r}_p' \widetilde{\Omega}_*^{-1} \bar{r}_p \right)^{-1}$, which further immediately implies Equation (A8). This completes the proof of Theorem 1. □

Original GRS statistic when $L = 1$. When $L = 1$, $\widetilde{\Omega}$ equals to $\frac{1}{T}\sum_{t=1}^{T}\tilde{r}_{pt}^2 - \bar{r}_p^2 = s_p^2$, the sample variance of $\tilde{r}_{pt}$ **without** d.f. defined by Gibbons et al. (1989, p. 1124). So, $\widetilde{W}$ equals the original GRS statistic when $L = 1$.

Over-rejection of $\widehat{W}$. Take the ratio between $\widehat{W}$ and $\widetilde{W}$, then by the relationship between $\widehat{\Omega}$ and $\widetilde{\Omega}$ in Equation (9), we get

$$\frac{\widehat{W}}{\widetilde{W}} = \frac{1 + \bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p}{1 + \frac{T-1}{T}\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p}, \tag{A9}$$

which measures how much the incorrect formula inflates the GRS statistic. Define a function $g(x) = \frac{1+x}{1+\frac{T-1}{T}x}$. Since the first-order derivative of this function is $g'(x) = \frac{1/T}{\left(1+\frac{T-1}{T}x\right)^2} > 0$, we know that $g(x)$ is a monotonically increasing function of x. This, combined with the facts that $g(0) = 1$ and $\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p > 0$, implies that $\widehat{W}/\widetilde{W} > 1$. As a result, when $\widehat{W}$ is gauged against the $F_{N,T-N-L}$, the distribution of $\widetilde{W}$, it will over-reject the null hypothesis of mean-variance efficiency of the L portfolios.

Model misranking by $\widehat{W}$. Some back-of-the-envelop calculation shows that $\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p$ tends to be larger for models with more factors. To see this, let $\mu_{\tilde{r}_p}$ denote the mean vector of $\tilde{r}_{pt}$ as above, then by the central limit theorem, we have $\sqrt{T}(\bar{r}_p - \mu_{\tilde{r}_p}) \xrightarrow{d.} \mathcal{N}(0, \Omega)$; and by the law of large numbers, we have $\widetilde{\Omega} \xrightarrow{p.} \Omega$. These two results imply that $T(\bar{r}_p - \mu_{\tilde{r}_p})'\widetilde{\Omega}^{-1}(\bar{r}_p - \mu_{\tilde{r}_p}) \xrightarrow{d.} \chi_L^2$. Note that $E(\chi_L^2) = L$, so this in turn implies that for fixed T, the mean of $\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p$ is approximately $E(\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p) \approx \frac{L}{T} + \mu_{\tilde{r}_p}'\Omega^{-1}\mu_{\tilde{r}_p}$, where $\mu_{\tilde{r}_p}'\Omega^{-1}\mu_{\tilde{r}_p}$ is expected to increase with L since the dimensions of both $\mu_{\tilde{r}_p}$ and Ω increase with L. As a result, the random variable $\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p$ tends to increase with L *on average*.[7]

Combined with Equation (A9), this means that the ratio $\widehat{W}/\widetilde{W}$ tends to be larger for larger models; that is, smaller models tend to be disproportionally favored if the incorrect GRS statistic $\widehat{W}$ is used to rank models, compared to the ranking based on the correct GRS statistic $\widetilde{W}$.

Additional errors in software packages. The R package `GRS.test` computes two different GRS statistics, see Kim (2022). One (function `GRS.test`) uses the unbiased estimators $\widehat{\Omega}$ and $\widehat{\Sigma}$ at the same time; the other (function `GRS.MLtest`) uses the MLEs $\widetilde{\Omega}$ and $\widetilde{\Sigma} \equiv \frac{1}{T}\sum_{t=1}^{T}\hat{\eta}_t'\hat{\eta}_t$ at the same time. The former is just $\widehat{W}$, and we denote the latter as

$$\breve{W} \equiv \frac{T(T-N-L)}{N(T-L-1)}\left(1 + \bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p\right)^{-1}\hat{\delta}_0'\widetilde{\Sigma}^{-1}\hat{\delta}_0, \text{ and note } \breve{W} = \frac{T}{T-L-1}\widetilde{W}. \tag{A10}$$

These two statistics are both incorrect and clearly stem from the interpretation of Ω and Σ as variance–covariance matrices in Gibbons et al. (1989).

Because of the relationship between $\widetilde{W}$ and $\breve{W}$ in Equation (A10), similar analysis as in that for $\widehat{W}$ indicates the same over-rejection and misranking problems for $\breve{W}$ as well, even to a worse extent than $\widehat{W}$ for typical data in empirical asset pricing studies.

The Stata packages `grstest` and `grstest2`, composed by different contributors,[8] make use of $\widehat{\Omega}$ and further compound this error by estimating Σ as $\frac{1}{T-1}\sum_{t=1}^{T}\hat{\eta}_t\hat{\eta}_t'$ and pre-multiplying the ratio $\frac{T-N-L}{N}$ instead of $\frac{T(T-N-L)}{N(T-L-1)}$. The result is a statistic that is difficult to justify and different from all those we discussed above.[9]

Asymptotic χ^2 test. First note that the distribution of the correct GRS statistic $\widetilde{W}$, when multiplied by N, converges to the χ_N^2 distribution as $T \to \infty$.[10] So, comparing $N\widetilde{W}$ with the critical value from the χ_N^2 distribution, rather than comparing $\widetilde{W}$ with the $F_{N,T-N-L}$ critical value, is by itself an asymptotically valid χ^2 test. The commonly used χ^2 test statistics in

empirical asset pricing research deviate from $N\widetilde{W}$, and the deviations are all positive,[11] so they will over-reject compared to $N\widetilde{W}$.[12] We find that the χ^2 statistics misrank models more often even than $\check{W}$ in our empirical studies, but we do not report the model ranking results for χ^2, because they do not have an intuitive economic interpretation in the model ranking context, and therefore are not commonly used for this purpose. The misrankings of these χ^2 can be easily shown by a similar analysis as for $\widehat{W}$ and $\check{W}$ and are therefore skipped here.

Appendix B. Details of the Empirical Results

We now turn to some empirical examples, focusing on how the different implementations of the GRS statistic, as well as the asymptotic χ^2 statistic, compare to the correct calculation, based on model testing and ranking outcomes, borrowing from Fama and French (2015, 2016) the choice of asset pricing models and the choice of test assets. The models we consider include the CAPM, the Fama–French three-factor model, two variations of a four-factor model, the Fama–French five-factor model and a six-factor model that includes momentum. The test assets we explore include 5×5 sortings based on market capitalization and various anomaly variables including operating profitability, return volatility, residual volatility, accruals and so on, up to as many as 32 ($2 \times 4 \times 4$) portfolio sortings. We also explore decile portfolio sortings based on size, operating profitability, momentum, book-to-market and investment. The number of test assets used in empirical work is commonly as large as 25, as we see in Fama and French (2015, 2016), though many studies use 30 to over 50 test assets. See, for instance, Lewellen et al. (2010), Kroencke (2017), Demaj et al. (2018), and Kleibergen and Zhan (2020). Recently, asset pricing models have typically contained at least four or five factors, though six are also commonly seen. See, for instance, Barillas and Shanken (2018), Fama and and French (2018), Kan et al. (2024) or Hanauer (2020). Given the state of the literature, our choice of test assets and factors sits comfortably amidst the typical empirical asset pricing applications.

We use data retrieved from the French data library, and we consider five, ten, fifteen, twenty, twenty-five, forty, and fifty-year periods drawn from 1963–2019 for our consideration of up to six factors in the competing asset pricing models, and from 1926–2019 for our consideration of up to four factors in the competing asset pricing models.[13]. We limit our sample window to no less than five years of monthly data because few studies use less than 60 observations; Gibbons et al. (1989) note that issues of stationarity can reasonably constrain the length of a time series used, so that "it is not uncommon to published work where T is around 60", Affleck-Graves and McDonald (1989) limited their analysis and simulations to 60 month periods; Ferson and Foerster (1994) studied 60, 120, and 720 monthly observations in their simulation study; Rouwenhorst (1999) used five years in sub-sample analysis; and among recent work that exploited as few as four or five years of data are Belimam et al. (2018), and Qin (2019). Leite et al. (2018) used as few as 98 months of data, Lewellen, Nagel, and Shanken (2010) used 168 observations of quarterly data, Choi et al. (2020) performed sub-sample stability tests using eight years of monthly data, and many studies of emerging economy markets have used ten to fifteen years of monthly data. See, for instance, Alhomaidi et al. (2019), Alshammari and Goto (2022), Merdad et al. (2015), and Sha and Gao (2019).

Our primary results, found in Tables A1–A7, make use of the full sample available to us by partitioning the data sample into overlapping periods. For instance, at the five year horizon over 1963–2019, we form five-year windows starting in 1963 and every year following, so that the first window extends from July 1963 to June 1967, the second from January 1964 to December 1968, January 1965 to December 1969, and so on, resulting in 53 five-year overlapping windows. For each of these windows over the period 1963–2019, we use 19 sets of test assets, listed in the first column of Table A1, and six competing asset pricing models. These models are the CAPM, the Fama–French three-factor model, four and five-factor models, as well as a six-factor model including momentum, all as considered in Fama and French (2015, 2016).

Appendix B.1. Results for Five-Year Windows

We first present a small subset of our empirical findings in Tables A1–A3. For convenience, Tables A1 and A2 replicate Tables 1 and 2 from the main text, and here we discuss them in greater depth. In these tables, we consider five-year windows, the minimum span of data the GRS statistic is commonly applied to. This short span of data shows the most serious over-rejection of asset pricing models from the application of the alternative formulations of the GRS test, as well as the highest frequency of misrankings relative to the correct formulation of the GRS statistic. We present evidence for longer spans of data, up to 50 year windows, in Tables A4–A7, and discuss them in Appendix B.2.

In Table A1, we present the number of excess test rejections at the 1%, 5%, and 10% levels, relative to the correct GRS statistic $\widetilde{W}$, for each of the alternative test statistics, the asymptotic χ^2, $\widehat{W}$ and $\check{W}$. The results for the largest number of test assets we considered, 32, are displayed in the first three rows of the table; the results for cases with 25 test assets follow on rows four through thirteen; the 17 test assets of the industry portfolios follow on row fourteen; and the remaining five rows present results for sets of 10 test assets. In Table A1, we use a total of 53 five-year overlapping windows starting from 1963 and consider six different asset pricing models, meaning that we have 318 cases for which an asset pricing model might be rejected, for each of the 19 sets of test assets.

Table A1. Number and proportion of subsamples with more rejections relative to the GRS test for five year windows sampled during 1963–2019, across 19 sets of test assets.

Test Assets **Nb of Subsamples = 53**	**Asymptotic χ^2**			$\check{W}$			$\widehat{W}$		
	1%	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
2 × 4 × 4 MExMEBExINV	289	275	247	4	17	21	0	1	0
2 × 4 × 4 MExMEBExOP	262	246	220	4	14	18	0	0	0
2 × 4 × 4 MExOPxINV	282	238	194	10	27	28	0	3	1
5 × 5 AccrualsxME	222	218	207	7	17	31	0	0	1
5 × 5 BExME	212	191	166	8	21	20	1	0	1
5 × 5 BetaxME	229	221	204	11	16	25	0	1	0
5 × 5 MExOP	238	227	188	9	27	42	0	0	0
5 × 5 MomentumxME	210	166	124	14	29	27	0	0	0
5 × 5 NetIssuexME	216	176	147	9	29	27	0	2	2
5 × 5 RVariancexME	147	94	65	18	28	12	0	2	0
5 × 5 VariancexME	137	81	45	26	27	12	0	3	1
5 × 5 BExInv	248	259	245	4	18	16	0	2	1
5 × 5 MExInv	214	189	171	14	17	29	0	0	0
Industry	126	153	152	12	9	22	0	1	0
Book-to-Market Deciles	48	70	67	5	22	21	0	1	1
Investment Deciles	22	30	56	4	4	6	0	0	0
Momentum Deciles	56	60	66	18	14	20	1	1	1
Size Deciles	49	90	68	6	23	27	0	0	2
Operating Profitability Deciles	48	70	65	4	24	15	0	1	0
Average	171.3	160.7	142	9.8	20.2	22.1	0.11	0.95	0.58
Proportion (%)	53.9	50.5	44.6	3.1	6.3	6.9	0.0	0.3	0.2

Notes: (1) The figures are the number of all decisions at the stated significance level for which the test statistic rejects the model when the correct GRS statistic does not reject, out of a total of 318 possible, with the exception of the last row for which the proportion is given. The sample periods of five years are sampled over July 1963 to December 2019 and the number of models tested are 6 for each window, which we sum over to obtain the total number of over-rejections. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. This means that there are 53 samples for the 5 year window. (2) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

The asymptotic χ^2 statistic fares the worst relative to the correct GRS statistic among the alternatives, over-rejecting roughly 50% of the time on average relative to $\widetilde{W}$, across

the common significance levels of 1%, 5%, and 10%. This over-rejection is worse when we consider more test assets.

The $\check{W}$ and $\widehat{W}$ do not display patterns related to the number of test assets or significance level, with the $\check{W}$ ($\widehat{W}$) over-rejecting relative to the correct GRS statistic about 5% (0.2%) of the time on average, across the common significance levels of 1%, 5%, and 10%. The over-rejection of the $\widehat{W}$ is unstable across test assets, however, with a few cases displaying close to 6% over-rejection (three times over 53 subsamples), and some with no over-rejection.

In Table A2, we present the number of cases for which each statistic misranks factor models relative to the correct GRS statistic ranking. Here, we rank the six models for each of the 53 five-year windows considered in Table A1. Misranking of models at the five year horizon by the $\check{W}$ statistic displays much worse performance than we saw for test rejections, with close to 60% of the cases displaying a misranking, and we saw even worse performance for the $\widehat{W}$, with close to 3% of the cases misranked. If we restrict our attention to cases for which the top model is misranked, the $\check{W}$ statistic misranks between 10% and 15% of the time, while the $\widehat{W}$ misranks the top model less than 0.5% of the time.

Table A2. Number and proportion of subsamples with different ranking outcomes from the GRS statistic for five year windows sampled during 1963–2019, across 19 sets of test assets.

Test Assets	**Any Model Mis-Ranked**		**Top Model Mis-Ranked**	
Nb of Subsamples = 53	$\check{W}$	$\widehat{W}$	$\check{W}$	$\widehat{W}$
2 × 4 × 4 MExMEBExINV	26	2	5	0
2 × 4 × 4 MExMEBExOP	37	1	6	0
2 × 4 × 4 MExOPxINV	31	1	12	0
5 × 5 AccrualsxME	36	1	12	0
5 × 5 BExME	26	1	7	0
5 × 5 BetaxME	31	1	10	0
5 × 5 MExOP	33	0	10	0
5 × 5 MomentumxME	33	0	14	0
5 × 5 NetIssuexME	27	4	9	2
5 × 5 RVariancexME	32	0	14	0
5 × 5 VariancexME	34	1	10	0
5 × 5 BExInv	35	3	6	0
5 × 5 MExInv	30	3	9	0
Industry	29	2	4	0
Book-to-Market Deciles	28	0	4	0
Investment Deciles	24	2	4	0
Momentum Deciles	30	2	5	0
Size Deciles	27	3	5	0
Operating Profitability Deciles	30	0	6	0
Average	30.47	1.42	8.00	0.11
Proportion (%)	57.5	2.7	15.1	0.2

Notes: (1) The figures are the number of all misrankings from a particular test statistic value across models relative to the GRS statistic ranking, out of a total of 53 possible, with the exception of the last row for which the proportion is given. The sample periods of five years are sampled over July 1963 to December 2019 and the number of models ranked are 6 for each window. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. This means that there are 53 samples for the 5 year window. (2) For detailed description of the factor and test asset construction see Fama and French (2015, 2016).

In Table A3, we present detailed results for the 5 × 5 net share issuance's crossed with size portfolio test asset set, for five overlapping five-year windows near the end of the 1963–2019 sample period, in order to give the reader a finer sense for the results in Tables A1 and A2. We present the average annualized and raw percentage alpha for each of the six asset pricing models and each window, as well as values of the correct GRS statistic $\widetilde{W}$, and the $\check{W}$ and $\widehat{W}$ statistics. Beside each test statistic value, we report the ranks of the six models, from one to six. A model ranked differently by $\widehat{W}$ or $\check{W}$ from $\widetilde{W}$ is indicated by a † next to the factor label and further identified with the appropriate column's rank

number being boldfaced. A misranked top model is indicated by an $*$ next to the factor label and further identified with the appropriate row's factor label being bolded. A test method producing different ranks using p-values from test statistic is indicated by a ‡ next to the window period and further identified with a ⇓ in the appropriate column. This issue of different ranking using the test statistic versus the p-value will be drawn out below.

Table A3. Summary statistics on factor models and test statistics for five year windows over 2007–2009 financial crisis sample period, for investment 5 × 5 NetIssuexME test assets.

Date/ Factor Model	Average Annualized /Raw % α	$\widetilde{W}$ Statistic /Rank	$\check{W}$ Statistic /Rank	$\widehat{W}$ Statistic /Rank
JAN 2005-DEC 2009				
Mkt	2.63/0.219	0.688/1	0.712/1	0.688/1
Mkt SMB HML	2.39/0.199	0.819/2	0.878/2	0.819/2
Mkt SMB HML UMD	2.37/0.198	0.837/3	0.913/3	0.837/3
Mkt SMB RMW CMA	2.54/0.212	0.910/4	0.992/4	0.912/4
Mkt SMB HML RMW CMA	2.52/0.210	0.997/6	1.108/6	0.999/6
Mkt SMB HML RMW CMA UMD	2.57/0.214	0.972/5	1.100/5	0.974/5
JAN 2006-DEC 2010 ‡			⇓	
Mkt †	3.23/0.269	1.114/3	1.152/**1**	1.114/3
Mkt SMB HML †	2.32/0.193	1.078/2	1.155/2	1.078/**1**
Mkt SMB HML UMD †	2.41/0.201	1.256/6	1.371/**5**	1.257/6
Mkt SMB RMW CMA $*$,†	2.64/0.220	1.077/1	1.174/**3**	1.080/**2**
Mkt SMB HML RMW CMA	2.65/0.221	1.130/4	1.256/4	1.134/4
Mkt SMB HML RMW CMA UMD †	2.81/0.234	1.253/5	1.418/**6**	1.257/5
JAN 2007-DEC 2011 ‡				⇓
Mkt †	3.39/0.283	1.329/2	1.375/**1**	1.329/**1**
Mkt SMB HML †	3.11/0.259	1.503/4	1.610/**3**	1.504/4
Mkt SMB HML UMD	3.26/0.272	1.888/5	2.060/5	1.890/5
Mkt SMB RMW CMA $*$,†	2.89/0.240	1.327/1	1.448/**2**	1.332/**2**
Mkt SMB HML RMW CMA †	2.92/0.244	1.478/3	1.643/**4**	1.484/3
Mkt SMB HML RMW CMA UMD	3.08/0.257	1.947/6	2.204/6	1.955/6
JAN 2008-DEC 2012				
Mkt	4.34/0.361	1.299/2	1.344/2	1.300/2
Mkt SMB HML	3.64/0.303	1.500/3	1.607/3	1.500/3
Mkt SMB HML UMD †	3.77/0.314	1.906/6	2.079/**5**	1.907/6
Mkt SMB RMW CMA	2.98/0.248	1.161/1	1.267/1	1.165/1
Mkt SMB HML RMW CMA	3.15/0.263	1.577/4	1.753/4	1.583/4
Mkt SMB HML RMW CMA UMD †	3.39/0.282	1.896/5	2.146/**6**	1.904/5
JAN 2009-DEC 2013 ‡		⇓		⇓
Mkt	2.74/0.228	1.678/2	1.736/2	1.681/2
Mkt SMB HML	3.05/0.254	2.938/5	3.148/5	2.946/5
Mkt SMB HML UMD	3.12/0.260	3.324/6	3.626/6	3.333/6
Mkt SMB RMW CMA	2.46/0.205	1.402/1	1.530/1	1.406/1
Mkt SMB HML RMW CMA	3.05/0.254	2.723/3	3.026/3	2.734/3
Mkt SMB HML RMW CMA UMD	2.72/0.227	2.733/4	3.093/4	2.745/4

Notes: (1) A test method producing different ranks using p-values versus test statistic is indicated by a ‡ in the assets label. This is further identified with a ⇓ in the appropriate column. Ranked test statistics different than $\widetilde{W}$ test ranked value is indicated by a † on the factor model label. This is further identified with the appropriate column's rank being bolded. Top ranked test statistics different than top ranked $\widetilde{W}$ test value are indicated by a $*$ in the assets label. This is further identified by the appropriate row's model being bolded. (2) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

What we see in Table A3 is fairly typical across the full set of empirical findings that Tables A1 and A2 are based on; the $\check{W}$ displays many misrankings, and misrankings of the top model are rare. Although not tabulated, only the asymptotic χ^2 is typically rejecting factor models in this particular small set of examples, so that the $\check{W}$, $\widehat{W}$, and correct $\widetilde{W}$ test

statistic are all consistent with each other. Common in studies that seek to rank models are rankings by the average absolute alpha, whether or not all models are rejected by the GRS test. See for instance Fama and French (2015). As we can see from Table A3, often the model with the smallest average absolute alpha is not top-ranked.

Appendix B.2. Summary Results for Longer Windows

We now present evidence for a longer span of data in Tables A4–A7, using two sets of date windows up to 50 years. In addition to the period of time 1963–2019 that we considered in Appendix B.1, now we add the period 1926–1962. The pre-1963 period lacks data for factors and test assets built using operating profitability, accruals etc., so we are left with six test assets constructed from book-to-market, size, industry classification, and momentum, and three different factor models, including the CAPM, the Fama–French three-factor model, and a four-factor model including momentum, all as considered in Fama and French (2015, 2016). For this restricted set of test assets and factors, we estimate test rejections and rankings using the entire 1926–2019 sample and we break out results separately from those constructed using the larger set of test assets and models on the 1963–2019 data alone. It is interesting to do this, as the chance of a misranking declines with fewer asset pricing models being considered.

In Table A4, we present the percentage of excess test rejections relative to the correct GRS statistic at each of the 1%, 5%, and 10% levels for each of the alternative test methods, the asymptotic χ^2, the $\check{W}$, and the $\widehat{W}$. This is performed for overlapping windows of 5, 10, 15, 20, 25, 40, and 50 years, using data that span either 1963–2019 or 1926–2019. Panel A displays the percentage of over-rejection rates averaged over six asset pricing models and 19 sets of test assets for the period 1963–2019; Panel B displays the same percentages for the period 1926–2019 using the smaller set of three factor models and six sets of test assets.

Table A4. Percentage of subsamples/models with different decision outcomes from the correct GRS test $\widetilde{W}$.

Window (Months)	χ^2 1%	χ^2 5%	χ^2 10%	$\check{W}$ 1%	$\check{W}$ 5%	$\check{W}$ 10%	$\widehat{W}$ 1%	$\widehat{W}$ 5%	$\widehat{W}$ 10%
Panel A: 1963–2019									
60	53.9	50.5	44.6	3.1	6.3	6.9	0.0	0.3	0.2
120	27.4	24.8	20.0	3.2	3.7	3.9	0.1	0.0	0.1
180	16.4	13.0	11.4	2.1	2.1	2.2	0.1	0.0	0.0
240	9.7	8.6	6.3	1.4	1.7	1.2	0.0	0.0	0.0
300	6.5	5.7	4.5	0.9	1.3	1.1	0.1	0.1	0.0
480	3.0	2.9	1.9	0.3	0.7	0.4	0.0	0.0	0.0
600	2.9	0.8	1.6	0.3	0.2	0.3	0.0	0.0	0.0
Panel B: 1926–2019									
60	34.6	36.1	33.2	2.6	4.3	5.3	0.0	0.1	0.1
120	15.4	15.5	12.6	2.2	3.1	2.5	0.0	0.0	0.0
180	8.8	10.3	9.2	1.0	1.4	1.7	0.1	0.0	0.0
240	8.6	7.7	4.7	1.3	2.0	1.3	0.0	0.0	0.0
300	6.0	4.4	3.1	1.0	0.8	0.6	0.0	0.0	0.0
480	3.9	2.8	0.9	0.7	0.7	0.2	0.0	0.0	0.0
600	2.2	1.7	0.6	0.6	0.1	0.2	0.0	0.0	0.0

Notes: The figures are the percentage of all decisions at the stated significance level for which the test statistic rejects the model when the correct GRS statistic does not reject. (1) For Panel A, the sample periods cover July 1963 to December 2019, the sample window for the test is either 5, 10, 15, 20, 25, 40, or 50 years, the number of models tested are 6 for each window, and we average over 19 different sets of test assets, listed in Table A1. These windows overlap, adjusted in a rolling window, so that all but 1 year of data overlaps with the next sample window. This means that there are 53 samples for the 5-year window. The rejection rates are aggregated over the 6 models employed, so that a 54% value in the top leftmost cell corresponds to roughly 171 incorrect rejections on average. The factor models considered are the CAPM, the Fama–French 3 factor model, 4 and 5 factor models, as well as a six-factor model including momentum, all as considered in Fama and French (2016). (2) For Panel B, the sample periods cover January 1926 to December 2019. The models considered here number three, as factors for size, book-to-market, momentum and the market are all that are available. The test assets are constructed from industry classification, book-to-market crossed with size, momentum crossed with size, book-to-market, size and momentum decile portfolios. (3) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

For both Panels A and B, we see a virtually monotonic decline in over-rejections as sample size increases, albeit at a fairly slow rate. The asymptotic χ^2 statistic over-rejects roughly 5% of the time relative to the $\widetilde{W}$ statistic even with a 25-year window. The $\check{W}$ over-rejects roughly 5% of the time with 5 years of data, and over 1% of the time even with a 25-year window. The $\widehat{W}$, which is fairly close to the correct $\widetilde{W}$ statistic, does not appear to over-reject with more than 25 years of data, and over-rejects less than 0.1% of the time with 10 or more years of data. The simulations based on models and calibrations described in Appendix C.1 reveal that $\widehat{W}$ always over-rejects, at least in this experimental design. The over-rejection declines with sample size, but at a decreasing (non-linear) rate, increases with the number of factors, and is largely unrelated to the number of test assets.

In Tables A5 and A6, we present the number of cases for which each statistic misranks factor models relative to the correct $\widetilde{W}$ statistic's ranking by test value and p-value respectively. Misranking of models by the $\check{W}$ statistic is remarkably large, even at a 50-year data horizon if the set of models is as large as six, for either ranking method (statistic value or p-value). Replacing the correct form of the GRS test with $\check{W}$ scrambles rankings over 15% of the time even at the 50-year horizon, and over 50% of the time at the 5-year horizon, for cases with six models (Panel A). When there are only three models (Panel B), misrankings are naturally fewer, and for data horizons over 15 years misrankings occur mostly less than 5% of the time across methods. The $\widehat{W}$ is typically consistent with the correct $\widetilde{W}$ statistic, though misrankings occur even at the 40-year window length.

Table A5. Percentage of subsamples/models with different model rank outcomes from the correct GRS statistic $\widetilde{W}$, ranked by test statistic value.

Window (Months)	Any Model Misranked		Top Model Misranked	
	$\check{W}$	$\widehat{W}$	$\check{W}$	$\widehat{W}$
Panel A: 1963–2019				
60	57.5	2.7	15.1	0.2
120	37.6	0.9	9.4	0.0
180	21.4	0.0	4.4	0.0
240	12.5	0.0	3.0	0.0
300	9.7	0.2	2.4	0.2
480	14.3	0.3	1.8	0.0
600	15.1	0.0	0.0	0.0
Panel B: 1926–2019				
60	17.4	0.7	9.6	0.4
120	10.6	0.2	5.5	0.0
180	5.2	0.2	1.9	0.0
240	3.1	0.0	0.9	0.0
300	1.9	0.0	1.2	0.0
480	1.5	0.3	0.6	0.0
600	1.1	0.0	0.7	0.0

Notes: The figures are the percentage of misrankings from a particular test statistic value across models relative to the GRS statistic ranking. (1) For Panel A, the sample periods cover July 1963 to December 2019, the sample window for the test is either 5, 10, 15, 20, 25, 40, or 50 years, the number of models tested is 6 for each window, and we aggregate over 19 different sets of test assets, listed in Table A1. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. The factor models considered are the CAPM, the Fama–French 3-factor model, 4 and 5-factor models, as well as a 6-factor model including momentum, all as considered in Fama and French (2016). (2) For Panel B, the sample periods cover January 1926 to December 2019. The models considered here number three, as factors for size, book-to-market, momentum and the market are all that are available. The test assets are constructed from industry classification, book-to-market crossed with size, momentum crossed with size, book-to-market, size and momentum decile portfolios. (3) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

Table A6. Percentage of subsamples/models with different model rank outcomes from the correct GRS statistic $\widetilde{W}$, ranked by test p-value.

Window (Months)	Any Model Mis-Ranked $\check{W}$	Any Model Mis-Ranked $\widehat{W}$	Top Model Mis-Ranked $\check{W}$	Top Model Mis-Ranked $\widehat{W}$
Panel A: 1963–2019				
60	54.6	1.9	13.1	0.4
120	37.4	0.5	9.1	0.0
180	21.9	0.4	4.4	0.1
240	11.5	0.1	2.8	0.0
300	9.4	0.2	2.1	0.0
480	13.2	0.3	1.8	0.0
600	13.2	0.0	0.0	0.0
Panel B: 1926–2019				
60	15.0	0.7	8.1	0.7
120	10.2	0.0	5.5	0.0
180	5.4	0.0	2.1	0.0
240	3.3	0.2	0.9	0.0
300	1.9	0.0	1.2	0.0
480	1.2	0.0	0.6	0.0
600	1.1	0.0	0.7	0.0

Notes: The figures are the percentage of misrankings from a particular test statistic p-value across models relative to the GRS statistic ranking. (1) For Panel A the sample periods cover July 1963 to December 2019, the sample window for the test is either 5, 10, 15, 20, 25, 40, or 50 years, the number of models tested is 6 for each window, and we aggregate over 19 different sets of test assets, listed in Table A1. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. The factor models considered are the CAPM, the Fama–French 3-factor model, 4 and 5-factor models, as well as a 6-factor model including momentum, all as considered in Fama and French (2016). (2) For Panel B, the sample periods cover January 1926 to December 2019. The models considered here number three, as factors for size, book-to-market, momentum and the market are all that are available. The test assets are constructed from industry classification, book-to-market crossed with size, momentum crossed with size, book-to-market, size and momentum decile portfolios. (3) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

If we restrict our attention to cases for which the top model is misranked, the $\widehat{W}$ statistic is consistent with the correct $\widetilde{W}$ statistic once we have 40 or more years of data, but $\check{W}$ misranks even at 40 years of data. Again, when there are only three models (Panel B), misrankings are fewer, and for data horizons over 10 years misrankings occur less than 5% of the time across methods.

Finally, in Table A7, we present evidence on rankings from these different test statistics compared to rankings based on the p-values of these test statistics, to see if they are consistent. We can think of ranking by the test statistic as a sort of mean squared error or model R^2 ranking - perhaps helpful if we are interested in minimizing model prediction error even if models are false (see, for instance, Teräsvirta and Mellin 1986). Here, we see that all the rankings by test statistics are fragile, even at a 50 year horizon, averaging close to 10% misranked if we compare three or six asset pricing models to each other. Table A7 highlights that the statistically sound p-values, instead of the raw GRS statistics, should be used in model ranking.

Considering the case of a true model that includes a subset of the available factors, an untabulated analysis of the simulated test rankings formed using the magnitude of the GRS statistic confirms that the probability of an incorrect model, larger than the true model, achieving a high rank versus other models increases with the number of factors in the model, when the GRS statistic ranking differs from the p-value ranking. This bias is stronger when using an incorrect GRS statistic. In cases for which both the GRS statistic ranking and the p-value ranking agree, there is no such pattern to tilt to larger models than the true model.

The important insight to take away from these results is that the error in calculating the GRS statistic can have a material impact on empirical results, particularly when twenty

or fewer years of data are used, which is not uncommon in empirical asset pricing studies. For example, Barillas and Shanken (2018) performed model comparisons on a little less than 15 years of monthly data. Harvey and Liu (2021) considered tests of asset pricing models and report simulations for 20 and 40 years of monthly data. Sha and Gao (2019) used 144 months of data and exploited 6 metrics to evaluate factor model performance, including the GRS statistic. Baek and Bilson (2015) considered 234 months of data in a subsample estimation. Chiah et al. (2016) used 23 years of data when comparing models using the GRS statistic. One takeaway from these papers is that many situations involving specialized data (like Sha and Gao 2019, exploring mutual fund returns in China) or subsample robustness checks (like Baek and Bilson 2015) are necessarily constrained to shorter samples than fifty or even twenty years, so that the bias from an incorrectly calculated GRS statistic becomes large.

Table A7. Percentage of subsamples/models with different model rankings if ranked by *p*-values rather than test statistics.

Window (Months)	$\widetilde{W}$	$\check{W}$	$\widehat{W}$
Panel A: 1963–2019			
60	16.3	21.8	16.8
120	4.9	5.8	5.3
180	4.3	3.9	3.9
240	4.6	5.4	4.4
300	4.8	4.8	4.8
480	2.6	3.8	2.6
600	6.6	7.9	6.6
Panel B: 1926–2019			
60	2.2	5.0	2.6
120	0.6	1.2	0.8
180	0.4	0.2	0.6
240	0.4	0.2	0.2
300	0.7	0.7	0.7
480	5.5	5.8	5.8
600	7.4	7.4	7.4

Notes: The figures are the percentage of misrankings from a particular test statistic value across models relative to the GRS statistic ranking. (1) For Panel A, the sample periods cover July 1963 to December 2019, the sample window for the test is either 5, 10, 15, 20, 25, 40, or 50 years; the number of models tested are 6 for each window; and we average over 19 different sets of test assets, listed in Table A1. These windows overlap, adjusted in a rolling window so that all but 1 year of data overlaps with the next sample window. The factor models considered are the CAPM, the Fama–French 3-factor model, 4 and 5-factor models, as well as a 6-factor model including momentum, all as considered in Fama and French (2016). (2) For Panel B, the sample periods cover January 1926 to December 2019. The models considered here number three, as factors for size, book-to-market, momentum and the market are all that are available. The test assets are constructed from industry classification, book-to-market crossed with size, momentum crossed with size, book-to-market, size and momentum decile portfolios. (3) For a detailed description of the factor and test asset construction see Fama and French (2015, 2016).

We do not recommend ranking models with the magnitude of the GRS statistic, and instead suggest the use of the *p*-value of the statistic from the exact F-distribution, since the *p*-value internalizes the different degrees of freedom of the GRS statistics computed for models with a different number of factors. We recognize that ranking of models by the GRS statistic has a desirable economic intuition—this is a direct, easy-to-understand metric tied to a model's factors spanning the test asset returns. But researchers need to at least understand that this ranking might have undesirable statistical properties. The detailed results that these tables are based on, and additional summary statistics are available on request.

Appendix C. Simulation Results

Here, we present simulation results to evaluate the size performance of the alternative tests $\widetilde{W}$, $\widehat{W}$, $\check{W}$ and χ^2. We conducted three sets of simulations, differentiated by the error generating process. Within each of these three sets of simulations, we look at two different factor models, one with three factors and one with six factors, and we look at two different groupings of test assets, one grouping being decile portfolios and one being a five by five set of twenty five portfolios. For each of these, we consider monthly samples of lengths 5, 10, 15, 20, 25, 40, and 50 years.

The first two sets of simulation results exploit the normal and t-distributions for the errors, calibrating return moments (mean, volatility, covariance) to French's portfolio returns. For the decile returns, we calibrate to the size-sorted decile monthly returns over 1963/7–2019/12, and for the 5 by 5 test asset case, we calibrate to the size by book-to-market returns, again over 1963/7–2019/12. The last set of simulations employed bootstrap methods and the French portfolio returns, described fully below.

Appendix C.1. Simulated Normal and t-Distributed Returns

Following the literature, we generate portfolio excess returns $\tilde{r}_{pt}$ as normal, independent and identically distributed, calibrated to monthly U.S. stock returns. We also consider the t-distribution with eight degrees of freedom and a bootstrap simulation.

The excess return for test asset i and time t is generated based on model (1), which is rewritten here:

$$\tilde{r}_{it} = \delta_{i0} + \sum_{j=1}^{L} \delta_{ij}\tilde{r}_{jt} + \tilde{\eta}_{it}, \tag{A11}$$

where $\tilde{\eta}_{it} \sim iid\ \ normal$ across t with mean 0 and volatility σ_{ii}, and $\tilde{r}_{jt} \sim iid\ \ normal$ across t with mean (μ_j/L), volatility σ_j and $E\left[\tilde{\eta}_{it}\tilde{r}_{jt}\right] = 0$. We set $\mu_j = 0.01$, $\sigma_j = 0.02$, $\sigma_{ii} = 0.08$, $\delta_{ij} = 1,\ \forall\ i, j$ and we explore only the case of $\delta_{i0} = 0,\ \forall\ i$.

We explore the size properties of the correct and incorrect formulas of the GRS statistic for numbers of portfolios (L) from 3 to 6, test assets (N) from 10 to 25, and sample sizes (T) from 60 (months) to 600. This spans typical applications of the GRS statistic. Our simulations show that the performance of the incorrect GRS formula generally suffers deterioration as the number of firms and factors increases, as one might expect. We present simulation results for the normal case in Tables A8 and A9 for the t-distribution case.

Table A8. Null rejection rates.

Test Assets/Portfolios N/L	Years	$\widetilde{W}$		$\widehat{W}$		$\check{W}$		χ^2	
		1%	10%	1%	10%	1%	10%	1%	10%
10 /3	5	0.010	0.104	0.011	0.106	**0.016**	**0.132**	**0.062**	**0.240**
	10	0.011	0.100	0.011	0.102	0.013	**0.119**	**0.026**	**0.162**
	15	0.010	0.099	0.010	0.100	0.012	0.109	**0.020**	**0.135**
	20	0.008	0.094	0.008	0.096	0.010	0.101	**0.016**	**0.120**
	25	0.010	0.100	0.010	0.101	0.011	0.107	**0.017**	**0.122**
	40	0.009	0.102	0.009	0.103	0.010	0.106	0.012	**0.113**
	50	0.009	0.096	0.009	0.096	0.009	0.099	0.012	0.105
10 /6	5	0.009	0.104	0.010	**0.110**	**0.021**	**0.158**	**0.065**	**0.246**
	10	0.010	0.095	0.010	0.097	**0.015**	**0.124**	**0.026**	**0.156**
	15	0.011	0.105	0.011	0.107	**0.014**	**0.125**	**0.020**	**0.143**
	20	0.009	0.094	0.009	0.096	0.011	**0.111**	**0.015**	**0.124**
	25	0.010	0.103	0.010	0.104	0.012	**0.114**	**0.015**	**0.125**
	40	0.009	0.097	0.010	0.098	0.011	0.104	0.012	**0.111**
	50	0.009	0.098	0.009	0.099	0.010	0.103	0.011	0.109

Table A8. *Cont.*

Test Assets/Portfolios N/L	Years	$\widetilde{W}$		$\widehat{W}$		$\check{W}$		χ^2	
		1%	10%	1%	10%	1%	10%	1%	10%
25 /3	5	0.011	0.103	0.011	0.107	**0.016**	**0.141**	**0.484**	**0.730**
	10	0.008	0.101	0.008	0.104	0.011	**0.122**	**0.124**	**0.367**
	15	0.010	0.099	0.010	0.101	0.012	**0.115**	**0.062**	**0.260**
	20	0.009	0.099	0.010	0.101	0.011	**0.110**	**0.043**	**0.206**
	25	0.010	0.097	0.011	0.098	0.012	0.106	**0.034**	**0.180**
	40	0.009	0.100	0.009	0.101	0.009	0.105	**0.020**	**0.147**
	50	0.008	0.093	0.008	0.094	0.009	0.098	**0.018**	**0.136**
25 /6	5	0.008	0.097	0.009	0.104	**0.020**	**0.163**	**0.524**	**0.763**
	10	0.010	0.102	0.011	0.106	**0.018**	**0.140**	**0.130**	**0.378**
	15	0.011	0.103	0.012	0.107	**0.016**	**0.131**	**0.069**	**0.261**
	20	0.012	0.103	0.012	0.105	**0.016**	**0.124**	**0.046**	**0.212**
	25	0.009	0.100	0.010	0.102	0.013	**0.116**	**0.031**	**0.181**
	40	0.009	0.099	0.009	0.100	0.011	**0.111**	**0.020**	**0.149**
	50	0.011	0.099	0.011	0.100	0.012	0.107	**0.019**	**0.135**

Notes: (1) Bold-faced numbers are rejection rates three standard deviations larger than the nominal values. (2) These results are based on 10,000 simulations. (3) The models are $\tilde{r}_{it} = \delta_{i0} + \sum_{j=1}^{L} \delta_{ij}\tilde{r}_{jt} + \tilde{\eta}_{it}$, where $\tilde{r}_{jt} \sim iid\ \mathcal{N}\left(\mu_j/L, \sigma_j^2\right)$, $\tilde{\eta}_{it} \sim iid\ \mathcal{N}(0, \sigma_{ii}^2)$ and $E\left[\tilde{\eta}_{it}\tilde{r}_{jt}\right] = 0$. For $\forall j$, $\mu_j = 0.01$, $\sigma_j = 0.02$, $\sigma_{ii} = 0.08$, and $\delta_{ij} = 1$.

Table A9. Null rejection rates.

Test Assets/Portfolios N/L	Years	$\widetilde{W}$		$\widehat{W}$		$\check{W}$		χ^2	
		1%	10%	1%	10%	1%	10%	1%	10%
10 /3	5	0.009	0.104	0.009	0.107	**0.015**	**0.137**	**0.061**	**0.236**
	10	0.011	0.098	0.011	0.100	0.012	**0.113**	**0.026**	**0.158**
	15	0.009	0.100	0.009	0.101	0.011	0.109	**0.018**	**0.135**
	20	0.010	0.100	0.010	0.101	0.011	0.107	**0.017**	**0.125**
	25	0.010	0.099	0.010	0.099	0.011	0.106	**0.014**	**0.122**
	40	0.010	0.106	0.010	0.106	0.010	**0.109**	0.013	**0.120**
	50	0.009	0.100	0.009	0.101	0.010	0.105	0.012	**0.112**
10 /6	5	0.010	0.104	0.012	**0.110**	**0.023**	**0.163**	**0.068**	**0.256**
	10	0.010	0.095	0.010	0.098	**0.015**	**0.122**	**0.027**	**0.155**
	15	0.010	0.099	0.011	0.101	**0.014**	**0.117**	**0.020**	**0.136**
	20	0.009	0.096	0.009	0.098	0.012	**0.111**	**0.016**	**0.123**
	25	0.012	0.097	0.012	0.098	**0.014**	0.109	**0.017**	**0.120**
	40	0.010	0.099	0.010	0.099	0.011	0.105	**0.014**	**0.111**
	50	0.010	0.102	0.011	0.103	0.011	0.109	0.012	**0.114**
25 /3	5	0.009	0.095	0.009	0.098	**0.014**	**0.131**	**0.477**	**0.725**
	10	0.011	0.105	0.011	0.107	**0.014**	**0.127**	**0.129**	**0.370**
	15	0.011	0.103	0.011	0.105	0.013	**0.117**	**0.064**	**0.254**
	20	0.011	0.094	0.011	0.096	0.012	0.106	**0.041**	**0.206**
	25	0.009	0.095	0.009	0.096	0.010	0.104	**0.031**	**0.180**
	40	0.009	0.098	0.009	0.099	0.010	0.104	**0.021**	**0.145**
	50	0.010	0.097	0.010	0.098	0.011	0.102	**0.017**	**0.137**
25 /6	5	0.009	0.100	0.010	0.106	**0.022**	**0.169**	**0.533**	**0.772**
	10	0.010	0.103	0.011	0.108	**0.020**	**0.140**	**0.131**	**0.380**
	15	0.011	0.104	0.012	0.108	**0.015**	**0.132**	**0.069**	**0.266**
	20	0.010	0.101	0.010	0.103	**0.015**	**0.117**	**0.044**	**0.209**
	25	0.010	0.099	0.010	0.100	0.012	**0.114**	**0.032**	**0.180**
	40	0.009	0.102	0.010	0.104	0.011	**0.114**	**0.021**	**0.150**
	50	0.009	0.101	0.010	0.101	0.011	0.109	**0.019**	**0.141**

Notes: (1) Bold-faced numbers are rejection rates three standard deviations larger than the nominal values. (2) The results are based on 10,000 simulations. (3) The models are $\tilde{r}_{it} = \delta_{i0} + \sum_{j=1}^{L} \delta_{ij}\tilde{r}_{jt} + \tilde{\eta}_{it}$, where $\tilde{r}_{jt} \sim iid\ \left(\mu_j/L, \sigma_j^2\right)$, t with 8 df, $\tilde{\eta}_{it} \sim iid\ \left(0, \sigma_{ii}^2\right)$, t with 8 df, and $E\left[\tilde{\eta}_{it}\tilde{r}_{jt}\right] = 0$. For $\forall j$, $\mu_j = 0.01$, $\sigma_j = 0.02$, $\sigma_{ii} = 0.08$, and $\delta_{ij} = 1$.

The correct formula of the GRS statistic generally presents no evidence of incorrect size in Table A8, as our simulation setting is one in which it should have the correct small-sample exact size. Even with the t-distribution, the GRS performs well. The $\widehat{W}$ formula shows some evidence of over-rejection with the t-distribution and a small sample size. The $\check{W}$ formula of the GRS statistic and the χ^2 show strong over-rejection under 20 years of data and the χ^2 shows evidence of over-rejection even with 50 years of data.

Table A10. Null rejection rates.

Test Assets/Portfolios N/L	Years	$\widetilde{W}$		$\widehat{W}$		$\check{W}$		χ^2	
		1%	10%	1%	10%	1%	10%	1%	10%
10 /3	5	**0.097**	**0.290**	**0.097**	**0.292**	**0.115**	**0.333**	**0.228**	**0.458**
	10	**0.140**	**0.344**	**0.140**	**0.344**	**0.153**	**0.367**	**0.204**	**0.425**
	15	**0.167**	**0.372**	**0.167**	**0.372**	**0.176**	**0.385**	**0.210**	**0.417**
	20	**0.189**	**0.395**	**0.189**	**0.395**	**0.198**	**0.405**	**0.222**	**0.433**
	25	**0.202**	**0.408**	**0.202**	**0.408**	**0.207**	**0.417**	**0.230**	**0.436**
	40	**0.234**	**0.431**	**0.234**	**0.431**	**0.237**	**0.435**	**0.250**	**0.447**
	50	**0.249**	**0.450**	**0.249**	**0.450**	**0.251**	**0.455**	**0.261**	**0.465**
10 /6	5	**0.080**	**0.266**	**0.080**	**0.268**	**0.114**	**0.342**	**0.213**	**0.451**
	10	**0.124**	**0.323**	**0.124**	**0.324**	**0.146**	**0.361**	**0.183**	**0.402**
	15	**0.148**	**0.356**	**0.149**	**0.357**	**0.164**	**0.381**	**0.192**	**0.406**
	20	**0.173**	**0.374**	**0.173**	**0.374**	**0.184**	**0.394**	**0.203**	**0.415**
	25	**0.186**	**0.387**	**0.186**	**0.387**	**0.194**	**0.403**	**0.208**	**0.418**
	40	**0.213**	**0.417**	**0.213**	**0.417**	**0.218**	**0.425**	**0.227**	**0.434**
	50	**0.234**	**0.437**	**0.234**	**0.437**	**0.239**	**0.444**	**0.246**	**0.450**
25 /3	5	**0.021**	**0.144**	**0.021**	**0.145**	**0.031**	**0.184**	**0.546**	**0.776**
	10	**0.023**	**0.156**	**0.024**	**0.157**	**0.031**	**0.182**	**0.185**	**0.452**
	15	**0.026**	**0.167**	**0.026**	**0.167**	**0.031**	**0.183**	**0.113**	**0.349**
	20	**0.032**	**0.175**	**0.032**	**0.176**	**0.036**	**0.188**	**0.091**	**0.304**
	25	**0.037**	**0.177**	**0.037**	**0.177**	**0.039**	**0.188**	**0.082**	**0.279**
	40	**0.042**	**0.188**	**0.042**	**0.188**	**0.044**	**0.196**	**0.069**	**0.248**
	50	**0.044**	**0.192**	**0.044**	**0.192**	**0.047**	**0.197**	**0.067**	**0.235**
25 /6	5	**0.018**	**0.132**	**0.018**	**0.134**	**0.034**	**0.208**	**0.583**	**0.793**
	10	**0.022**	**0.145**	**0.023**	**0.146**	**0.032**	**0.194**	**0.185**	**0.449**
	15	**0.023**	**0.150**	**0.023**	**0.151**	**0.033**	**0.183**	**0.103**	**0.329**
	20	**0.025**	**0.158**	**0.025**	**0.158**	**0.032**	**0.182**	**0.078**	**0.288**
	25	**0.027**	**0.160**	**0.027**	**0.161**	**0.031**	**0.179**	**0.067**	**0.255**
	40	**0.030**	**0.167**	**0.030**	**0.167**	**0.033**	**0.181**	**0.056**	**0.228**
	50	**0.034**	**0.167**	**0.034**	**0.167**	**0.036**	**0.176**	**0.052**	**0.209**

Notes: (1) Bold-faced numbers are rejection rates three standard deviations larger than the nominal values. (2) The results are based on 10,000 simulations. (3) The models are $\tilde{r}_{it} = \delta_{i0} + \sum_{j=1}^{L} \delta_{ij}\tilde{r}_{jt} + \tilde{\eta}_{it}$, where the data were generated through a block-bootstrap approach.

Appendix C.2. Bootstrap Simulation

There are two main categories of bootstrapping in the regression context, the random X approach, which resamples the complete set of variables including the dependent variable for each observation, and the fixed X approach, which resamples residuals and explanatory variable values and forms simulated dependent variable values. That is, the fixed X approach builds simulated dependent variable values from the explanatory variables and either simulated or resampled regression residuals. The choice of using either simulated or resampled residuals is what distinguishes the major variations of the fixed X bootstrap approach. The non-parametric fixed X bootstrap approach, which we employ, uses resampled regression residuals.

Suppose we have a sample of T observations of a dependent variable $r_{i,t}$, $(i = 1, \ldots, N)$, a $K \times 1$ vector of factor portfolios $r_{p,t}$, and a regression model $E[r_{i,t}] = \alpha_i + \beta_i' r_{p,t}$. Define $\hat{\alpha}_i$ and $\hat{\beta}_i$ as the OLS estimates of α_i and β_i and, noting that we wish to explore the null

hypothesis that $\alpha_i = 0$, define $\hat{r}^*_{i,t} = \hat{\beta}_i r_{p,t}$ and $\hat{e}_{i,t} = r_{i,t} - (\hat{\alpha}_i + \hat{\beta}_i r_{p,t})$. We then form R resamples of $\hat{r}^*_{i,t}$, $(i = 1, \ldots, N)$, and $r_{p,t}$ with each resample containing T observations. Separately and independently, we form R resamples of $\hat{e}_{i,t}$ with each resample also containing T observations, and finally we form $r^*_{i,t} = \hat{r}^*_{i,t} + \hat{e}_{i,t}$ for each of the R resamples. Using $r^*_{i,t}$ and $r_{p,t}$, we fit the model $E[r^*_{i,t}] = \alpha_i + \beta_i r_{p,t}$ on each of these resampled datasets, and retrieve the various GRS test statistics for each resampled dataset.

To deal with a well-documented property of financial returns, lack of independence across time, we also employ block bootstrap resampling which allows for data dependence. See, for instance, Politis and Romano (1994), White (2000), and Gonçalves and White (2002, 2005). It is the resampling in (random-length) blocks from the original data that produces results incorporating data dependence. Politis and Romano (1994) used blocks of data with lengths distributed according to the geometric distribution. The mean block length b is data-dependent. Politis and Romano (1994) recommended a length proportional to $T^{1/3}$, where T = sample size, which is what we use.

Again, we exploit French's portfolio returns. For the decile returns, we resample from the size-sorted decile monthly returns over 1963/7–2019/12, and for the 5 by 5 test asset case we resample from the size by book-to-market returns, over 1963/7–2019/12. Bootstrap simulations show persistent over-rejection of the null hypothesis in all these tests, though the correct GRS F test shows the smallest over-rejection. Similarly to Harvey and Liu (2021), we find little or no over-rejection when we evaluate t-tests on the intercept with bootstrapped data, rather than a joint test across intercepts of the test assets, but joint tests appear much more fragile than the one-at-a-time t-tests on intercepts.

Appendix D. Software Packages

The SAS and R packages used to implement our generalized GRS test can be found at the authors' websites: http://markkamstra.com/data.html (accessed on 31 August 2023) (SAS) and https://ruoyaoshi.github.io/ (accessed on 31 August 2023) (R). A Stata package `grsftest` coded by Mengnan (Cliff) Zhu can be found at https://ideas.repec.org/c/boc/bocode/s458828.html (accessed on 31 August 2023). See Zhu (2020).

Appendix E. Detailed Model Statistics

Appendix E.1. 5 Year Windows over 1963–2019

Appendix E.2. 10 Year Windows over 1963–2019

Appendix E.3. 15 Year Windows over 1963–2019

Appendix E.4. 20 Year Windows over 1963–2019

Appendix E.5. 25 Year Windows over 1963–2019

Appendix E.6. 40 Year Windows over 1963–2019

Appendix E.7. 50 Year Windows over 1963–2019

Appendix E.8. 5 Year Windows over 1926–2019

Appendix E.9. 10 Year Windows over 1926–2019

Appendix E.10. 15 Year Windows over 1926–2019

Appendix E.11. 20 Year Windows over 1926–2019

Appendix E.12. 25 Year Windows over 1926–2019

Appendix E.13. 40 Year Windows over 1926–2019

Appendix E.14. 50 Year Windows over 1926–2019

Available on Supplementary Materials.

Appendix F. Summary Statistics

Appendix F.1. 5 Year Windows over 1963–2019
Appendix F.2. 10 Year Windows over 1963–2019
Appendix F.3. 15 Year Windows over 1963–2019
Appendix F.4. 20 Year Windows over 1963–2019
Appendix F.5. 25 Year Windows over 1963–2019
Appendix F.6. 40 Year Windows over 1963–2019
Appendix F.7. 50 Year Windows over 1963–2019
Appendix F.8. 5 Year Windows over 1926–2019
Appendix F.9. 10 Year Windows over 1926–2019
Appendix F.10. 15 Year Windows over 1926–2019
Appendix F.11. 20 Year Windows over 1926–2019
Appendix F.12. 25 Year Windows over 1926–2019
Appendix F.13. 40 Year Windows over 1926–2019
Appendix F.14. 50 Year Windows over 1926–2019

Available on Supplementary Materials.

Notes

1 See, for instance, Cakici et al. (2013, eq. (4)) and Mosoeu and Kodongo (2022, eq. (2)).

2 Asymptotic versions are commonly employed or promoted. See, for instance, MacKinlay and Richardson (1991), Cochrane (2005, p. 234), Zaremba and Czapkiewicz (2017), Demaj et al. (2018), Belimam et al. (2018), Qin (2019), and Verbeek (2021, Sct. 2.3).

3 We acknowledge that it is difficult to take seriously the assumption of normality of returns—returns are bounded below by -100% due to limited liability in financial markets for publicly traded assets and returns are known to be heteroskedastic and dependent over time. Here we adopt the Gibbons et al. (1989) setting for comparison purposes and to develop small sample results. Knez and Ready (1997) develop some interesting approaches for factor model estimation allowing for non-normality.

4 For related analysis on an extension to the GRS test, see Kleibergen and Zhan (2020) and Kleibergen et al. (2023).

5 Cochrane (2005, eq. (12.6)) uses $\widetilde{\Omega}$ for Ω and $\widetilde{\Sigma}$ for Σ, but pre-multiplies by $\frac{T-N-L}{N}$, so that the resulting GRS statistic equals to $\widetilde{W}$ in this paper. The d.f. adjustment (or lack of it) in the estimators of Σ can be easily offset by pre-multiplying an appropriate factor, but this is not the case for Ω.

6 Perhaps another outcome of Cochrane (2005), the Stata packages pre-multiply the ratio $\frac{T-N-L}{N}$ used by Cochrane (2005), but fail to use the corresponding $\widetilde{\Sigma}$.

7 We need to point out that the argument here is based on an approximation, as $E(\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p)$ is a non-linear function of $\bar{r}_p$ and $\widetilde{\Omega}$. Moreover, $\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p$ may deviate from its mean for a particular sample. Therefore, it is entirely possible that the incorrect formula of the GRS statistic favors larger models in some cases.

8 See Tharyan (2009) and Ibert (2014).

9 For comparison, the generalized GRS statistic formula given in Cochrane (2005, p. 230) uses $\widetilde{\Omega}$ and $\widetilde{\Sigma}$, but Cochrane (2005) is careful to properly pre-multiply the ratio $\frac{T-N-L}{N}$ so that the statistic equals $\widetilde{W}$ in this paper.

10 This is because the F_{d_1,d_2} and the $\chi^2_{d_1}$ distributions are related through $d_1 F_{d_1,d_2} \xrightarrow{d.} \chi^2_{d_1}$ as $d_2 \to \infty$, where d_1 and d_2 are the degrees of freedom.

11 Several different statistics are all commonly used in empirical research. For example, the usual Wald statistic equals $T\left(1+\bar{r}_p'\widetilde{\Omega}^{-1}\bar{r}_p\right)^{-1}\hat{\delta}_0'\hat{\Sigma}^{-1}\hat{\delta}_0$, which deviates from $N\widetilde{W}$ by a factor of $\frac{T-L-1}{T-N-L}$. Another example is the χ^2 statistic formula in Cochrane (2005, p. 230), which deviates from $N\widetilde{W}$ by a factor of $\frac{T}{T-N-L}$. Both factors are larger than 1, meaning that both χ^2 statistics are larger than $N\widetilde{W}$ for any sample size, with Cochrane's (2005, p. 230) formula being the largest. Note that Cochrane (2005, p. 230), like Gibbons et al. (1989), only gives the formula for the $L = 1$ case, and here we refer to its generalized version for the $L \geq 1$ case.

12 In Section 3, we only report empirical results using the Wald statistic for the asymptotic χ^2 test. Since the formula in Cochrane (2005, p. 230) is even larger, it will lead to worse over-rejection.

13 We thank Ken French for making this valuable resource freely available.

References

Affleck-Graves, John, and Bill McDonald. 1989. Nonnormalities and tests of asset pricing theories. *The Journal of Finance* 44: 889–908. [CrossRef]

Alhomaidi, Asem, M. Kabir Hassan, William J. Hippler, and Abdullah Mamun. 2019. The impact of religious certification on market segmentation and investor recognition. *Journal of Corporate Finance* 55: 28–48. [CrossRef]

Alshammari, Saad, and Shingo Goto. 2022. What factors drive Saudi stock markets? Firm characteristics that attract retail trades. *International Review of Economics and Finance* 80: 994–1011. [CrossRef]

Anderson, Theodore Wilbur. 2003. *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken: Wiley-Interscience.

Baek, Seungho, and John F. O. Bilson. 2015. Size and value risk in financial firms. *Journal of Banking and Finance* 55: 295–326. [CrossRef]

Barillas, Francisco, and Jay Shanken. 2018. Comparing asset pricing models. *The Journal of Finance* 73: 715–54. [CrossRef]

Bartlett, Maurice. 1951. An Inverse Matrix Adjustment Arising in Discriminant Analysis. *The Annals of Mathematical Statistics* 22: 107–11. [CrossRef]

Belimam, Doha, Yong Tan, and Ghizlane Lakhnati. 2018. An empirical comparison of asset-pricing models in the Shanghai a-share exchange market. *Asia-Pacific Financial Markets* 25: 249–65. [CrossRef]

Cakici, Nusret, Frank J. Fabozzi, and Sinan Tan. 2013. Size, value, and momentum in emerging market stock returns. *Emerging Markets Review* 16: 46–65. [CrossRef]

Chiah, Mardy, Daniel Chai, Angel Zhong, and Song Li. 2016. A Better Model? An empirical investigation of the Fama-French five-factor model in Australia. *International Review of Finance* 16: 595–638. [CrossRef]

Choi, Seo Joon, Kanghyun Kim, and Sunyoung Park. 2020. Is systemic risk systematic? Evidence from the US stock markets. *International Journal of Finance and Economics* 25: 642–63. [CrossRef]

Cochrane, John. 2005. *Asset Pricing: Revised Edition*. Princeton: Princeton University Press.

Demaj, Arber, Bora Oskay, Betim Lushtaku, and Thedo Linssen. 2018. Asset Pricing Models and Anomalies: An Empirical Analysis. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3128562 (accessed on 4 April 2022).

Fama, Eugene F., and Kenneth R. French. 2015. A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116: 1–22. [CrossRef]

Fama, Eugene F., and Kenneth R. French. 2016. Dissecting Anomalies with a Five-Factor Model. *The Review of Financial Studies* 29: 69–103. [CrossRef]

Fama, Eugene F., and Kenneth R. French. 2018. Choosing factors. *Journal of Financial Economics* 128: 234–52. [CrossRef]

Ferson, Wayne E., and Stephen R. Foerster. 1994. Finite sample properties of the generalized method of moments in tests of conditional asset pricing models. *Journal of Financial Economics* 36: 29–55. [CrossRef]

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. 1989. A Test of the Efficiency of a Given Portfolio. *Econometrica* 57: 1121–52. [CrossRef]

Gonçalves, Sílvia, and Halbert White. 2002. The Bootstrap of the Mean for Dependent Heterogenous Arrays. *Econometric Theory* 18: 1367–84. [CrossRef]

Gonçalves, Sílvia, and Halbert White. 2005. Bootstrap Standard Error Estimates for Linear Regressions. *Journal of the American Statistical Association* 100: 970–79. [CrossRef]

Hanauer, Matthias X. 2020. A Comparison of Global Factor Models. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3546295 (accessed on 19 April 2022).

Harvey, Campbell R., and Yan Liu. 2021. Lucky factors. *Journal of Financial Economics* 141: 413–35. [CrossRef]

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

Ibert, Markus. 2014. *GRSTEST2: Stata Module to Implement the Gibbons, Ross, Shanken (1989) Test*. Statistical Software Components S457786. Boston: Boston College Department of Economics.

Kan, Raymond, Xiaolu Wang, and Xinghua Zheng. 2024. In-sample and out-of-sample Sharpe ratios of multi-factor asset pricing models. *Journal of Financial Economics* 155: 103837. [CrossRef]

Kim, Jae H. 2022. GRS.test: GRS Test for Portfolio Efficiency, Its Statistical Power Analysis, and Optimal Significance Level Calculation. R Package Version 1.2. Available online: https://CRAN.R-project.org/package=GRS.test (accessed on 1 April 2022).

Kleibergen, Frank, and Zhaoguo Zhan. 2020. Robust inference for consumption-based asset pricing. *The Journal of Finance* 75: 507–50. [CrossRef]

Kleibergen, Frank, Lingwei Kong, and Zhaoguo Zhan. 2023. Identification robust testing of risk premia in finite samples. *Journal of Financial Econometrics* 21: 263–97. [CrossRef]

Knez, Peter J., and Mark J. Ready. 1997. On the robustness of size and book-to-market in cross-sectional regressions. *The Journal of Finance* 52: 1355–82.

Kroencke, Tim A. 2017. Asset pricing without garbage. *The Journal of Finance* 72: 47–98. [CrossRef]

Leite, André Luis, Marcelo Cabus Klotzle, Antonio Carlos Figueiredo Pinto, and Aldo Ferreira da Silva. 2018. Size, value, profitability, and investment: Evidence from emerging markets. *Emerging Markets Review* 36: 45–59. [CrossRef]

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken. 2010. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics* 96: 175–94. [CrossRef]

MacKinlay, A. Craig, and Matthew P. Richardson. 1991. Using generalized method of moments to test mean-variance efficiency. *The Journal of Finance* 46: 511–27.

Merdad, Hesham Jamil, M. Kabir Hassan, and William J. Hippler, III. 2015. The Islamic risk factor in expected stock returns: An empirical study in Saudi Arabia. *Pacific-Basin Finance Journal* 34: 293–314. [CrossRef]
Mosoeu, Selebogo, and Odongo Kodongo. 2022. The Fama-French five-factor model and emerging market equity returns. *The Quarterly Review of Economics and Finance* 85: 55–76. [CrossRef]
Politis, Dimitris N., and Joseph P. Romano. 1994. The Stationary Bootstrap. *Journal of the American Statistical Association* 89: 1303–13. [CrossRef]
Qin, Rui. 2019. Study on Applicability of Fama-French Five-Factor Model in Chinese A-Share Market. Paper presented at the 2nd International Symposium on Social Science and Management Innovation (SSMI 2019), Xi'an, China, November 29–30; Amsterdam: Atlantis Press, pp. 491–500.
Rouwenhorst, K. Geert. 1999. Local return factors and turnover in emerging stock markets. *The Journal of Finance* 54: 1439–64. [CrossRef]
Sha, Yezhou, and Ran Gao. 2019. Which is the best: A comparison of asset pricing factor models in Chinese mutual fund industry. *Economic Modelling* 83: 8–16. [CrossRef]
Sharpe, William F. 1964. Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance* 19: 425–42.
Sharpe, William F. 1966. Mutual fund performance. *The Journal of Business* 39: 119–38. [CrossRef]
Teräsvirta, Timo, and Ilkka Mellin. 1986. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, 159–71.
Tharyan, Rajesh. 2014. *GRSTEST: Stata Module to Implement the Gibbons et al. (1989) Test in a Single-Factor or Multi-Factor Setting*. Statistical Software Components S457069. Boston: Boston College Department of Economics.
Verbeek, Marno. 2021. Panel Methods for Finance. In *Panel Methods for Finance*. Berlin: De Gruyter.
White, H. 2000. A Reality Check for Data Snooping. *Econometrica* 68: 1097–26.
Zaremba, Adam, and Anna Czapkiewicz. 2017. Digesting anomalies in emerging European markets: A comparison of factor pricing models. *Emerging Markets Review* 31: 1–15. [CrossRef]
Zhu, Mengnan. 2020. *GRSFTEST: Stata Module to Perform the Gibbons, Ross, Shanken Test of Mean-Variance Efficiency of Asset Returns*. Statistical Software Components S458828. Boston: Boston College Department of Economics.