A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory: A Reply
Author(s): Richard Roll and Stephen A. Ross
Source: *The Journal of Finance*, Vol. 39, No. 2 (Jun., 1984), pp. 347-350
Published by: Blackwell Publishing for the American Finance Association
Stable URL: http://www.jstor.org/stable/2327864
Accessed: 27/11/2009 07:42

THE JOURNAL OF FINANCE • VOL. XXXIX, NO. 2 • JUNE 1984

A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory: A Reply

RICHARD ROLL and STEPHEN A. ROSS*

THE PAPER BY Dhrymes, Friend, and Gultekin (DFG) in this issue criticizes empirical work on the Arbitrage Pricing Theory (APT) conducted by ourselves, (RR [3]), and by a number of others. We are honored that such eminent professors as DFG have taken an interest in our work. Unfortunately, the DFG paper contains analysis which might foster misconceptions in the casual reader. Our purpose here is to present a brief, nontechnical comment. We are confident that those who are inclined to study the DFG paper in detail will be able to discern its merit for themselves.

The introductory section and the concluding Section VIII of DFG stress three points: first, that the method of RR has "major" pitfalls and is "seriously" flawed; second, that individual factors should not be tested for their pricing influence; and third, that more than three to five factors can be found by increasing the size of the group analyzed. We will comment on the reasonableness of the second and third point in some detail. The first is a value judgment and, as such, cannot be disputed on logical grounds. We would be the last to object to a reasoned opinion judiciously drawn by dispassionate and unbiased scholars, but, of course, we do not agree with the particular opinion expressed by DFG.

In Sections I and II, DFG repeat the standard development of the APT and make much of the argument that the APT, as a null hypothesis in a statistical test, is independent of any rotation of the factors. They stress their view that the only meaningful tests are those of whether *any* factors are priced, "rather than those which test whether some of them are priced and others are not."

RR raise and analyze precisely this same point,[1] and, in fact, our awareness of this issue led us to conduct F-tests of the joint significance of all factor prices (see our Tables IV and VI). Despite the rotation problem, tests of individual factor pricing have meaning. Since the factors are extracted in the order of their importance in explaining the covariance matrix of returns, it is not only perfectly valid, it is also interesting to ascertain if they each have an influence upon pricing. The number of statistically significant *estimated* priced factors can be different than the number of true factors because of chance rotation in the

* University of California at Los Angeles and Yale University, respectively.

[1] The rotation problem was emphasized several different times in RR (cf. pp. 1083–84, 1089, 1099). Compare our language, "In particular, if G is an orthogonal transformation matrix . . . [and] . . . If B is to be estimated . . . then all transforms BG will be equivalent," (p. 1083), with their language ". . . B is identified by factor analytic procedures only to the extent of left multiplication by an orthogonal matrix," (p. 326).

sample. There is no reason, however, that the same chance event should happen in every sample.

We were reassured by the "proof" in DFG's Section II that tests of the intercept (and of extraneous variables such as the "own" variance) are not affected by the rotation problem. We agree and we conducted such tests in Sections III and IV, where we pointed out, *inter alia*, that "... there is one parameter, the intercept term ... which should be identical across groups, whatever the sample rotation of the generating factors" (p. 1099).

In Section III, DFG present what they apparently feel is a debilitating criticism. The central point seems to be that the RR grouping procedure, which factor analyzed subsets of the data, does not produce the same estimates as would be obtained if one were able to perform a factor analysis on the entire data set at once.

Different samples of data generally do result in different sample estimates. But the issue is not how close the estimates obtained by a grouping procedure are to the estimates one would obtain from a full factor analysis, rather the question is how far the estimates are from the true but unknown factor structure parameters. A grouping procedure obviously does not use the full information in the sample, but the resulting estimates are consistent, which is all that one can generally hope for in many statistical applications. Grouping is not only statistically sound, it also has one important advantage: it is the only *feasible* alternative (for computational reasons).

In discussing this issue, DFG say that, "In their paper, RR ... make only the caveat that ... it may be different factors that correspond to different groups. The remainder of their discussion appears to ignore this point ..." (p. 331). In fact, we explicitly dealt with this point in the last section (IV) of our paper entitled, "A Test for the Equivalence of Factor Structure Across Groups." This section begins with the statement, "One of the most troubling econometric problems in the two preceding sections was due to the technological necessity of splitting assets into groups ... there is no good way to ascertain whether the *same* three (or four) factors generate the returns in every group" (p. 1099). We go on to provide a test for factor structure equivalence across groups which, although perhaps not statistically powerful, gives no indication whatsoever that the groups *do* differ in the identity of their respective factors. We think this can hardly be characterized as "ignoring the point."

Perhaps some of the confusion can be traced to DFG's belief that "the factors postulated in such models [as the APT] are not *ab initio* tangible concrete entities" (p. 331). Apparently, DFG do not accept the APT's fundamental assumption that returns are generated by causative, concrete, economic forces (the $\tilde{\delta}_t$'s in our notation), which move all asset prices. Factor analysis is simply one possible tool by which the effects of these forces can be measured; but, in principle, they could be measured directly. Although factor analysis does not have to assume that there are "concrete, *ab initio*" entities, the APT *does* make this assumption and, after all, it is the APT which is being tested.

Perhaps DFG's confusion on this point is responsible for simply false statements such as, "It is quite important to realize that factor analyzing 30 security groups *is not equivalent in any way to factor analyzing a 240- or 1260-security*

group if we impose the condition of extracting the same number of factors . . ." (pp. 333–34, their emphasis, our double emphasis). If there actually are fewer than 30 "pervasive" factors generating returns, then factor analyzing groups of size thirty or more is equivalent *in every way* except statistical power and computational cost.[2]

In Section IV, DFG test for the one rather bizarre case in which sample estimates obtained from grouped and nongrouped data coincide, namely when the covariances between assets in different groups are zero. It must be stressed that the RR procedure does not, even implicitly, assume that these covariances are zero. To do so would be ridiculous since the asset groups are formed alphabetically! Rather, the grouping procedure of RR simply leaves these covariances undetermined. To observe that the RR procedure coincides with the full factor analysis when the groups are independent tells us nothing about the relevant properties of the grouping procedure itself.

In Section V, DFG perform tests on the number of factors and obtain results different from those of RR. Like them, we also do not understand why their results differ from our own.[3] They do, however, cite the paper by Cho et al. [1] as being supportive of their findings. In fact, Cho et al. [1] report that in simulation tests the RR method proves to be robust in extracting the true factor structure. Of course, DFG's emphasis on the number of extracted factors is, at best, a secondary issue since the acid text of the APT is how well the factors explain pricing and, in particular, how well they fare against alternative hypotheses.

We want to take this opportunity to emphasize the irrelevance of the point that factor analysis extracts more factors with larger groups of securities or with larger time series sample sizes. This is exactly what one would expect; firms are in industries together, or inhabit the same region of the country, or produce substitute or complement products, or compete for the same labor, etc. There are many reasons why the number of *nonpriced* factors will increase with the group size. To illustrate, suppose that a group of 30 securities contains just one cosmetics company. Factor analysis produces, say, three significant factors. Now add a 31st company, a second cosmetics producer. If the time series sample is large enough, we would certainly anticipate finding a fourth significant factor, a factor for the cosmetics industry.[4] But, this is irrelevant for asset pricing theory. Since investors can diversify across industries, and since the cosmetics factor is not pervasive, it

[2] On page 333, DFG make the unbelievable assertion that a particular test of the importance of using grouped subsets has *infinite* statistical power. Using their notation, suppose $\hat{B}_i'$ is an estimated matrix of factor loadings obtained by analyzing all assets at once and $\hat{B}_i'$ is an estimate of the same matrix obtained from analyzing the assets in groups. Then according to DFG, "the length (inner product) of the columns of $\hat{B}_i'$ and $\hat{B}_i'$ must be identical. *This identity is not a matter of statistical significance and departures from it (beyond round-off errors) cannot be attributed to sampling errors,* . . . . Consequently, there is no such simple relationship linking the 'factor loadings' obtained from . . . [grouped and all data]. Thus, the RR procedure is not equivalent to testing the universe of securities" (pp. 333–34, their emphasis)!

[3] We doubt, however, that the difference is due to "the greater precision (sic) of our [DFG's] computer software (SAS)," pp. 338–39.

[4] Statistical power is why the number of detected factors should increase with the time series sample size.

will not be priced, i.e., it will have no associated risk premium (or one that is immeasurably small). We expect there are as many factors as there are sets of assets (pairs, triplets, and so on), and that they could all be detected with a sufficiently powerful test; but almost all of them are diversifiable and thus are just as irrelevant as if the idiosyncratic disturbance term in the APT were *really* purely random. To argue that the detection of more factors with larger groups of assets is economically meaningful ignores the intuition of modern financial theory.

Section VI is entitled, "How Well Does the APT Model Explain Daily Returns?" The purpose of asset pricing theories, APT or others, is to explain expected returns; actual returns are determined by unexpected as well as expected forces. As a consequence, a test of the "goodness-of-fit" between "predicted" and "actual" returns tells us more about the noisy nature of stock returns than about the ability of any particular model to explain expected returns. DFG report an improvement in explaining actual returns as the number of factors increases. But this is to be expected even when the factors are *not* priced. For instance, an industry factor such as the cosmetics factor mentioned above will improve the R^2 in a time series market model.

We cannot fully discuss the tests of Section VII since they are not reported in full, but it is interesting to note that DFG adopted tests "like those used by RR," even though such procedures are "subject to the basic limitations ... discussed earlier in the paper" (p. 345). Despite these alleged limitations, however, DFG rely on them to produce results which they interpret as "... inconsistent with the APT model" (p. 345). So, having spent their entire paper criticizing the RR test procedures, DFG finally report results for which the tests are apparently satisfactory.

In summary, the RR paper was certainly not the definitive test of the APT; it was merely a first step that others have extended by constructive suggestions for improving the testing procedures.

REFERENCES

1. D. Chinhyung Cho, Edwin J. Elton, and Martin J. Gruber. "On the Robustness of the Roll and Ross APT Methodology." *Journal of Financial and Quantitative Analysis*, forthcoming.
2. Phoebus J. Dhrymes, Irwin Friend, and N. Bulent Gultekin. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory." *Journal of Finance* 39 (June 1984), 323–46.
3. Richard Roll and Stephen A. Ross. "An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance* 35 (December 1980) 1073–1103.