A Possible Explanation of the Small Firm Effect

RICHARD ROLL*

ABSTRACT

Recent empirical studies have found that small listed firms yield higher average returns than large firms even when their riskiness is equal. The riskiness of small firms, however, has been improperly measured. Apparently, the error is due to auto-correlation in portfolio returns caused by infrequent trading. Other anomalous predictors of risk-adjusted returns, such as price/earnings ratios and dividend yields, may also derive some of their apparent power from this spurious source.

I. Introduction

The "small-firm" effect has recently received wide attention in both the financial and the academic literature. Small firms seem to have larger average returns than large firms even after adjusting for risk.

The two most sophisticated papers are by Reinganum [15, 14] which differ in that [14] uses the capital asset pricing model (CAPM) to measure risk while [15] uses the arbitrage pricing model (APT). The results were very similar. Market capitalization is a significant predictor of average return and the effect is large; about 18 percent per annum in excess return for the decile of smallest firms.

The size effect has been noted by others (e.g., Banz [2]) and the potential for apparent profit has not been lost on the investment community (cf. Fortune, 25 Feb. 1980, pp. 153, 156). However, most scholars and practitioners realized that the strong and persistent difference in average returns between small and large firms probably meant that small firms really were riskier; but, evidently, that risk measures were incomplete and that the CAPM and APT were at best misspecified and at worst simply false. Indeed, Reinganum regards his work as tests of the CAPM and the APT.

Before concluding, however, that the CAPM and APT are false, there is an econometric problem that should be investigated. This problem has the potential to explain the small firm effect. Because small firms are traded less frequently, risk measures obtained from short interval returns data (such as daily), seriously understate the actual risk from holding a small firm portfolio, whatever the model investors use to assess risk.

II. A simple experiment concerning the size effect on risk

Data for the market capitalizations of individual firms are available but are difficult to use because of the tedious programming involved. Fortunately, there

* Graduate School of Management, University of California, Los Angeles, Los Angeles, CA 90024  
(213) 825-6118

879
Table I
Equally-Weighted Index (E) and Standard & Poor’s 500 Index (S) According to Holding Period (Continuously-Compounded Returns)
(1962–1977)

| Holding Period in Trading Days (Returns) | Sample Size | Difference in Mean Returns \(E - S\) (%/Annum) | \(\hat{\beta}_{E|S}\) (Std. Err.) | \(\sigma^2_{E}/\sigma^2_{S}\) |
|------------------------------------------|-------------|-----------------------------------------------|---------------------------------|-------------------|
| 1 (Daily)                                | 388        | 12.56                                         | .865                            | .879              |
| 5 (Weekly)                               | 776        | 12.53                                         | .861                            | 1.06              |
| 10 (Bi-Weekly)                           | 388        | 12.53                                         | .845                            | 1.16              |
| 21 (Monthly)                             | 184        | 12.63                                         | .862                            | 1.25              |
| 42 (Bi-Monthly)                          | 92         | 12.63                                         | .825                            | 1.36              |
| 63 (Quarterly)                           | 61         | 12.30                                         | .842                            | 1.39              |
| 126 (Semi-Annual)                        | 30         | 12.20                                         | .834                            | 1.48              |

* Simple ordinary least squares “beta” of equally-weighted on S & P 500.
* Ratio of total return variances, equally-weighted to S & P 500.

is one set of data which is widely available and easy to use: the time series of returns for value-weighted and equally-weighted indexes. A value-weighted index such as the S&P 500 is obviously more heavily invested in large firms than is an equally-weighted index. Thus, comparing the behavior of two such indexes will enable us to study, with very little effort, the size effect.

Table I provides some relevant information for the S&P 500 and an equally-weighted index of New York and American listed common stocks for the period July 1962 through December 1977. As the table shows, the average returns are over 12 percent per annum higher for the equally-weighted index. Moreover, based on daily data, there is little, if any, reason to conclude that the riskinesses of the indexes are different. Their variances of total returns are approximately equal while the simple OLS beta of the equally-weighted index on the S&P 500 is actually significantly below unity based on a standard \(t\)-test.\(^1\) This agrees with the results of Reinganum that small firms have higher returns even when their measured risk is no greater than that of large firms.

Of course, neither the variance of a well-diversified portfolio nor the beta of one index on another is necessarily the correct risk measure. For our purpose, however, the precision of risk measure is less important than its alteration in response to using longer data intervals. In Table I, notice that when weekly, bi-weekly, monthly, bi-monthly, quarterly and semi-annual returns are employed,

\(^1\) The measured beta is .879 and its standard error is .00857. A simple \(t\)-test for significant difference from unity has the value \((1 - .879) / .00857 = 14.1\). However, we shall see later that this overstates the significance of beta’s difference from unity.
the correlation coefficient of returns stays about the same. In contrast, the beta and the ratio of total variances increase uniformly and materially. The daily variance of an equally-weighted portfolio is only six percent larger than the daily variance of the S&P 500, but the variance of semi-annual returns is 217 percent larger! Similarly, the simple beta increases from .879 to 1.48.

Very similar results are obtained when using the value weighted index of NYSE and AMEX listed securities in place of the S&P 500. The mean returns differ by only about 8.7 percent (probably because the S&P 500 does not include dividends) but the daily to semi-annual changes in beta and in variance ratio are .893 to 1.48 and 1.09 to 2.72 respectively.

Based on this evidence, no practical investor who is obliged to hold his portfolio for more than a day would regard a well-diversified small-firm portfolio as equal in risk to a similarly well-diversified large-firm portfolio.

To ascertain whether the error in risk assessment from using daily data is sufficient to explain the entire observed difference in average returns, we can deduce the implied market price of risk and judge whether it is in a reasonable range. For example, assuming that the value-weighted index is mean/variance efficient and the true $\beta_{RI}$ is about 1.5, the value-weighted index's implied expected return exceeds its zero-beta portfolio expected return by $8.7/1.5 = 5.8$ percent per annum. An alternative estimate can be obtained by assuming that the equally-weighted index is perfectly diversified and that expected return is proportional to standard deviation. This yields an estimate of the expected return on the value weighted index in excess of the risk-free rate of $8.7/(\sqrt{2.72} - 1) = 13.4$ percent per annum.

Both estimates rely on several rather dubious assumptions and both must be regarded as very rough indeed, particularly in view of the difficulty in detecting true differences in expected returns. Nevertheless, they seem too high and we must therefore conclude tentatively that perhaps only part of the observed risk-adjusted excess returns related to size can be explained by mis-assessment of risk. Further detailed study with individual firms is required to determine whether the size effect is entirely or just largely spurious.

III. The Cause of Risk Alteration with Data Interval is Serial Dependence in Returns

Why does the equally-weighted index display such dramatic increases in relative variance with lengthening data interval? The answer is simple: the daily returns of the equally-weighted index are more auto-correlated than those of the S&P 500. An auto-regression of daily returns is reported in Table II. The F-test of serial dependence yields an F value of 48.8 for the equally-weighted index and 13.9 for the S&P 500. Both values indicate highly significant auto-correlation but the equally-weighted index displays much more positive dependence and this causes a greater downward bias in the daily returns' variance.

\[\text{See Brown and Warner [6].}\]
Since the returns are continuously compounded, daily returns can be added to obtain longer interval returns\(^3\) and it is easy to show that the variance \(\sigma^2 = \sigma^2 [n + 2(n - 1)\rho_1 + 2(n - 2)\rho_2 + 2(n - 3)\rho_3 + \ldots]\) by

\[
\sigma^2 = \sigma^2 \left[ n + 2(n - 1)\rho_1 + 2(n - 2)\rho_2 + 2(n - 3)\rho_3 + \ldots \right] \tag{1}
\]

where \(\rho_j\) is the auto-correlation coefficient for a lag of \(j\) days.

If the auto-correlation coefficients of the equally-weighted portfolio are generally larger (algebraically) than those of the same lag for the S&P 500, the daily variance of returns will be more downward biased for the smaller firm (i.e., equally-weighted) portfolio. As Table II reports, the daily auto-correlation coefficients for the equally-weighted index are larger than those for the S&P 500 at every lag out to at least 21 days. In fact, the auto-correlation coefficients are all positive for the equally-weighted index while 10 of 21 are negative for the S&P 500. This pattern is maintained for even longer lags.\(^4\)

Further suspicion that auto-correlation may be a culprit is suggested by two separate clues in the Reinganum and Banz papers. In Reinganum [15] auto-correlations of excess returns are reported for ten portfolios ranked by size. In the middle portfolios whose constituent securities are medium-size firms, there is virtually no auto-correlation. In the extremes, portfolios containing both the largest and smallest firms display significant and positive auto-correlation up to three daily lags. Reinganum's small-firm portfolio auto-correlation is similar to the equally-weighted portfolio auto-correlation reported here. His large-firm portfolio auto-correlation can be explained by remembering that the returns are excess returns relative to a control portfolio which is itself equally weighted. Thus, auto-correlation is induced in the large firm excess return by the equally-weighted control portfolio even though the large-firm portfolio probably has a low degree of serial dependence in its own returns. As for the middle-size firm excess returns, there is no observed auto-correlation because the two positive serial dependence terms (one for each component) and the two positive serial cross-dependence terms are just offsetting.\(^5\)

The clue in Banz's study is discerned by relating it to Reinganum's. Banz used monthly data, so his implied small-firm premium after risk adjustment should be smaller than Reinganum's, who used daily beta. The reason for this can be understood by reference to Table I. Using monthly data (21 trading days), the

\(^3\) The same empirical results obtain for discretely-compounded returns but the algebra in this paragraph is more complex in that case.

\(^4\) A similar auto-correlation pattern as the S&P 500 is displayed by the value-weighted index of NYSE and AMEX stocks. The F-test of serial dependence has a value of 17.9 and eight of the first 21 auto-correlation coefficients are negative. The first three are .278, .069, .0406. The value-weighted index has a bit more weighting in small stocks than the S&P 500 and therefore, as would be expected, it is slightly closer to the equally-weighted index.

\(^5\) In other words, the excess return is \(r_t = R_{pt} - R_{ct}\), where \(p\) indicates the portfolio under study and \(c\) indicates the control portfolio. The serial covariance in excess returns, \(\text{Cov}(r_t, r_{t-1})\), for example, is composed of \(\text{Cov}(R_{pt}, R_{pt-1}) + \text{Cov}(R_{ct}, R_{ct-1}) - \text{Cov}(R_{pt}, R_{ct-1}) - \text{Cov}(R_{ct}, R_{pt-1})\). When trading frequency is different for \(p\) and \(c\), one of the first two terms dominates. Otherwise, the four mutually cancel.
Table II

Auto-Correlation of Returns on S & P 500 (S) and Equally-Weighted (E) Indexes (1962-1977)

<table>
<thead>
<tr>
<th>Lag (Days)</th>
<th>S Auto-Correlation Coefficient</th>
<th>E Auto-Correlation Coefficient</th>
<th>S Auto-Regressiona Coefficient (t-value)</th>
<th>E Auto-Regressiona Coefficient (t-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.241</td>
<td>.326</td>
<td>.259</td>
<td>.458</td>
</tr>
<tr>
<td></td>
<td>(16.0)</td>
<td>(28.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.0055</td>
<td>.126</td>
<td>-.0787</td>
<td>-.131</td>
</tr>
<tr>
<td></td>
<td>(-.4.72)</td>
<td>(-7.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.0250</td>
<td>.147</td>
<td>.0463</td>
<td>.131</td>
</tr>
<tr>
<td></td>
<td>(2.77)</td>
<td>(7.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.0139</td>
<td>.144</td>
<td>-.00410</td>
<td>.0177</td>
</tr>
<tr>
<td></td>
<td>(-.0245)</td>
<td>(.962)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-.0151</td>
<td>.0985</td>
<td>-.00242</td>
<td>.0962</td>
</tr>
<tr>
<td></td>
<td>(-.144)</td>
<td>(2.01)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P value for significance of all 21 lagged terms: F_S = 48.8, F_E = 13.9. Degrees of freedom = (21,3853).
Percentage of auto-correlation coefficients algebraically larger for equally weighted index = 100%. (21 out of 21).
a From a regression with 21 lagged trading days (up to one month). Coefficients for lags greater than 5 days not reported for reasons of space. They are generally not statistically significant.

The auto-correlation effect has already been reflected partially in the betas and in total variances. However, it has not been fully incorporated compared to returns as long as six months. Thus, a small-firm portfolio constructed to have CAPM market-level risk (β = 1) will have a lower observed premium with monthly than with daily data.

Banz reported a difference between two CAPM β risk-adjusted portfolios, one with the 50 smallest NYSE-listed firms and the other with the 50 largest, of 1.01 percent per month, slightly more than twelve percent per annum (Banz [2], Table III). In Reinganum [15, Table VIII], the average difference in mean excess return between the smallest firm portfolio and the largest firm portfolio was .000500 + .000343 per trading day or about 21 percent per annum (assuming 252 trading days per year). The difference accords rather nicely with what might be predicted from Table I as being the auto-correlation effect's difference for daily versus monthly returns. Notice that the monthly $\beta_{BS}$ is somewhat more than halfway between the daily and the semi-annual $\beta$'s.

This may be overstating the case somewhat since Reinganum also included AMEX-listed and thus smaller firms. On the other hand, the case may be understated because of Reinganum's later calendar period which coincided with more active trading and because Reinganum had more firms per portfolio.

a When Banz used the 10 smallest and largest firms instead of the 50 largest and smallest, his estimated difference in returns went up to over 18 percent per annum. This suggests that the Reinganum-Banz difference is actually understated.
IV. Trading Frequency, Auto-Correlation, Riskiness

Small-firm portfolios have higher auto-correlation of returns because their constituent securities are less-frequently traded. The effect is easy to see when an entire day passes without a trade; then that day's implicit return will be recorded on the day when its first subsequent trade takes place. This return is correlated, of course, with the returns of other firms which did register trades on the first day. The auto-correlation thereby induced in a portfolio of such securities is completely spurious and is simply the result of a defect in our record of prices. A similar spurious auto-correlation is induced even if firms trade every day but not continuously. The longer the average time between trades, the greater the induced auto-correlation in portfolios of such firms.\(^7\)

At first, one might think there is a knotty theoretical question of how risk is related to the investment "horizon." A moment's reflection, however, reveals that no such problem is present. The true riskiness is exactly the same for all data intervals; it is simply underestimated for the shorter ones. Most of our statistical measures are based on the assumption of a random sample in which there is no dependence among sample observations. With auto-correlation present, however, the sample observations are not independently distributed.

The impact on measures of risk is easy to understand intuitively. Take the estimated variance, for instance. Its sampling properties are related to the distribution of a sum of squared independent standardized variates. If the underlying distribution is Gaussian, the sum of \(n\) squared standardized independent observations has a \(\chi^2\) distribution with \(n\) degrees of freedom. The expected value of this sum is also \(n\). With positive dependence among the observations, however, there are (intuitively speaking) actually fewer than \(n\) degrees of freedom in the sample and the expected value of the sample variance is less than the population variance.

This was explained several years ago by Fisher [8]. More recently, in their detailed analysis of noncontinuous trading, Scholes and Williams [17] noted

"That measured variances for daily returns on large portfolios typically understate true variances. For portfolios more heavily weighted with securities trading on average less frequently,—e.g., an equally-weighted portfolio—these effects are even more pronounced." (p. 314).

A similar statement appears in the introductory section of Dimson [7]. Evidently, the magnitude of these effects and their application to the apparent return premia of small firms have been overlooked, with some notable exceptions such as Saloner and Strebel [16].

Some direct empirical information on the impact can be obtained by using

\(^7\) Positive serial dependence is induced in portfolio returns by nonsynchronous trading whereas the dependence generally observed in individual security returns is negative (and very small), [cf. Scholes and Williams].

This suggests rather strongly that portfolio return dependence is indeed spurious due to nonsynchronous trading, and is not caused by genuine dependence in individual returns.
daily data and Dimson's [7] method of estimating beta. His estimator is obtained simply by calculating a multiple regression of the equally-weighted return on contemporaneous, leading and lagged S&P 500 returns and then summing the slope coefficients.

Dimson suggests that "When shares are being regressed on an index which is composed of large companies and/or which is value weighted, it is the lagged coefficients which are of importance," (p. 205). Accordingly, I used 21 lags (up to one month) and only five leads (up to one week). The Dimson estimator of beta thereby obtained was 1.44. This is quite close to the beta obtained with semi-annual data and reported in Table I.

Interestingly, in the multiple regression, none of the five leading coefficients were significant and all five had $t$-ratios below 1.0. The contemporaneous term had a $t$-ratio of 100.9, the first lagged $t$-ratio was 23.8, and the next ten were as follows: 1.28, 8.29, 7.55, 6.70, 2.14, 3.10, 2.89, 1.88, 3.35, and 1.54. Even further lags were sometimes significant. There seems to be a lot of infrequent trading revealed in these data.

The Dimson method is particularly useful for checking on whether the pattern displayed in Table I is due to some peculiarity in the overall sample period which is not revealed in subperiods and, by inference, would not be present in the future. Note that the simple expedient of computing betas and variances for long holding periods, as was reported in Table I, rapidly runs out of observations as the subperiods become shorter. For instance, there are only two semi-annual returns per calendar year; so the ratio of semi-annual equally-weighted to value-weighted variances would be subject to a large sampling error. The Dimson method, however, uses daily data and thus retains its ability to detect non-synchronous trading with calendar samples of short duration.

Table 3 gives the ordinary beta and the Dimson beta of the equally-weighted index on the S&P 500 for each full calendar year present in the overall sample. There is some sampling variability in both betas over time but in every year, the Dimson beta is larger; it is significantly larger since the standard error of the OLS beta ranges from .025 to .035. The ratio of Dimson to OLS beta is never less than 1.25 and was as high as 2.37 (in 1976). This evidence indicates that the nonsynchronous trading effect is persistent.

One might be tempted to infer that the effect has increased over time because the largest differences in the two betas occur during the last six of the fifteen years. Moreover, there is some a priori plausibility to the increase. Exchange listing requirements have not changed very much in nominal terms and the number of listed issues has grown. Since the S&P 500 is still composed only of the largest 500 securities, perhaps the difference in nonsynchronicity has also grown between it and an equally-weighted index of a larger number of relatively small firms. A more sophisticated test is necessary, of course, before this pattern can be considered definitive.

* Dimson's method, unlike Scholes's and Williams's, can be applied easily and directly to higher-order serially dependent returns. Thus, the Dimson method was chosen here because the data display higher than first-order dependence. Fowler and Rocke [9] argue that Dimson's method is biased. However, I have been unable to ascertain whether they are correct and whether the bias is material, if it exists. The empirical results below indicate that the bias is probably not very large.
V. Infrequent Trading and Other "Anomalies" in Average Returns

If the size effect on risk-adjusted returns can be explained by auto-correlation-induced biases in measured risk, other puzzling empirical findings might also be susceptible to explanations by the same phenomenon. Perhaps the two most widely-known are the price/earnings (P/E) effect, e.g., Basu [3], and the dividend yield effect, e.g., Litzenberger and Ramaswamy [10], Ball [1] provides a survey of these and other anomalies.

Actually, the P/E need not be of much further concern after Reinganum [14]. He demonstrated convincingly that the P/E effect is attributable to P/E and size being strongly related. If both variables are used to predict risk-adjusted returns, size dominates completely and the P/E makes no additional contribution. This result is highly suggestive of auto-correlation as the root of both effects: since P/E and size are highly but not perfectly correlated, the one most closely related to infrequency of trading would have the greatest ability to predict risk-adjusted returns. Clearly, size is a better proxy for infrequent trading than P/E.

The dividend yield effect is an open issue because of the mixed empirical evidence. Although Litzenberger and Ramaswamy [10] find a strong positive effect of yield on risk-adjusted returns, Morgan [13] finds only a weak effect (which he suggests is due to nonstationarity); Black and Scholes [4] find no significant effect whatsoever, and Long [11] finds a negative effect in a very careful study with unusual data not subject to the infrequency of trading problem.

The question to be addressed here (but not answered definitely) is whether any of the positive effect findings might be explained at least partly by auto-
correlation in portfolio returns and concomitant underassessment of risk. When the NYSE and AMEX listed assets are considered, there is good reason to predict a rather complex and non-monotonic relation between dividend yield and trading frequency. At one end of the spectrum are zero yield stocks, most of which are small firms. Assets at the high end of the dividend yield spectrum are probably pseudoequities (such as leaseholds) and utilities. None is frequently traded relative to typical listed assets which pay positive but modest dividends.

A recent paper by Blume [5] supports this conjecture. He finds a positive relation between dividend yield and risk-adjusted return except for stocks which pay no dividends at all. This zero-dividend group also has a high average excess return after adjustment for CAPM β risk and the overall relationship between dividend yield and excess return is a lopsided U.

Dividend yield is also specific to industrial groupings. For example, electric utilities generally have high yields. Furthermore, trading frequency also is related to industry; oil firms are large companies (now) while retail stores and textile products are small. Thus, we should anticipate a complex relationship between dividend yield and trading frequency with a correspondingly complex relationship between yield and the bias in risk measures. Perhaps the mixed evidence on dividend yield is due partly to this complexity, with different statistical approaches having picked out or not picked out an overall relationship. At the very least, the potential for risk mis-assessment depending in some way on dividend yield is clearly present. Future studies may wish to make appropriate corrections to individual security risk parameters, à la Dimson, before proceeding to investigate the dividend yield effect.

VI. Conclusions

Trading infrequency seems to be a powerful cause of bias in risk assessments with short-interval data. Rather horrendous bias is induced in daily data and the bias is still large and significant with returns measured over intervals as long as one month.

The mis-assessment of risk has the potential to explain why small firms, low price/earnings ratio firms, and possibly high dividend yield firms display large excess returns (after adjustment for risk). Positive auto-correlation induced in portfolios of such firms because of infrequent trading results in downward biased measures of portfolio risk and corresponding overestimates of “risk adjusted” average returns.

Acknowledgement

Without implicating him in any remaining errors, the suggestions of the referee, Stephen J. Brown, are gratefully acknowledged.

5 Some and perhaps most of the positive dividend effect, particularly that found by Litzenberger/Ramaswamy, is due to the unexpurgated information content of dividends. See Miller and Scholes [12].
REFERENCES


13. Morgan, I.G., "Dividends and Capital Asset Prices," (School of Business, Queen’s University, Kingston, Ontario, June 1980).


15. Reinganum, Marc R., "A Simple Test of the Arbitrage Pricing Theory," (Graduate School of Business, University of Southern California, August, 1980).
