Journal of Financial Economics 9 (1981) 3–18. North-Holland Publishing Company

THE RELATIONSHIP BETWEEN RETURN AND MARKET VALUE OF COMMON STOCKS*

Rolf W. BANZ

Northwestern University, Evanston, IL 60201, USA

Received June 1979, final version received September 1980

This study examines the empirical relationship between the return and the total market value of NYSE common stocks. It is found that smaller firms have had higher risk adjusted returns, on average, than larger firms. This 'size effect' has been in existence for at least forty years and is evidence that the capital asset pricing model is misspecified. The size effect is not linear in the market value; the main effect occurs for very small firms while there is little difference in return between average sized and large firms. It is not known whether size *per se* is responsible for the effect or whether size is just a proxy for one or more true unknown factors correlated with size.

1. Introduction

The single-period capital asset pricing model (henceforth CAPM) postulates a simple linear relationship between the expected return and the market risk of a security. While the results of direct tests have been inconclusive, recent evidence suggests the existence of additional factors which are relevant for asset pricing. Litzenberger and Ramaswamy (1979) show a significant positive relationship between dividend yield and return of common stocks for the 1936–1977 period. Basu (1977) finds that price–earnings ratios and risk adjusted returns are related. He chooses to interpret his findings as evidence of market inefficiency but as Ball (1978) points out, market efficiency tests are often joint tests of the efficient market hypothesis and a particular equilibrium relationship. Thus, some of the anomalies that have been attributed to a lack of market efficiency might well be the result of a misspecification of the pricing model.

This study contributes another piece to the emerging puzzle. It examines the relationship between the total market value of the common stock of a firm and its return. The results show that, in the 1936–1975 period, the common stock of small firms had, on average, higher risk-adjusted returns

*This study is based on part of my dissertation and was completed while I was at the University of Chicago. I am grateful to my committee, Myron Scholes (chairman), John Gould, Roger Ibbotson, Jonathan Ingersoll, and especially Eugene Fama and and Merton Miller, for their advice and comments I wish to acknowledge the valuable comments of Bill Schwert on earlier drafts of this paper

than the common stock of large firms. This result will henceforth be referred to as the 'size effect'. Since the results of the study are not based on a particular theoretical equilibrium model, it is not possible to determine conclusively whether market value *per se* matters or whether it is only a proxy for unknown true additional factors correlated with market value. The last section of this paper will address this question in greater detail.

The various methods currently available for the type of empirical research presented in this study are discussed in section 2. Since there is a considerable amount of confusion about their relative merit, more than one technique is used. Section 3 discusses the data. The empirical results are presented in section 4. A discussion of the relationship between the size effect and other factors, as well as some speculative comments on possible explanations of the results, constitute section 5.

2. Methodologies

The empirical tests are based on a generalized asset pricing model which allows the expected return of a common stock to be a function of risk β and an additional factor ϕ, the market value of the equity.[1] A simple linear relationship of the form

$$E(R_i)=\gamma_0+\gamma_1\beta_i+\gamma_2[(\phi_i-\phi_m)/\phi_m], \tag{1}$$

is assumed, where

$E(R_i)$ = expected return on security i,
γ_0 = expected return on a zero-beta portfolio,
γ_1 = expected market risk premium,
ϕ_i = market value of security i,
ϕ_m = average market value, and
γ_2 = constant measuring the contribution of ϕ_i to the expected return of a security.

If there is no relationship between ϕ_i and the expected return, i.e., $\gamma_2=0$, (1) reduces to the Black (1972) version of the CAPM.

Since expectations are not observable, the parameters in (1) must be estimated from historical data. Several methods are available for this purpose. They all involve the use of pooled cross-sectional and time series regressions to estimate γ_0, γ_1, and γ_2. They differ primarily in (a) the assumption concerning the residual variance of the stock returns (homoscedastic or heteroscedastic in the cross-sectional), and (b) the treatment of the

[1] In the empirical tests, Φ_i and Φ_m are defined as the market proportion of security i and average market proportion, respectively. The two specifications are, of course, equivalent.

errors-in-variables problem introduced by the use of estimated betas in (1). All methods use a constrained optimization procedure, described in Fama (1976, ch. 9), to generate minimum variance (m.v.) portfolios with mean returns γ_i, $i=0,\ldots,2$. This imposes certain constraints on the portfolio weights, since from (1)

$$E(R_p)\equiv\gamma_i=\gamma_0\sum_j w_j+\gamma_1\sum_j w_j\beta_j + \gamma_2\left[\left(\sum_j w_j\phi_j-\phi_m\sum_j w_j\right)\Big/\phi_m\right],\qquad i=0,\ldots,2, \tag{2}$$

where the w_j are the portfolio proportions of each asset j, $j=1,\ldots,N$. An examination of (2) shows that $\hat{\gamma}_0$ is the mean return of a standard m.v. portfolio ($\sum_j w_j=1$) with zero beta and $\phi_p\equiv\sum_j w_j\phi_j=\phi_m$ [to make the second and third terms of the right-hand side of (2) vanish]. Similarly, $\hat{\gamma}_1$ is the mean return on a zero-investment m.v. portfolio with beta of one and $\phi_p=0$, and $\hat{\gamma}_2$ is the mean return on a m.v. zero-investment, zero-beta portfolio with $\phi_p=\phi_m$. As shown by Fama (1976, ch. 9), this constrained optimization can be performed by running a cross-sectional regression of the form

$$R_{it}=\gamma_{0t}+\gamma_{1t}\beta_{it}+\gamma_{2t}[(\phi_{it}-\phi_{mt})/\phi_{mt}]+\varepsilon_{it},\qquad i=1,\ldots,N, \tag{3}$$

on a period-by-period basis, using estimated betas $\hat{\beta}_{it}$ and allowing for either homoscedastic or heteroscedastic error terms. Invoking the usual stationarity arguments the final estimates of the gammas are calculated as the averages of the T estimates.

One basic approach involves grouping individual securities into portfolios on the basis of market value and security beta, reestimating the relevant parameters (beta, residual variance) of the portfolios in a subsequent period, and finally performing either an ordinary least squares (OLS) regression [Fama and MacBeth (1973)] which assumes homoscedastic errors, or a generalized least squares (GLS) regression [Black and Scholes (1974)] which allows for heteroscedastic errors, on the portfolios in each time period.[2] Grouping reduces the errors-in-variables problem, but is not very efficient because it does not make use of all information. The errors-in-variables problem should not be a factor as long as the portfolios contain a reasonable number of securities.[3]

Litzenberger and Ramaswamy (1979) have suggested an alternative method which avoids grouping. They allow for heteroscedastic errors in the cross-section and use the estimates of the standard errors of the security

[2]Black and Scholes (1974) do not take account of heteroscedasticity, even though their method was designed to do so.

[3]Black, Jensen and Scholes (1972, p. 116).

betas as estimates of the measurement errors. As Theil (1971, p. 610) has pointed out, this method leads to unbiased maximum likelihood estimators for the gammas as long as the error in the standard error of beta is small and the standard assumptions of the simple errors-in-variables model are met. Thus, it is very important that the diagonal model is the correct specification of the return-generating process, since the residual variance assumes a critical position in this procedure. The Litzenberger–Ramaswamy method is superior from a theoretical viewpoint; however, preliminary work has shown that it leads to serious problems when applied to the model of this study and is not pursued any further.[4]

Instead of estimating equation (3) with data for all securities, it is also possible to construct arbitrage portfolios containing stocks of very large and very small firms, by combining long positions in small firms with short positions in large firms. A simple time series regression is run to determine the difference in risk-adjusted returns between small and large firms. This approach, long familiar in the efficient markets and option pricing literature, has the advantage that no assumptions about the exact functional relationships between market value and expected return need to be made, and it will therefore be used in this study.

3. Data

The sample includes all common stocks quoted on the NYSE for at least five years between 1926 and 1975. Monthly price and return data and the number of shares outstanding at the end of each month are available in the monthly returns file of the Center for Research in Security Prices (CRSP) of the University of Chicago. Three different market indices are used; this is in response to Roll's (1977) critique of empirical tests of the CAPM. Two of the three are pure common stock indices — the CRSP equally- and value-weighted indices. The third is more comprehensive: a value-weighted combination of the CRSP value-weighted index and return data on corporate and government bonds from Ibbotson and Sinquefield (1977) (henceforth 'market index').[5] The weights of the components of this index are derived from information on the total market value of corporate and government bonds in various issues of the *Survey of Current Business* (updated annually) and from the market value of common stocks in the CRSP monthly index file. The stock indices, made up of riskier assets, have both higher returns

[4]If the diagonal model (or market model) is an incomplete specification of the return generating process, the estimate of the standard error of beta is likely to have an upward bias, since the residual variance estimate is too large The error in the residual variance estimate appears to be related to the second factor. Therefore, the resulting gamma estimates are biased.

[5]No pretense is made that this index is complete, thus, the use of quotation marks It ignores real estate, foreign assets, etc ; it should be considered a first step toward a comprehensive index. See Ibbotson and Fall (1979)

and higher risk than the bond indices and the 'market index'.[6] A time series of commercial paper returns is used as the risk-free rate.[7] While not actually constant through time, its variation is very small when compared to that of the other series, and it is not significantly correlated with any of the three indices used as market proxies.

4. Empirical results

4.1. Results for methods based on grouped data

The portfolio selection procedure used in this study is identical to the one described at length in Black and Scholes (1974). The securities are assigned to one of twenty-five portfolios containing similar numbers of securities, first to one of five on the basis of the market value of the stock, then the securities in each of those five are in turn assigned to one of five portfolios on the basis of their beta. Five years of data are used for the estimation of the security beta; the next five years' data are used for the reestimation of the portfolio betas. Stock price and number of shares outstanding at the end of the five year periods are used for the calculation of the market proportions. The portfolios are updated every year. The cross-sectional regression (3) is then performed in each month and the means of the resulting time series of the gammas could be (and have been in the past) interpreted as the final estimators. However, having used estimated parameters, it is not certain that the series have the theoretical properties, in particular, the hypothesized beta. Black and Scholes (1974, p. 17) suggest that the time series of the gammas be regressed once more on the excess return of the market index. This correction involves running the time series regression (for $\hat{\gamma}_2$)

$$\hat{\gamma}_{2t} - R_{Ft} = \hat{\alpha}_2 + \hat{\beta}_2(R_{mt} - R_{Ft}) + \hat{\varepsilon}_{2t}. \tag{4}$$

It has been shows earlier that the theoretical β_2 is zero. (4) removes the effects of a non-zero $\hat{\beta}_2$ on the return estimate $\hat{\gamma}_2$ and $\hat{\alpha}_2$ is used as the final estimator for $\widehat{\gamma_2 - R_F}$. Similar corrections are performed for γ_0 and γ_1. The

[6] Mean monthly returns and standard deviations for the 1926–1975 period are.

	Mean return	Standard deviation
'Market index'	0.0046	0.0178
CRSP value-weighted index	0.0085	0.0588
CRSP equally-weighted index	0.0120	0.0830
Government bond index	0.0027	0.0157
Corporate bond index	0.0032	0.0142

[7] I am grateful to Myron Scholes for making this series available. The mean monthly return for the 1926–1975 period is 0.0026 and the standard deviation is 0.0021.

derivations of the $\hat{\beta}_i$, $i=0,\ldots,2$, in (4) from their theoretical values also allow us to check whether the grouping procedure is an effective means to eliminate the errors-in-beta problem.

The results are essentially identical for both OLS and GLS and for all three indices. Thus, only one set of results, those for the 'market index' with GLS, is presented in table 1. For each of the gammas, three numbers are reported: the mean of that time series of returns which is relevant for the test of the hypothesis of interest (i.e., whether or not $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are different from the risk-free rate and the risk premium, respectively), the associated t-statistic, and finally, the estimated beta of the time series of the gamma from (4). Note that the means are corrected for the deviation from the theoretical beta as discussed above.

The table shows a significantly negative estimate for γ_2 for the overall time period. Thus, shares of firms with large market values have had smaller returns, on average, than similar small firms. The CAPM appears to be misspecified. The table also shows that γ_0 is different from the risk-free rate. As both Fama (1976, ch. 9) and Roll (1977) have pointed out, if a test does not use the true market portfolio, the Sharpe–Lintner model might be wrongly rejected. The estimates for γ_0 are of the same magnitude as those reported by Fama and MacBeth (1973) and others. The choice of a market index and the econometric method does not affect the results. Thus, at least within the context of this study, the choice of a proxy for the market portfolio does not seem to affect the results and allowing for heteroscedastic disturbances does not lead to significantly more efficient estimators.

Before looking at the results in more detail, some comments on econometric problems are in order. The results in table 1 are based on the 'market index' which is likely to be superior to pure stock indices from a theoretical viewpoint since it includes more assets [Roll (1977)]. This superiority has its price. The actual betas of the time series of the gammas are reported in table 1 in the columns labeled $\hat{\beta}_i$. Recall that the theoretical values of β_0 and β_1 are zero and one, respectively. The standard zero-beta portfolio with return $\hat{\gamma}_0$ contains high beta stocks in short positions and low beta stocks in long positions, while the opposite is the case for the zero-investment portfolio with return $\hat{\gamma}_1$. The actual betas are all significantly different from the theoretical values. This suggests a regression effect, i.e., the past betas of high beta securities are overestimated and the betas of low beta securities are underestimated.[8] Past beta is not completely uncorrelated with the error of the current beta and the instrumental variable approach to the error-in-variables problem is not entirely successful.[9]

[8]There is no such effect for β_2 because that portfolio has both zero beta and zero investment, i.e., net holdings of both high and low beta securities are, on average, zero.

[9]This result is first documented in Brenner (1976) who examines the original Fama–McBeth (1973) time series of $\hat{\gamma}_{0t}$.

Table 1

Portfolio estimators for γ_0, γ_1 and γ_2 based on the 'market index' with generalized least squares estimation [a]

$$R_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\hat{\beta}_{it} + \hat{\gamma}_{2t}[(\phi_{it} - \phi_{mt})/\phi_{mt}]$$

Period	$\hat{\gamma}_0 - R_F$	$t(\hat{\gamma}_0 - R_F)$	$\hat{\beta}_0$	$\hat{\gamma}_1 - (R_M - R_F)$	$t(\hat{\gamma}_1 - (R_M - R_F))$	$\hat{\beta}_1$	$\hat{\gamma}_2$	$t(\hat{\gamma}_2)$	$\hat{\beta}_2$
1936–1975	0.00450	2.76	0.45	−0.00092	−1.00	0.75	−0.00052	−2.92	0.01
1936–1955	0.00377	1.66	0.43	−0.00060	−0.80	0.80	−0.00043	−2.12	0.01
1956–1975	0.00531	2.22	0.46	−0.00138	−0.82	0.73	−0.00062	−2.09	0.01
1936–1945	0.00121	0.30	0.63	−0.00098	−0.77	0.82	−0.00075	−2.32	−0.01
1946–1955	0.00650	2.89	0.03	−0.00021	−0.26	0.75	−0.00015	−0.65	0.06
1956–1965	0.00494	2.02	0.34	−0.00098	−0.56	0.96	−0.00039	−1.27	−0.01
1966–1975	0.00596	1.43	0.49	−0.00232	−0.80	0.69	−0.00080	−1.55	0.01

[a] $\hat{\gamma}_0 - R_F$ = mean difference between return on zero beta portfolio and risk-free rate. $\hat{\gamma}_1 - (R_M - R_F)$ = mean difference between actual risk premium ($\hat{\gamma}_1$) and risk premium stipulated by Sharpe–Lintner model $(R_M - R_F)$. $\hat{\gamma}_2$ = size premium. $\hat{\beta}_i$ = actual estimated market risk of $\hat{\gamma}_i$ (theoretical values. $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0$), all β_0, β_1 are significantly different from the theoretical values. $t(\ )$ = t-statistic.

The deviations from the theoretical betas are largest for the 'market index', smaller for the CRSP value-weighted index, and smallest for the CRSP equally-weighted index. This is due to two factors: first, even if the true covariance structure is stationary, betas with respect to a value-weighted index change whenever the weights change, since the weighted average of the betas is constrained to be equal to one. Second, the betas and their standard errors with respect to the 'market index' are much larger than for the stock indices (a typical stock beta is between two and three), which leads to larger deviations -- a kind of 'leverage' effect. Thus, the results in table 1 show that the final correction for the deviation of $\hat{\beta}_0$ and $\hat{\beta}_1$ from their theoretical values is of crucial importance for maket proxies with changing weights.

Estimated portfolio betas and portfolio market proportions are (negatively) correlated. It is therefore possible that the errors in beta induce an error in the coefficient of the market proportion. According to Levi (1973), the probability limit of $\hat{\gamma}_1$ in the standard errors-in-the-variables model is

$$\operatorname{plim} \hat{\gamma}_1 = \gamma_1/(1+(\sigma_u^2 \cdot \sigma_2^2)/D) < \gamma_1,$$

with

$$D = (\sigma_1^2 + \sigma_u^2) \cdot \sigma_2^2 - \sigma_{12}^2 > 0,$$

where σ_1^2, σ_2^2 are the variances of the true factors β and ϕ, respectively, σ_u^2 is the variance of the error in beta and σ_{12} is the covariance of β and ϕ. Thus, the bias in $\hat{\gamma}_1$ is unambiguously towards zero for positive γ_1. The probability limit of $\hat{\gamma}_2 - \gamma_2$ is [Levi (1973)]

$$\operatorname{plim} (\hat{\gamma}_2 - \gamma_2) = (\sigma_u^2 \cdot \sigma_{12} \cdot \gamma_1)/D.$$

We find that the bias in $\hat{\gamma}_2$ depends on the covariance between β and ϕ and the sign of γ_1. If σ_{12} has the same sign as the covariance between $\hat{\beta}$ and ϕ, i.e., $\sigma_{12} < 0$, and if $\gamma_1 > 0$, then $\operatorname{plim}(\hat{\gamma}_2 - \gamma_2) < 0$, i.e., $\operatorname{plim} \hat{\gamma}_2 < \gamma_2$. If the grouping procedure is not successful in removing the error in beta, then it is likely that the reported $\hat{\gamma}_2$ overstates the true magnitude of the size effect. If this was a serious problem in this study, the results for the different market indices should reflect the problem. In particular, using the equally-weighted stock index should then lead to the smallest size effect since, as was pointed out earlier, the error in beta problem is apparently less serious for that kind of index. In fact, we find that there is little difference between the estimates.[10]

[10]For the overall time period, $\hat{\gamma}_2$ with the equally-weighted CRSP index is -0.00044, with the value weighted CRSP index -0.00044 as well as opposed to the -0.00052 for the 'market index' reported in table 1 The estimated betas of $\hat{\gamma}_0$ and $\hat{\gamma}_1$ which reflect the degree of the error in beta problems are 0 07 and 0 91, respectively, for the equally-weighted CRSP index and 0 13 and 0.87 for the value-weighted CRSP index.

Thus, it does not appear that the size effect is just a proxy for the unobservable true beta even though the market proportion and the beta of securities are negatively correlated.

The correlation coefficient between the mean market values of the twenty-five portfolios and their betas is significantly negative, which might have introduced a multicollinearity problem. One of its possible consequences is coefficients that are very sensitive to addition or deletion of data. This effect does not appear to occur in this case: the results do not change significantly when five portfolios are dropped from the sample. Revising the grouping procedure — ranking on the basis of beta first, then ranking on the basis of market proportion — also does not lead to substantially different results.

4.2. A closer look at the results

An additional factor relevant for asset pricing — the market value of the equity of a firm — has been found. The results are based on a linear model. Linearity was assumed only for convenience and there is no theoretical reason (since there is no model) why the relationship should be linear. If it is nonlinear, the particular form of the relationship might give us a starting point for the discussion of possible causes of the size effect in the next section. An analysis of the residuals of the twenty-five portfolios is the easiest way to look at the linearity question. For each month t, the estimated residual return

$$\hat{\varepsilon}_{it} = R_{it} - \hat{\gamma}_{0t} - \hat{\gamma}_{1t}\hat{\beta}_{it} - \hat{\gamma}_{2t}[(\phi_{it} - \phi_{mt})/\phi_{mt}], \qquad i = 1, \ldots, 25, \tag{5}$$

is calculated for all portfolios. The mean residuals over the forty-five year sample period are plotted as a function of the mean market proportion in fig. 1. Since the distribution of the market proportions is very skewed, a logarithmic scale is used. The solid line connects the mean residual returns of each size group. The numbers identify the individual portfolios within each group according to beta, '1' being the one with the largest beta, '5' being the one with the smallest beta.

The figure shows clearly that the linear model is misspecified.[11] The residuals are not randomly distributed around zero. The residuals of the portfolios containing the smallest firms are all positive; the remaining ones are close to zero. As a consequence, it is impossible to use $\hat{\gamma}_2$ as a simple size premium in the cross-section. The plot also shows, however, that the misspecification is not responsible for the significance of $\hat{\gamma}_2$ since the linear model underestimates the true size effect present for very small firms. To illustrate this point, the five portfolios containing the smaller firms are

[11] The nonlinearity cannot be eliminated by defining ϕ_i as the log of the market proportion

deleted from the sample and the parameters reestimated. The results, summarized in table 2, show that the $\hat{\gamma}_2$ remain essentially the same. The relationship is still not linear; the new $\hat{\gamma}_2$ still cannot be used as a size premium.

Fig. 1 suggests that the main effect occurs for very small firms. Further support for this conclusion can be obtained from a simple test. We can regress the returns of the twenty-five portfolios in each result on beta alone and examine the residuals. The regression is misspecified and the residuals contain information about the size effect. Fig. 2 shows the plot of those residuals in the same format as fig. 1. The smallest firms have, on average, very large unexplained mean returns. There is no significant difference between the residuals of the remaining portfolios.

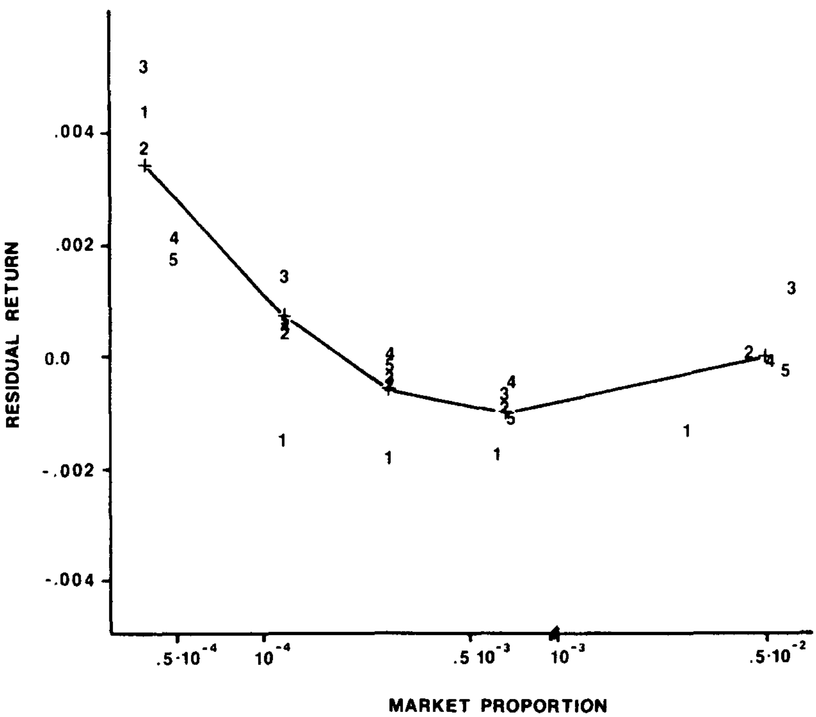

Fig. 1. Mean residual returns of portfolios (1936–1975) with equally-weighted CRSP index as market proxy. The residual is calculated with the three-factor model [eq. (3)]. The numbers 1,...,5 represent the mean residual return for the five portfolios within each size group (1: portfolio with largest beta,. .,5: portfolio with smallest beta) + represents the mean of the mean residuals of the five portfolios with similar market values.

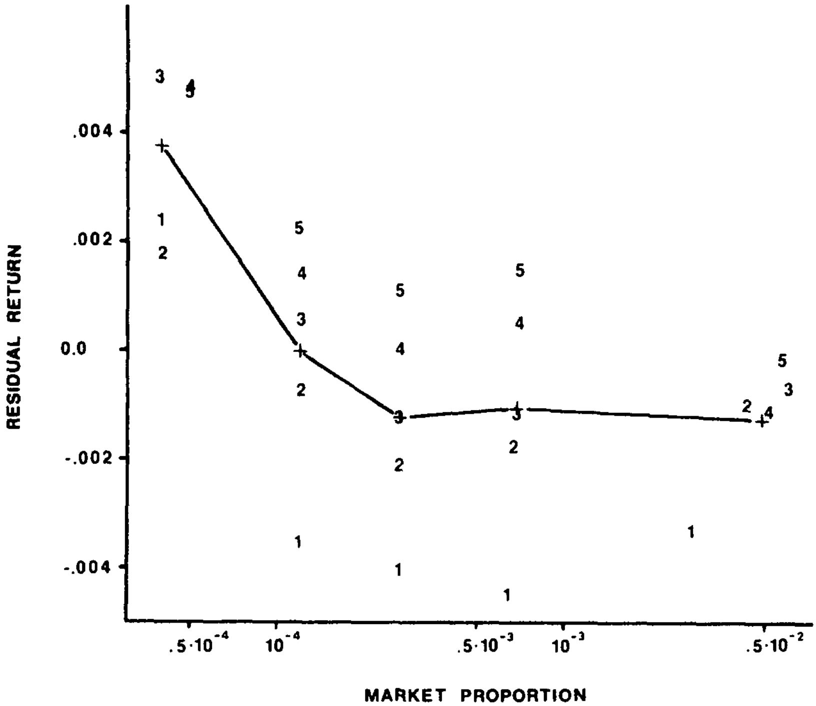

Fig. 2. Mean residual returns of portfolios (1936–1975) with equally-weighted CRSP index as market proxy. The residual is calculated with the two-factor model ($\hat{\varepsilon}_{it} = R_{it} - \hat{\gamma}_{0t} - \hat{\gamma}_{1t}\hat{\beta}_{it}$) The symbols are as defined for fig. 1.

4.3. *'Arbitrage' portfolio returns*

One important empirical question still remains: How important is the size effect from a practical point of view? Fig. 2 suggests that the difference in returns between the smallest firms and the remaining ones is, on average, about 0.4 percent per month. A more dramatic result can be obtained when the securities are chosen solely on the basis of their market value.

As an illustration, consider putting equal dollar amounts into portfolios containing the smallest, largest and median-sized firms at the beginning of a year. These portfolios are to be equally weighted and contain, say, ten, twenty or fifty securities. They are to be held for five years and are rebalanced every month. They are levered or unlevered to have the same beta. We are then interested in the differences in their returns,

$$R_{1t} = R_{st} - R_{lt}, \qquad R_{2t} = R_{st} - R_{at}, \qquad R_{3t} = R_{at} - R_{lt}, \tag{6}$$

Table 2
Portfolio estimators for γ_2 for all 25 portfolios and for 20 portfolios (portfolios containing smallest firms deleted) based on CRSP equally weighted index with generalized least-squares estimation.[a]

Period	Size premium $\hat{\gamma}_2$ with 25 portfolios	20 portfolios
1936–1975	−0.00044 (−2.42)	−0 00043 (−2.54)
1936–1955	−0.00037 (−1 72)	−0.00041 (−1.88)
1956–1975	−0.00056 (−1.91)	−0.00050 (−1.91)
1936–1945	−0.00085 (−2.81)	−0.00083 (−2.48)
1946–1955	0.00003 (0.12)	−0.00003 (−0 13)
1956–1965	−0.00023) (−0.81)	−0.00017 (−0.65)
1966–1975	−0.00091 (−1.78)	−0.00085 (−1.84)

[a] t-statistic in parentheses

where R_{st}, R_{at} and R_{lt} are the returns on the portfolios containing the smallest, median-sized and largest firms at portfolio formation time (and $R_{1t} = R_{2t} + R_{3t}$). The procedure involves (a) the calculation of the three differences in raw returns in each month and (b) running time series regressions of the differences on the excess returns of the market proxy. The intercept terms of these regressions are then interpreted as the $\bar{R}_i$, $i=1,\ldots,3$. Thus, the differences can be interpreted as 'arbitrage' returns, since, e.g., R_{1t} is the return obtained from holding the smallest firms long and the largest firms short, representing zero net investment in a zero-beta portfolio.[12] Simple equally weighted portfolios are used rather than more sophisticated minimum variance portfolios to demonstrate that the size effect is not due to some quirk in the covariance matrix.

Table 3 shows that the results of the earlier tests are fully confirmed. $\bar{R}_2$, the difference in returns between very small firms and median-size firms, is typically considerably larger than $\bar{R}_3$, the difference in returns between median-sized and very large firms. The average excess return from holding very small firms long and very large firms short is, on average, 1.52 percent

[12] No *ex post* sample bias is introduced, since monthly rebalancing includes stocks delisted during the five years. Thus, the portfolio size is generally accurate only for the first month of each period

Table 3
Mean monthly returns on 'arbitrage' portfolios.[a]
$R_j - R_k = \hat{\alpha}_i + \hat{\beta}_i (R_m - R_F)$

	$\bar{\alpha}_1$[b]			$\bar{\alpha}_2$[c]			$\bar{\alpha}_3$[d]		
	$n=10$	$n=20$	$n=50$	$n=10$	$n=20$	$n=50$	$n=10$	$n=20$	$n=50$
Overall period									
1931–1975	0.0152	0.0148	0.0101	0.0130	0.0124	0.0089	0.0021	0.0024	0.0012
	(2.99)	(3.53)	(3.07)	(2.90)	(3.56)	(3.64)	(1.06)	(1.41)	(0.85)
Five-year subperiods									
1931–1935	0.0589	0.0597	0.0427	0.0462	0.0462	0.0326	0.0127	0.0134	0.0101
	(2.25)	(2.81)	(2.35)	(1.92)	(2.55)	(2.46)	(1.09)	(1.49)	(1.42)
1936–1940	0.0201	0.0182	0.0089	0.0118	0.0145	0.0064	0.0084	0.0037	0.0025
	(0.82)	(0.97)	(0.67)	(0.55)	(0.90)	(0.65)	(1.20)	(0.62)	(0.49)
1941–1945	0.0430	0.0408	0.0269	0.0381	0.0367	0.0228	0.0049	0.0038	0.0041
	(2.29)	(2.46)	(2.17)	(2.29)	(2.54)	(2.02)	(1.25)	(1.09)	(1.68)
1946–1950	−0.0060	−0.0046	−0.0036	−0.0058	−0.0059	−0.0029	−0.0002	−0.0104	−0.0007
	(−1.17)	(−0.97)	(−0.97)	(−1.03)	(−1.29)	(−0.83)	(−0.07)	(−0.50)	(−0.38)
1951–1955	−0.0067	−0.0011	0.0013	−0.0004	0.0026	0.0010	−0.0062	−0.0037	0.0003
	(−0.89)	(−0.21)	(0.32)	(−0.07)	(0.72)	(0.39)	(−1.29)	(−0.99)	(0.11)
1956–1960	0.0039	0.0008	0.0037	0.0007	−0.0027	0.0011	0.0031	0.0035	0.0026
	(0.67)	(0.15)	(0.89)	(0.14)	(−0.64)	(0.45)	(0.88)	(1.16)	(0.97)
1961–1965	0.0131	0.0060	0.0024	0.0096	0.0046	0.0036	0.0035	0.0014	−0.0012
	(1.38)	(0.67)	(0.31)	(1.11)	(0.72)	(0.77)	(0.59)	(0.24)	(−0.24)
1966–1970	0.0121	0.0117	0.0077	0.0129	0.0110	0.0071	0.0008	0.0007	0.0006
	(1.64)	(2.26)	(1.91)	(1.93)	(2.71)	(2.43)	(0.23)	(0.22)	(0.27)
1971–1975	0.0063	0.0108	0.0098	0.0033	0.0077	0.0083	0.0030	0.0031	0.0015
	(0.60)	(1.23)	(1.45)	(0.39)	(1.18)	(1.79)	(0.64)	(0.72)	(0.43)

[a]Equally-weighted portfolios with n securities, adjusted for differences in market risk with respect to CRSP value-weighted index, t-statistics in parentheses.
[b]Small firms held long, large firms held short.
[c]Small firms held long, median-size firms held short.
[d]Median-size firms held long, large firms held short.

per month or 19.8 percent on an annualized basis. This strategy, which suggests very large 'profit opportunities', leaves the investor with a poorly diversified portfolio. A portfolio of small firms has typically much larger residual risk with respect to a value-weighted index than a portfolio of very large firms with the same number of securities [Banz (1978, ch. 3)]. Since the fifty largest firms make up more than 25 percent of the total market value of NYSE stocks, it is not surprising that a larger part of the variation of the return of a portfolio of those large firms can be explained by its relation with the value-weighted market index. Table 3 also shows that the strategy would not have been successful in every five year subperiod. Nevertheless, the magnitude of the size effect during the past forty-five years is such that it is of more than just academic interest.

5. Conclusions

The evidence presented in this study suggests that the CAPM is misspecified. On average, small NYSE firms have had significantly larger risk adjusted returns than large NYSE firms over a forty year period. This size effect is not linear in the market proportion (or the log of the market proportion) but is most pronounced for the smallest firms in the sample. The effect is also not very stable through time. An analysis of the ten year subperiods show substantial differences in the magnitude of the coefficient of the size factor (table 1).

There is no theoretical foundation for such an effect. We do not even know whether the factor is size itself or whether size is just a proxy for one or more true but unknown factors correlated with size. It is possible, however, to offer some conjectures and even discuss some factors for which size is suspected to proxy. Recent work by Reinganum (1980) has eliminated one obvious candidate: the price–earnings (P/E) ratio.[13] He finds that the P/E-effect, as reported by Basu (1977), disappears for both NYSE and AMEX stocks when he controls for size but that there is a significant size effect even when he controls for the P/E-ratio, i.e., the P/E-ratio effect is a proxy for the size effect and not vice versa. Stattman (1980), who found a significant negative relationship between the ratio of book value and market value of equity and its return, also reports that this relationship is just a proxy for the size effect. Naturally, a large number of possible factors remain to be tested.[14] But the Reinganum results point out a potential problem with some of the existing negative evidence of the efficient market hypothesis. Basu believed to have identified a market inefficiency but his P/E-effect is

[13]The average correlation coefficient between P/E-ratio and market value is only 0.16 for individual stocks for thirty-eight quarters ending in 1978. But for the portfolios formed on the basis of P/E-ratio, it rises to 0.82. Recall that Basu (1977) used ten portfolios in his study.

[14]E.g., debt–equity ratios, skewness of the return distribution [Kraus and Litzenberger (1976)].

just a proxy for the size effect. Given its longevity, it is not likely that it is due to a market inefficiency but it is rather evidence of a pricing model misspecification. To the extent that tests of market efficiency use data of firms of different sizes and are based on the CAPM, their results might be at least contaminated by the size effect.

One possible explanation involving the size of the firm directly is based on a model by Klein and Bawa (1977). They find that if insufficient information is available about a subset of securities, investors will not hold these securities because of estimation risk, i.e., uncertainty about the true parameters of the return distribution. If investors differ in the amount of information available, they will limit their diversification to different subsets of all securities in the market.[15] It is likely that the amount of information generated is related to the size of the firm. Therefore, many investors would not desire to hold the common stock of very small firms. I have shown elsewhere [Banz (1978, ch. 2)] that securities sought by only a subset of the investors have higher risk-adjusted returns than those considered by all investors. Thus, lack of information about small firms leads to limited diversification and therefore to higher returns for the 'undesirable' stocks of small firms.[16] While this informal model is consistent with the empirical results, it is, nevertheless, just conjecture.

To summarize, the size effect exists but it is not at all clear why it exists. Until we find an answer, it should be interpreted with caution. It might be tempting to use the size effect, e.g., as the basis for a theory of mergers — large firms are able to pay a premium for the stock of small firms since they will be able to discount the same cash flows at a smaller discount rate. Naturally, this might turn out to be complete nonsense if size were to be shown to be just a proxy.

The preceding discussion suggests that the results of this study leave many questions unanswered. Further research should consider the relationship between size and other factors such as the dividend yield effect, and the tests should be expanded to include OTC stocks as well.

[15] Klein and Bawa (1977, p 102)

[16] A similar result can be obtained with the introduction of fixed holding costs which lead to limited diversification as well. See Brennan (1975), Banz (1978, ch. 2) and Mayshar (1979)

References

Ball, Ray, 1978, Anomalies in relationships between securities' yields and yield surrogates, Journal of Financial Economics 6, 103–126

Banz, Rolf W., 1978, Limited diversification and market equilibrium: An empirical analysis, Ph D dissertation (University of Chicago, Chicago, IL)

Basu, S., 1977, Investment performance of common stocks in relation to their price-earnings ratios. A test of market efficiency, Journal of Finance 32, June, 663–682.

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, Journal of Business 45, July, 444–454.

Black, Fischer, and Myron Scholes, 1974, The effects of dividend yield and dividend policy on common stock prices and returns, Journal of Financial Economics 1, May, 1–22.

Black, Fischer, Michael C. Jensen and Myron Scholes, 1972, The capital asset pricing model Some empirical tests, in: M.C. Jensen, ed., Studies in the theory of capital markets (Praeger, New York) 79–121.

Brennan, Michael J., 1975, The optimal number of securities in a risky asset portfolio when there are fixed costs of transacting: Theory and some empirical evidence, Journal of Financial and Quantitative Analysis 10, Sept., 483–496.

Brenner, Menachem, 1976, A note on risk, return and equilibrium: Empirical tests, Journal of Political Economy 84, 407–409.

Fama, Eugene F., 1976, Foundations of finance (Basic Books, New York).

Fama, Eugene F. and James D. MacBeth, 1973, Risk return and equilibrium. Some empirical tests, Journal of Political Economy 71, May–June, 607–636.

Ibbotson, Roger G. and Carol L. Fall, 1979, The United States market wealth portfolio, Journal of Portfolio Management 6, 82–92.

Ibbotson, Roger G. and Rex A. Sinquefield, 1977, Stocks, bonds, bills and inflation: The past (1926–1976) and the future (1977–2000) (Financial Analysis Research Foundation)

Klein, Roger W. and Vijay S. Bawa, 1977, The effect of limited information and estimation risk on optimal portfolio diversification, Journal of Financial Economics 5, Aug., 89–111.

Kraus, Alan and Robert H Litzenberger, 1976, Skewness preference and the valuation of risk assets, Journal of Finance 31, 1085–1100

Levi, Maurice D., 1973, Errors in the variables bias in the presence of correctly measured variables, Econometrica 41, Sept., 985–986.

Litzenberger, Robert H and Krishna Ramaswamy, 1979, The effect of personal taxes and dividends on capital asset prices: Theory and empirical evidence, Journal of Financial Economics 7, June, 163–195.

Mayshar, Joram, 1979, Transaction costs in a model of capital market equilibrium, Journal of Political Economy 87, 673–700.

Reinganum, Marc R, 1980, Misspecification of capital asset pricing: Empirical anomalies based on earnings yields and market values, Journal of Financial Economics, this issue

Roll, Richard, 1977, A critique of the asset pricing theory's tests: Part I, Journal of Financial Economics 4, Jan., 120–176.

Stattman, Dennis, 1980, Book values and expected stock returns, Unpublished M.B.A. honors paper (University of Chicago, Chicago, IL).

Theil, Henri, 1971, Principles of econometrics (Wiley, New York).

U.S. Department of Commerce, Office of Business Economics, 1969, 1970, Survey of current business 49, May, 11–12, 50, May, 14.