ETLA
ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B 00120 Helsinki Finland Tel. 358-9-609 900
Telefax 358-9-601 753 World Wide Web: http://www.etla.fi/

Keskusteluaiheita - Discussion papers

No. 691

Hans Gerhard Heidle*

MARKET MICROSTRUCTURE AND ASSET PRICING: A SURVEY

* Ph.D. Candidate in Finance at the Owen Graduate School of Management, Vanderbilt University. I am indebted to Hans Stoll for stimulating my interest in this subject and for valuable discussions and suggestions. Email: hans.heidle@owen.vanderbilt.edu.

ISSN 0781-6847 25.10.1999

Heidle, Hans Gerhard – **Market Microstructure and Asset Pricing: A Survey**, Helsinki: ETLA, Elinkeinoelämän Tutkimuslaitos, The Research Institute of the Finnish Economy, 1999, 57 p. (Keskusteluaiheita – Discussion Papers, ISSN 0781-6847, No. 691).

Abstract: This paper investigates the relation between market microstructure and asset pricing. First, it gives an extensive review of previous work in this area, theoretical as well as empirical. The papers are contrasted according to their different approaches and results. Interestingly, there is no unanimous opinion in this area. While Constantinides (1986) states that transaction costs have only second-order effects on asset returns, others like Amihud and Mendelson (1986) claim that transaction costs are significant in determining asset returns. Then, I present a simple model which yields results similar to Amihud and Mendelson (1986) and I discuss possible extensions and modifications. Finally, some empirical issues are raised and discussed briefly.

Keywords: market microstructure, asset pricing, capital markets, transaction costs, market friction.

JEL codes: D40, G10, G12, G14.

Heidle, Hans Gerhard – **Markkinoiden mikrorakenne ja arvopapereiden hinnoittelu: kirjallisuuskatsaus**, Helsinki: ETLA, Elinkeinoelämän Tutkimuslaitos, The Research Institute of the Finnish Economy, 1999, 57 s. (Keskusteluaiheita – Discussion Papers, ISSN 0781-6847, No. 691).

Tiivistelmä: Tämä tutkimus käsittelee markkinoiden mikrorakenteen ja arvopapereiden hinnoittelun välistä suhdetta. Aluksi käsitellään aiempaa teoreettista ja empiiristä kirjallisuutta korostaen eri lähestymistapoja ja toisistaan poikkeavia tuloksia. Kirjallisuudessa esitetyt näkemykset ovat ristiriitaisia: Constantinideksen (1986) mukaan transaktiokustannuksilla on vain sekundäärinen merkitys arvopapereiden hinnoitteluun; toisaalta esim. Amihud ja Mendelson (1986) väittävät että transaktiokustannuksilla on merkittävä vaikutus arvopapereiden tuottoihin. Seuraavaksi määritellään Amihudin ja Mendelsonin (1986) työhön pohjautuva malli ja keskustellaan sen mahdollisista laajennuksista ja muunnelmista. Lopuksi nostetaan esiin muutamia empiirisiä kysymyksiä.

Avainsanat: markkinoiden mikrorakenne, arvopapereiden hinnoittelu, pääomamarkkinat, transaktiokustannukset, markkinoiden epätäydellisyys.

Preface

Mr. Hans Gerhard Heidle, a Ph.D. Candidate in Finance at the *Owen Graduate School of Management* (Vanderbilt University), kindly accepted our invitation to visit ETLA and share his thoughts on the latest developments in finance. His presentation *Issues in Finance – Market Microstructure and its Relevance* (25 August 1999) gave us an excellent introduction to the topic as well as information on the cutting edge research in the field by Mr. Heidle and others. From the outset it seemed that the topic would be of little interest to somebody not familiar with the field, but Mr. Heidle convincingly showed that market microstructure related issues may indeed be of interest to regulators, investors, companies, and even to the society as a whole.

There have been significant research in the 'Great Depression' of Finland in the early 1990s and many authors have concluded that the stock market played a considerable role in the boom of the 1980s and the bust that followed. Thus, one could take a historical perspective in studying market microstructure in Finland. On the other hand, great changes also lie ahead: Helsinki Stock Exchange is seeking alliances with its counterparts in other European countries. Even relatively small Finnish companies will soon have a real option to be listed in a number of foreign stock exchanges or abroad altogether.

Although this survey dates back a few years, it serves as a good introduction to the literature on market microstructure and its relation to asset pricing. We were delighted to have Mr. Heidle as our guest and are looking forward to hearing from him in the future.

Pentti Vartia, CEO
ETLA – The Research Institute of the Finnish Economy
October 25, 1999

Contents

1. Introduction

One of the central questions in finance is how prices and returns of assets are determined. Hence, it is natural to ask if and how the structure of financial markets may affect the pricing of assets. Intuitively, there are two basic views. First, one can think that a high bid-ask spread of an security is due to high risk of the underlying asset. That is asset pricing influences market microstructure via the bid-ask spread. A second view is that a higher return on an asset is not only due to the risk of the underlying asset rather it is also due to a higher bid-ask spread on this security. That is market microstructure affects asset pricing via the bid-ask spread. Which view is correct is an empirical issue.

Banz (1981) and Reinganum (1981) document a size anomaly in observed average returns of equity, empirical evidence not consistent with the Sharpe-Lintner-Mossin CAPM [SLM-CAPM].[1] Their work kicked off a discussion of the size effect and its possible explanations. The significance of the size effect in returns is also documented by Fama and French (1992, 1993). Fama and French try to explain asset returns with three factors, a broad index, the book-to-market ratio of the firm, and the size of the firm. Their model proves to be quite useful and the three factors introduced by Fama and French are used in subsequent work to adjust for risk, i.e. their factors replace the traditional beta-risk adjustment, e.g., James and Woodward (1995) and Brennan and Subrahmanyam (1994). However, Fama and French (1992, 1993) leave the question of the economical justification for their model open.

Stoll and Whaley (1983) and Day, Stoll, and Whaley (1985) suggest transaction costs as a possible explanation for the size effect. Stoll and Whaley (1983) find that the total market value of the firm is inversely related to risk-adjusted returns (Stoll and Whaley adjust for risk using the traditional beta from the SLM-CAPM). Besides this size effect they also document a price per share effect, that is the price per share is also inversely related to the risk-adjusted returns. With an arbitrage portfolio methodology, they demonstrate that transaction costs can at least partially explain the size effect. Thus, they conclude that transaction costs could be the (or one) missing factor in the single-period, two-parameter CAPM. Further, Stoll and Whaley show that the size effect reverses when the analysis uses one month returns

[1] To be exact, it should be noted that the empirical evidence either suggests that the market is not efficient or that the CAPM does not hold or both. It is always a joint hypothesis respectively a joint test of market efficiency and the CAPM. However, most authors tend to assume market efficiency and reject the SLM-CAPM.

net of transaction costs, that is small firms yield negative abnormal returns after transaction costs for a holding period of one month. However, with longer holding periods the size effect diminishes and with four months returns there is virtually no size effect detectable. They conclude that the data is consistent with the CAPM applied to after transaction costs returns over longer investment horizons. Day et al. (1985) also show that the price per share seems to work better than size of the firms in explaining the anomaly found in average returns.

This brief introduction indicates the considerable interest the finance community has in the question if and how asset pricing is affected by market microstructure. If this link can explain empirical anomalies found in the stock markets has also found recognition in the finance literature. In the next section, I present and review important papers in this area[2] and I try to group them according to the different approaches they take as well as according to the conclusions they draw.[3] Section 3 provides a discussion of possible extensions of the significant Amihud and Mendelson (1986) model, presents a simple model, and discusses some empirical issues. Finally, Section 4 contains concluding remarks.

2. Previous Work

In this section, I review previous work dealing with the effect of transaction costs on asset pricing and portfolio selection. First, I present purely theoretical papers by Chen, Kim, and Kon (1975), Magill and Constantinides (1976), Constantinides (1986), and Amihud, Mendelson, and Yu (1992). The authors focus on the intertemporal portfolio selection of investors facing (proportional) transaction costs. Interestingly, they do not necessarily agree in their conclusions. Chen et al. (1975) as well as Amihud et al. (1992) show that transaction costs have significant effects on asset pricing (expected returns), whereas Constantinides (1986) concludes that transaction costs only have second-order effects on the liquidity premia implied by equilibrium asset returns. That means the small firm effect cannot be explained by transaction costs, which is contradictive to empirical evidence found by Stoll and Whaley

[2]I do not claim that this is a complete discussion nor that the list of the papers is a complete one.

[3]Some authors conclude that transaction costs are an important component in returns, e.g. Stoll and Whaley (1983) or Amihud and Mendelson (1986), others think that transaction costs have only a second-order effect on asset pricing, e.g., Constantinides (1986). However, the empirical evidence is mainly in favor of the first group, see Stoll and Whaley (1983) and Amihud and Mendelson (1986), whereas Constantinides (1986) does not provide empirical evidence.

(1983) and Amihud and Mendelson (1986).

The second group of papers provide theoretical foundation as well as empirical analysis. I will discuss Amihud and Mendelson (1986), Brennan and Subrahmanyam (1994), and James and Woodward (1995). The last group is formed by purely empirical papers such as Amihud and Mendelson (1989), Stoll and Whaley (1983), Day et al. (1985), Reinganum (1990), and Eleswarapu and Reinganum (1993). With the exception of Eleswarapu and Reinganum (1993), all empirical investigations seem to support the view that transaction costs are a determinant of asset returns and that the small firm effect may be explained by transaction costs. Eleswarapu and Reinganum (1993) shed some doubt on this view by analyzing seasonalities and get different results compared to Amihud and Mendelson (1986), although they are using the same time period.

2.1. Theoretical Work

2.1.1. Chen, Kim, and Kon (1975)

Chen et al. (1975)–CKK–try to unify the portfolio demand and the liquidity demand for money of individual investors. The motivation is clear since money reduces the risk of the portfolio as well as the liquidation costs in meeting cash demands. Their model extends the SLM-CAPM by introducing stochastic cash demand and liquidation costs, i.e., it is a single period model and dynamic portfolio adjustments over time are not modelled. Their model shows that not all investors will hold the same risky portfolio in equilibrium as the SLM-CAPM predicts, i.e., the separation property no longer holds. Thus, transaction costs are a possible explanation for discrepancies between empirical evidence and the traditional CAPM. CKK rely on the standard assumptions of the SLM-CAPM. They measure the degree of liquidity of assets to an investor by (a)the liquidation costs the investor incurs when selling the asset to meet his (stochastic) cash demands, and (b)the covariance between the return of the asset and the investor's stochastic cash demands. They modify and add to the standard assumptions of the SLM-CAPM in the following way:

- Investors are risk-averse and single-period expected utility maximizers, where their preference functions are defined in terms of mean and variance of the ending portfolio value net of the liquidation costs incurred in meeting cash demands.

- Stochastic cash demands among investors can differ.

- There exists a riskfree liquid asset with a certain rate of return and without any liquidation costs. Risky assets carry a proportional penalty cost when liquidated. This proportional penalty cost is equal among risky assets.

CKK distinguish between internal and external liquidity. The internal liquidity of an asset is the sum of the liquidity services yielded by an asset to its holder. The aggregated liquidity services yielded by an asset to all market participants are called its external liquidity. Liquidity services are defined as the asset's marginal contribution to the investor's expected penalty cost, variance of penalty cost of liquid asset shortage, and liquidity risk (covariance between asset return and cash demands).

CKK formulate and solve the maximization problem for the individual investor. The result is a capital asset pricing model with cash demands (CAPMCD):

$$\begin{aligned} E(\widetilde{R}_k) &= \alpha + \lambda[Scov(\widetilde{R}_k, \widetilde{R}_m) - cov(\widetilde{R}_k, \widetilde{\Phi})], \\ \alpha &= R_f + \left\{ \frac{\sum_i \left[-\frac{1}{2}\left(\frac{\partial V_i}{\partial E_i}\right)_{B_i} E(\widetilde{\Phi}_i) - \frac{1}{2} {}_{B_i} V(\widetilde{\Phi}_i) + \sum_{h=1}^{n} S_{ih} \frac{\partial cov(\widetilde{R}_h, \widetilde{\Phi}_i)}{\partial B_i} \right]}{\frac{1}{2}\left(\sum_i \frac{\partial V_i}{\partial E_i}\right)} \right\}, \\ \lambda &= \frac{E(\widetilde{R}_m) - \alpha}{Svar(\widetilde{R}_m) - cov(\widetilde{R}_m, \widetilde{\Phi})}. \end{aligned} \tag{2.1}$$

where:

R_f	1+ riskfree rate of interest
$\widetilde{R}_k,\ \widetilde{R}_m$	1+ expected rate of return on the kth asset/market pf
$E(\cdot), cov(\cdot)$	expectation and covariance operator, respectively
λ	market price for risk
$\widetilde{\Phi}_i$	stochastic penalty function of individual investor
$\widetilde{\Phi} = \sum_i \widetilde{\Phi}_i$	aggregate stochastic penalty function
$E_i = E(\widetilde{W}_i)$	expected net ending portfolio value
$V_i = var(\widetilde{W}_i)$	variance of net ending portfolio value

S_{ih}	market value of the ith investor's holding of the hth risky asset
S	aggregate market value of all risky assets
B_i	market value of the ith investor's holding of the riskfree liquid asset
$E(\widetilde{\Phi}_i)$	expected penalty cost
${}_{B_i}E(\widetilde{\Phi}_i) = \partial E(\widetilde{\Phi}_i)/\partial B_i$	marginal contribution to the investor's expected penalty cost by transferring an investment dollar from risky assets to the riskfree liquid asset
$V(\widetilde{\Phi}_i)$	variance of penalty cost function
${}_{B_i}V(\widetilde{\Phi}_i) = \partial V(\widetilde{\Phi}_i)/\partial B_i$	marginal contribution to the variance of the investor's penalty cost by transferring an investment dollar from risky assets to the riskfree liquid asset

Equation 2.1 shows that the standard CAPM is altered in several ways. First, the intercept, α, is now the sum of one plus the riskfree rate of interest, R_f, and the risk-adjusted value of the liquidity services of the riskfree liquid asset, $\sum_i \left[-\frac{1}{2}\left(\frac{\partial V_i}{\partial E_i}\right)_{B_i} E(\widetilde{\Phi}_i) - \frac{1}{2}{}_{B_i}V(\widetilde{\Phi}_i) + \sum_{h=1}^{n} S_{ih}\frac{\partial cov(\widetilde{R}_h,\widetilde{\Phi}_i)}{\partial B_i}\right] \Big/ \frac{1}{2}\left(\sum_i \frac{\partial V_i}{\partial E_i}\right)$. The riskfree asset reduces the expected penalty cost, ${}_{B_i}E(\widetilde{\Phi}_i) < 0$, as well as the variance of the penalty cost, ${}_{B_i}V(\widetilde{\Phi}_i) < 0$. The last term in the numerator of the right-hand side, $\sum_{h=1}^{n} S_{ih}\frac{\partial cov(\widetilde{R}_h,\widetilde{\Phi}_i)}{\partial B_i}$, is the aggregate adjustment for liquidity risk. Second, the market price for risk, λ, has the excess return over the riskfree return plus the value of the liquidity services of the riskfree asset as numerator, $E(\widetilde{R}_m) - \alpha$. The denominator consists of the variability risk of the market portfolio, $Svar(\widetilde{R}_m)$, and the aggregate external liquidity risk premium, $-cov(\widetilde{R}_m, \widetilde{\Phi})$.

Compared to the SLM-CAPM, there is an additional risk premium included which is priced by the market, $-cov(\widetilde{R}_k, \widetilde{\Phi})$. The additional risk premium is due to the penalty costs, which in turn depend on the stochastic aggregate cash demand. CKK show that this additional component is positive for liquidity preferred assets [$cov(\widetilde{R}_k, \widetilde{J}) > 0$, where $\widetilde{J}$is the aggregate stochastic cash demand], zero for liquidity neutral assets [$cov(\widetilde{R}_k, \widetilde{J}) = 0$], and negative for liquidity averse assets [$cov(\widetilde{R}_k, \widetilde{J}) < 0$]. Depending on the characteristic, the asset's total systematic risk increases or decreases and the market demands a higher or lower

expected return. Hence, if $cov(\widetilde{R}_m, \widetilde{J}) < 0$the traditional CAPM is very likely to overstate the market price for risk.

CKK also analyze the individual investor's portfolio choice. They find that the demand for a specific asset is dependent on the characteristic of this asset (if it is liquidity preferred or not), and on the standard deviation of the asset. More important they find that the risky portfolio an investor chooses is now dependent on his individual characteristics, i.e., his specific stochastic cash demands, his individual taste, and his initial wealth. Thus, the separation theorem no longer holds, and not all investors hold identical risky portfolios.

Summarized, CKK find that there is still a linear risk-return relation, but liquidation costs and stochastic cash demands do matter. The portfolio choice of individual investors as well as the pricing formula are altered. The SLM-CAPM tends to overstate the market price for risk. Individual investors do not hold the identical risky portfolio, i.e., the separation theorem no longer holds. CKK conclude that their model rationalizes financial intermediation, since the pooling of different individual stochastic cash demands can lead to external economies.

However, their model includes some features which are questionable. All risky assets carry the same proportional penalty costs. To get a deeper understanding of the impact of liquidation costs on individual asset returns, one has to assign different penalty costs to the different assets. A more severe critique can be found in the comment by Constantinides (1976).

Constantinides (1976). Constantinides claims that CKK are inconsistent in their approach, since CKK's model is neither a single-period nor a multi-period setting. He points out further that CKK apply their liquidation costs inconsistently. First, investors do not incur any transaction costs when investing their initial wealth at time zero. Second, investors maximize a risk-averse preference function which depends on the mean and variance of $\widetilde{W}_i$, the terminal wealth, which consists of the initial holdings of assets, plus returns, less the cash demand and less the liquidation costs incurred when meeting the cash demand in excess of the holding in the riskfree liquid asset.

To illustrate the second point of criticism, Constantinides first considers a single-period setting. In a single-period framework, after the investor has satisfied the cash demand, he

will consume his entire remaining wealth, i.e., he has to liquidate $\widetilde{W}_i$and thus additional liquidation costs will be incurred. CKK ignore those additional liquidation costs. In a multi-period framework, a similar problem arises, when we assume that $\widetilde{W}_i$at the end of the period is only partly consumed and partly invested for future consumption. First, CKK again ignore liquidation costs on the consumed part. Second, the value which matters at the end of the period is the utility of future consumption induced by $\widetilde{W}_i$, but this implies that not only the value, but also the composition of the portfolio at the end of the period matters. The induced utility is a function of total wealth $\widetilde{W}_i$as well as of the portfolio composition at the end of the period.

Due to these inconsistencies, Constantinides questions the CKK model and the implied results. He points out that the use of risk-averse expected utility of wealth maximization is not justified when introducing liquidation costs.

2.1.2. Magill and Constantinides (1976)

In a fairly technical paper, Magill and Constantinides (1976)–MC–introduce transaction costs into continuous time portfolio selection. Merton (1969, 1971) has shown that using a continuous time framework simplifies the extension of the one-period mean-variance analysis to the dynamic case. Thus, in contrast to Chen et al. (1975), MC model the rebalancing of the portfolio over time explicitly. The use of a continuous time framework also seems justified since most capital markets allow for trading at any moment in time. However, a potential weakness is, that as long as trading is costless, the amount of trading induced by constantly changing prices is unrealistic high compared with observed investor behavior. MC show that this weakness can be resolved by introducing transaction costs. They point out that a continuous time model with transaction costs seems to be the most realistic image of investor behavior on theoretical as well as empirical grounds.

Their model implies that each investor does not have *the* optimal portfolio, rather she has a *region* of optimal portfolios. Like Chen et al. (1975), this implies that not all investors hold the same risky portfolio. A second result is that investors trade at discrete and random points in time, which is consistent with observed trading behavior.

The solution of the portfolio selection problem for the individual investor is fairly com-

plicated and referring to the original paper, I just emphasize some assumptions and implications. Interestingly, the problem and the solution is related to inventory control problems. Some of the main assumptions are:

- The market is continuous and competitive. There are m securities and the riskfree rate r allows investors to deposit or borrow unlimited amounts.
- Information about probability distributions of security prices and current price quotations are perfect and costless.
- Transaction costs are proportional to the value of the transaction and are incurred on sales and purchases, and the proportional transaction costs vary across securities.
- Investors know their respective expected life-span $[0, T]$ and expects to earn income $y(t)$ as a continuous function over $[0, T]$. Investors act as if they knew T and $y(t)$ with certainty at $t = 0$.
- Securities prices are modeled as geometric Brownian motions, thus prices are lognormally distributed.
- Investors are assumed to have decreasing absolute risk aversion and they choose their consumption and transaction policy such that they maximize their utility over time.

As mentioned, the derivation is complicated and lengthy, and for this survey, the implications are of more interest. The intuition behind the results is that there is a trade-off between transaction costs and the benefits obtained from rebalancing the portfolio (which leads to improved diversification). MC show that there is a region of portfolio weights around the optimum in absence of transaction costs where the portfolio remains unchanged, and in this sense this region contains optimal portfolios. In this region, the benefits from rebalancing are not big enough to outweigh the associated transaction costs. Interestingly, this region of optimal portfolio weights is independent of the life-span T and the wealth of the investor. Furthermore, the proportions of the assets held in the portfolio are independent of the holdings in all other securities and transaction costs. The investor only rebalances her portfolio when the proportions are outside the boundaries of the optimal region, and in the course of rebalancing, she will bring them back to the closest boundary.

This implies that trades occur at randomly spaced instants in time. The frequency of trading in any security per unit of time is a declining function of the transaction costs of this security. The optimal consumption policy depends on the current portfolio policy, investor wealth and remaining life-span.

MC conclude that transaction costs have an impact on the optimal portfolio policy of an investor, in that the investor only rebalances the portfolio proportions when the benefits outweigh the transaction costs. They feel that the applied methods could be helpful in determining the impact of transaction costs on capital market equilibrium. However, their analysis is restricted to a special class of utility functions and the transaction costs are restricted to be proportional, more complex transaction cost functions or other utility functions may alter the results. They do derive testable hypotheses, but do not provide empirical evidence, although the result that investors trade at randomly spaced instants in time can be observed in the capital markets.

The papers discussed so far focused on the impact of transaction costs on the intertemporal portfolio choice of individual investors. The next paper by Constantinides (1986) relates capital market equilibrium and transaction costs.

2.1.3. Constantinides (1986)

Constantinides (1986) considers again an intertemporal portfolio selection model with proportional transaction costs and he focuses on the decision problem of a representative investor. Trades are generated by the adjustment of the investor's portfolio to maximize expected utility of the infinite lifetime consumption. He restricts the analysis to two assets. The main result is that the demand for assets is sensitive to transaction costs, but transaction costs have only second-order effects on the liquidity premia implied by equilibrium asset returns. Furthermore, he shows that the liquidity premia implied by a single-period model with appropriately chosen period length will deviate from the liquidity premia implied by an intertemporal model since the appropriate time period is specific to each asset.[4]

The model is an exchange economy with a single consumption good as numeraire. There

[4]Stoll and Whaley (1983) also point out the importance of the length of the time period or investment horizon in their study of the small-firm effect and the possible explanation by transaction costs, see discussion below.

is one risky and one riskless asset. The investor is a pricetaker and continuous trading is possible. The returns are in form of capital gains only and there are no taxes. Short sales are permitted and assets are infinitely divisible. The riskless asset follows the following process:

$$dP_0(t) = P_0(t)rdt, \tag{2.2}$$

with solution:

$$P_0(t) = P_0(0)e^{rt}. \tag{2.3}$$

The risky asset follows a geometric Brownian motion:

$$dP_1(t) = P_1(t)[\mu dt + \sigma dw(t)], \tag{2.4}$$

where $dw(t)$ is the increment of a Wiener process.

The investor has wealth $W(t)$ at time t, he consumes $c(t)dt$ over $[t, t+dt]$ and he has an exponential instantaneous utility function. Constantinides denotes the investor's holding in the riskless asset prior to a transaction at time t with $x(t)$ and the holding in the risky asset with $y(t)$. The proportional transaction cost rate is a constant k. Thus if the investor buys $v(t)$ of the risky asset, the holding in the riskless asset will be $x(t) - v(t) - |v(t)|k$. Constantinides defines an investment policy as simple if there are reflecting barriers $\underline{\lambda}$ and $\overline{\lambda}$ with $\underline{\lambda} < \overline{\lambda}$ such that the investor does not transact as long as $\lambda = y(t)/x(t)$, the ratio of risky asset holdings to riskfree asset holdings lies in the interval $[\underline{\lambda}, \overline{\lambda}]$. The investor transacts to the closest boundary when the ratio $\lambda = y(t)/x(t)$ lies outside this interval. Similarly, he defines a simple consumption policy: The consumption rate is a constant fraction of the holding in the riskless asset, that is $\beta \equiv c(t)/x(t)$. Constantinides restricts his analysis to the set of simple policies. This limitation leads to underestimation of the maximized expected utility of consumption and thus overestimation of the liquidity premium. Therefore, the restriction to simple policies does not affect the conclusion that transaction costs have only second-order effects on the liquidity premium implied by asset returns, since the liquidity premium is even overestimated.

The expected derived utility function of $x(t)$ and $y(t)$ is

$$J[x(t), y(t); \beta, \underline{\lambda}, \overline{\lambda}] \equiv E_t \int_t^\infty e^{-\rho(\tau - t)} \gamma^{-1} c^\gamma(\tau) d\tau, \tag{2.5}$$

where E_t denotes the expectation at time t over the Wiener process $w(\tau)$ and $\beta, \underline{\lambda}, \overline{\lambda}$ are parameters characterizing a given simple investment policy and a given simple consumption policy. Using the Bellman equation, Constantinides is able to solve (2.5) and gets

$$J[x(t), y(t); \beta, \underline{\lambda}, \overline{\lambda}] = \frac{\beta^\gamma}{\rho - \gamma(r - \beta)} \left(\frac{x^\gamma}{\gamma} + A_1 x^{\gamma - s_1} \gamma^{s_1} + A_2 x^{\gamma - s2} \gamma^{s_2} \right), \tag{2.6}$$

where A_1 and A_2 are free parameters and s_1 and s_2 are the roots of the quadratic equation

$$\frac{\sigma^2}{2} s^2 + \left(\mu - \frac{\sigma^2}{2} - r + \beta \right) s - [\rho - \gamma(r - \beta)] = 0. \tag{2.7}$$

Constantinides further shows that the parameters A_1 and A_2 are uniquely determined by the controls $\beta, \underline{\lambda}$, and $\overline{\lambda}$. He also is able to prove that the optimal controls $\underline{\lambda}$ and $\overline{\lambda}$, which maximize (2.6) are independent of $x(t)$ and $y(t)$. However, he cannot prove that also the optimal control β is independent of $x(t)$ and $y(t)$. Constantinides concludes anyway that the optimal control triplet $(\beta, \underline{\lambda}, \overline{\lambda})$ is independent of $x(t)$ and $y(t)$ and determined by

$$J^*[x(t), y(t)] \equiv \max_{\beta, \underline{\lambda}, \overline{\lambda}} J[x(t), y(t); \beta, \underline{\lambda}, \overline{\lambda}], \qquad \underline{\lambda} \leq \frac{y(t)}{x(t)} \leq \overline{\lambda} \tag{2.8}$$

He suggests a numerical procedure to calculate the optimal controls for various model parameter values. This procedure yields some interesting results. I just highlight the most important implications for the topic of this review paper. The introduction of transaction costs broadens the region of no transactions, i.e., the difference between $\underline{\lambda}$ and $\overline{\lambda}$ widens. The region of no transactions is also shifted toward the riskless asset. This implies that the average demand for the risky asset is decreasing with the transaction costs. Hence, transaction costs have a first-order effect on the demand for assets.

Consider two assets with perfectly correlated rates of return and equal variances in their rates of return. Now, if trading in one of these assets is subject to transaction costs and trading in the other asset not, then the first asset has to offer a higher expected rate of return than the second, it has to offer a liquidity premium. Constantinides defines the liquidity premium, $\delta(k)$, on the risky asset in the presence of proportional transaction costs k as the increase in expected return on the risky asset which is necessary to give the investor the same expected utility in the economy with transaction costs and the economy without transaction costs at the ratio $y/x = \lambda^*$ (the optimal ratio without transaction costs). With this definition

he ignores the transaction costs incurred to initially adjust the portfolio weights to λ^*. Constantinides calculates the liquidity premium for several parameter values. He shows that $\delta(k)/k \cong .15$, that is the liquidity premium is one order of magnitude smaller than the proportional transaction costs. This allows Constantinides to conclude that transaction costs have only a second-order effect on the expected return of risky assets. He also states that transaction costs cannot explain the small firm effect documented by Banz (1981).[5] A different definition of the liquidity premium which accounts for the initial transaction costs yields similar results.

Furthermore, Constantinides points out the difference between one-period models and intertemporal models. In intertemporal models, the points in time when trading occurs is determined endogenously. For example in his model, trading only occurs if the ratio of the risky and the riskless asset is outside the region of no transacions. In a one-period model, the investment horizon is fixed and thus the assumed length of time to amortize the transaction costs is arbitrary. The liquidity premium implied by a one-period model is a function of this arbitrary investment horizon. To choose the "correct" period length cannot solve this problem, since the correct length is specific to each asset. He shows that for a simple one-period model the liquidity premium is

$$\delta(k) = -\frac{1}{T}\ln(1-k). \tag{2.9}$$

The variance of the asset's return does not show up in this equation, in contrast he showed that in the case of the intertemporal model, there is a strong positive relation between the liquidity premium and the variance of the asset's rate of return. Thus, the liquidity premium in the one-period model is too high for low-variance stocks and too low for high-variance stocks.

The major results are that there is a region of optimal ratios between the risky and the riskless asset where no trading occurs. Transaction costs have first-order effects on the demand for assets: First, there is no trading as long as the ratio is in the optimal region and second, the demand is shifted toward the asset without transaction costs. The intertemporal model shows that the frequency and the volume of trade decreases with transaction costs. On the other hand, transaction costs have only second-order effects on equilibrium asset

[5] In contrast to for example Stoll and Whaley (1983).

returns, the expected utility is insensitive to deviations of the asset proportions from the optimal proportions in the absence of transaction costs. Transaction costs may have a significant role in dissipating arbitrage profits, but Constantinides rules out that transaction costs can explain deviations of risk-adjusted rates of return (e.g., the small firm effect).

He offers possible extensions of his model. There could be more than one risky asset, he states that this would lower the liquidity premia even further. The model could incorporate fixed transaction costs, unfortunately, this would make it intractable. The process how firms supply their shares as well as the market making process could be included, i.e., the security prices would be endogenized.

Constantinides uses several simplifications and approximations in his model, which may effect the results in an unfavorable way. The numerical solution is only an approximation and he cannot prove the independence of β and $x(t), y(t)$. A major problem is that he does not provide any empirical evidence and his results are in contradiction with empirical studies by Stoll and Whaley (1983), Day et al. (1985) and Amihud and Mendelson (1986).

2.1.4. Amihud, Mendelson, and Yu (1992)

This paper investigates an intertemporal portfolio selection model. Amihud, Mendelson, and Yu (1992)–AMY–introduce proportional transaction costs and labor income risk. This setup is related to Chen et al. (1975)–CKK–and Constantinides (1986). The economy has two assets, a liquid asset and an illiquid asset, and they consider the decision problem of the representative consumer. The model differs from CKK in that it is a continuous time rather than a single-period model. In AMY's model, liquidation is not forced, it is endogenously determined by (uncertain) labor income. Hence, trading is not only induced by the need to rebalance the portfolio like in Constantinides (1976), it is also induced by the uncertainty in the labor income. The representative consumer has to trade-off the higher return on the illiquid asset and the expected transaction costs related to his investment policy. This investment policy is determined by two barriers, a lower and an upper bound, for the holding in the liquid asset. If the holding falls below the lower bound, the consumer sells the illiquid asset to increase the holding in the liquid asset to the lower bound. If the holding is above the upper bound, he invests the portion above the upper bound in the illiquid asset.

AMY first present a simple model without income uncertainty. They consider an infinitely-lived investor with repetitive two-period income pattern (certain). With an income of zero in the first period and an income of I in the second period. Assuming a CARA utility function

$$U = -\sum_{t=0}^{\infty} \frac{1}{\beta^t} e^{-C_t} \tag{2.10}$$

and a liquid asset with a constant return of 10% and illiquid assets with different proportional transaction costs,[6] AMY can show that the investor invests in exactly one of the assets when he has income I, and he consumes everything when his income is 0. The evolving investment and consumption plan for the two-period subperiods is the same over his lifetime. This yields the following maximization problem:

$$\begin{aligned} \max \quad & U = -\tfrac{\beta}{\beta-1}\left(e^{-C_1} + \tfrac{1}{\beta} e^{-C_2}\right) \\ s.t. \quad & R_k X(1-k) = C_1, \\ & I = C_2 + (1+k)X \end{aligned} \tag{2.11}$$

where R_k is the gross return on asset k, and X is the investment in asset k, and the asset k has proportional transaction costs of k. This problem can be solved easily. When setting k equal to zero in 2.11, AMY obtain a benchmark utility level when investing in the liquid asset: U_0. They define the liquidity premium as the additional return, m, the investor requires to be willing to invest in the illiquid asset rather than the liquid asset. That means the additional return on the illiquid asset with proportional transaction costs which leaves him with the utility U_0, he would reach by investing in the liquid asset. AMY compute liquidity premiums for different transaction costs rates. The liquidity premium increases with the transaction costs rate (as expected). Furthermore, the effect is not a second-order effect in contrast to Constantinides (1986). The relation between k and m is almost linear, due to the absence of income uncertainty. When income uncertainty is introduced, they still expect an increasing relation, but the relationship will be concave. The investor will reduce the trading in the illiquid asset when transaction costs rise thus the effect of transaction costs on his utility diminishes and he will require a smaller liquidity premium compared to the case without income uncertainty.[7]

[6] In their model, there is no uncertainty in the returns on assets.

[7] The increasing and concave relationship between transaction costs and liquidity premium is also documented in Amihud and Mendelson (1986), see discussion below.

In the complete continuous time model, consumers also have the CARA utility function 2.10 with risk aversion $\beta > 0$. Consumers have infinite horizon and they experience independent income shocks. The investment opportunities for the consumers consist of a liquid traded asset and an illiquid traded asset with proportional transaction costs on buying $k > 0$, and on selling $\pi > 0$. The income distribution of a price-taking, "representative" consumer follows a Gaussian process,

$$dI_t = \mu dt + \sigma dz_t, \tag{2.12}$$

where z_t is a standard Brownian motion.[8] With S_t as the consumer's total holding of the liquid asset after the realizations of labor income, dividend payouts (both assets have equal instantaneous dividend rates), the consumption decision, but before the investment decision at time t, and X_t as the time t holding of the illiquid asset before the investment decision, the consumer's problem at time 0 and with endowment (S_0, X_0) is:

$$\begin{array}{lrcl} \max & E_{(S_0,X_0)} \int_0^\infty -\frac{1}{\beta} e^{-\rho t - \beta C_t} dt, & & \\ s.t. & dS_t & = & (r_1 S_t + r_2 X_t + \mu - C_t)dt - (1+k)dL_t + (1-\pi)dU_t + \sigma dz_t, \\ & dX_t & = & dL_t - dU_t, \end{array} \tag{2.13}$$

with r_1, r_2 denoting the (certain) returns to the liquid respectively illiquid asset.[9] L_t, U_t are the cumulative purchases and sales of the illiquid asset up to time t. The consumer maximizes in his consumption and his investment. AMY restrict the consumption-investment policies (C_t, L_t, U_t) in the following way: $L_0 = U_0 = 0$, and

$$L_t = \int_0^t A_s ds, \qquad U_t = \int_0^t B_s ds, \qquad 0 \leq A_s, B_s \leq \kappa, \tag{2.14}$$

where κ is the upper bound of trading in the illiquid asset. Under the restriction (2.14), (2.13) can be solved using the Bellman equation. AMY show that the optimal trading policies are barrier policies, i.e., the consumer either buys or sells at the maximum level or she does not trade at all. They prove the necessity and the sufficiency of such policies and show (similar to other papers like Constantinides (1986) or Magill and Constantinides (1976)) that there

[8] Note, the formulation as arithmetic Brownian motion allows for negative income realizations to occur with positive probability.

[9] Note again, AMY abstract from return uncertainty, the assets offer certain returns.

are two constants, $Q_2 > Q_1$, such that as long as the holdings in the liquid asset, S_t, are in the region $[Q_1, Q_2]$, the consumer will not transact. If $S_t \geq Q_2$,she buys $\$A_t$ of the illiquid asset and if $S_t \geq Q_2$, she sells $\$B_t$ of the illiquid asset, where the purchase or the sale of the illiquid asset brings the holdings of the liquid asset exactly back to the violated boundary.

Hence, the consumer tries to smooth his consumption with her investment policy by using the illiquid asset to hedge the long-term income uncertainty, whereas the holding in the liquid asset are used for temporary imbalances between the consumer's consumption needs and her current income. AMY solve the complete consumption-investment problem and offer some numerical results to show the following: (1)Higher income uncertainty leads to a wider no-transaction region, i.e., the frequency and the volume of trading in the illiquid asset decreases. (2)The higher the return of the illiquid asset, the narrower the no-transaction region. (3)The higher the transaction costs, the wider the no-trading region.[10]

The most interesting question for the purpose of this survey is, if and how transaction costs affect the liquidity premia implied in the returns of assets. AMY define the liquidity premium as the minimum return premium which makes the investor indifferent between trading and holding only the liquid asset and trading both assets (the liquid and the illiquid). Their numerical computation suggest the following relations between the illiquidity premium, m, the transaction cost rate, k, and the income volatility, σ:

1. Keeping k constant, the higher σ, the higher the illiquidity premium, m. Furthermore, for high σ, an increase in k leads to an increase in m, and this effect is of the same order, in contrast to Constantinides (1986). For very high σ, the illiquidity premium becomes very large and for low σ, the effect of an increase in k on m is small. Seemingly, there is an increasing and convex relation between the required illiquidity premium and σ.

2. AMY find that m is an increasing and concave function of k. As the transaction cost rate increases the investor demands a higher return. On the other hand, as k increases, the no-trading region widens and thus the frequency if trading in the illiquid asset decreases and in turn the effect of the higher transaction costs is mitigated, which yields the concave relationship.

AMY's model offers an interesting implication on the behavior of investors. As stated, the

[10]These results seem to be intuitively clear, and similar results are found by Constantinides (1986).

illiquidity premium is the minimum excess return on the illiquid asset which compensates for the higher transaction costs. In equilibrium, the market gives the illiquidity premia on illiquid assets and the investors choose the assets which maximize their utility. Thus, investors with high income uncertainty may choose assets with low transaction costs to get the appropriate compensation. Investors with lower income uncertainty may prefer to invest in assets with high transaction costs, since they do not have to trade that often and thus the higher transaction costs may not affect them as much. A clientele effect may evolve. Actually, what should matter is the relative magnitude of an investor's income uncertainty compared to the income uncertainty of the marginal investor in a specific asset, who sets the illiquidity premium embedded in the return on the asset. However, they leave an exact investigation for further work.

The main results in this paper are the increasing and convex relationship between m and σ, and the increasing and concave relationship between m and k. Moreover, the effects of the transaction costs can be first-order. This is contradictive to Constantinides (1986), but it is consistent with the empirical findings in Amihud and Mendelson (1986). In addition, AMY's intertemporal model suggests a clientele effect which deserves further investigation.

AMY restrict their analysis to a two-asset setting with one liquid and one illiquid asset. They also consider only one representative investor, thus they do not derive equilibrium asset prices. However, one can argue that at least a "relative" pricing is possible, i.e., one can always focus on two assets and consider the one with the lower transaction costs as an asset without transaction costs. More severe, they also abstract from return risk of the assets. The assets offer a certain return, the only uncertainty in their model comes from the income uncertainty of the investor. The introduction of multiple risky assets with different transaction cost rates may produce a richer model and different conclusions. In the setting of AMY, a traditional portfolio selection problem is not existent, since there is no return uncertainty.

The papers discussed above were purely theoretical in their nature, they focused on intertemporal portfolio selection theory and tried to capture the dynamics in the optimal behavior of investors. The last two papers (Constantinides (1986) and AMY(1992)) tried to infer implications for the equilibrium returns on assets. Both models focus on a "representative" investor and investigate her specific problem in choosing between one liquid and one

illiquid asset. They both do not offer an equilibrium model and they do not try to differentiate between several risky and illiquid assets. Constantinides (1986) restricts his models to a riskless and a risky asset, whereas AMY only consider riskless assets, they abstract from return uncertainty. Interestingly, the models come to different conclusions. Constantinides (1986) states that the effect of transaction costs on asset returns is of second order in magnitude and not sufficient to explain an abnormality like the size effect documented by Banz (1981) and Reinganum (1981). AMY in contrast come to the conclusion that transaction costs may play a significant role in determining asset returns, respectively liquidity premia. Which conclusion is correct is an empirical question and thus I turn now to work which offers empirical evidence for the hypotheses derived from the theory.

2.2. Theoretical Work with Empirical Evidence

2.2.1. Amihud and Mendelson (1986)

In an important paper Amihud and Mendelson (1986)–AM86–investigate the effect of illiquidity, measured by the bid-ask spread on asset returns. The economy of their model has investors with different expected holding periods and there are assets with different relative spreads. They find that the expected returns are increasing in the relative bid-ask spread and that there is a clientele effect, that is investors with longer time horizons will prefer assets with higher transaction costs. The expected returns net of transaction costs increase with the holding period, this implies that higher-spread assets yield higher net returns (if held long enough). Investors with longer expected holding periods can increase the net return on their investment by investing in higher-spread assets.

They test the hypothesis that the expected asset returns are increasing in the spread and that this relationship is concave. The evidence provided by AM86 supports this hypothesis. They point out that this relation between asset returns and the bid-ask spread is not an anomaly of market inefficiency, rather it is due to rational investor behavior.

The model. AM86 assume M investor types ($i = 1, 2, ..., M$) and $N + 1$ assets ($j = 0, 1, ..., N$) with relative spreads S_j, where $S_0 = 0$. The spreads are increasing in j. Asset 0 has unlimited supply and there is one unit available for each of the assets $1, ..., N$. Com-

petitive market makers quote bid and ask prices, the expected inventory position is zero. There is no uncertainty in the value of the firm, i.e., the uncertainty in asset returns is due to the uncertainty in holding period rather than to uncertainty about the underlying value. The holding period of each type-i investor, T_i, is exponentially distributed with mean $E[T_i] = 1/\mu_i$. Each investor enters the market, buys assets and liquidates his portfolio at T_i, and leaves the market, thus, each investor solves a single-period investment problem. Type-i investors arrive according to a Poisson process with arrival rate λ_i. Investors maximize their expected discounted net cashflows, that is they are risk-neutral. The expected present value of holding portfolio i is:

$$\begin{aligned} &E_{T_i}\left\{\int_0^{T_i} e^{-\rho y}\left[\sum_{j=0}^{N} x_{ij} d_j\right] dy\right\} + E_{T_i}\left\{e^{-\rho T_i}\sum_{j=0}^{N} x_{ij} V_j (1-S_j)\right\} \\ &\qquad = (\mu_i + \rho)^{-1} \sum_{j=0}^{N} x_{ij}\left[d_j + \mu_i V_J (1-S_j)\right]. \end{aligned} \tag{2.15}$$

where ρ is the spread-free, risk-adjusted rate of return on asset 0.[11] x_{ij} denotes the quantity of asset j in investor i's portfolio. Each asset pays a perpetual cash flow of $\$d_j$ per unit of time. V_j is the quoted ask price for asset j and $V_j(1-S_j)$ the respective bid price. Assuming the investors to be price-taking, the objective function for a type-i investor becomes

$$\begin{aligned} \max \quad & \textstyle\sum_{j=0}^{N} x_{ij}\left[d_j + \mu_i V_J (1-S_j)\right], \\ s.t. \quad & \textstyle\sum_{j=0}^{N} x_{ij} V_j \;\leq\; W_i \\ & x_{ij} \;\geq\; 0 \quad \text{for all} \quad j = 0, 1, 2, ..., N, \end{aligned} \tag{2.16}$$

where W_i is the investor i's wealth at the beginning of the period. The constraints are the budget constraint and the restriction of no short sales. The market clearing condition reads:

$$\sum_{i=1}^{M} m_i x_{ij} = 1, \qquad j = 1, 2, ..., N, \tag{2.17}$$

where $m_i = \lambda_i/\mu_i$ is the mean of a Poisson process and denotes the expected number of type-i investors in the market. AM86 rely on Gale (1960) to state that their model has an equilibrium allocation and a unique price vector. They solve the model using the expected

[11] As noted, the values of the firms, or at least the bid and ask quotes of the dealers, are assumed to be constant over time. It is not clear why AM adjust for risk in determining ρ for asset 0, since they do not adjust for risk on the other assets. Implicitely, they assume that all assets bear the same risk, and in their model they even assume that there is no risk at all in the firm values.

spread-adjusted return of asset j to investor-type i:

$$r_{ij} = d_j/V_j - \mu_i S_j, \tag{2.18}$$

where d_j/V_j is the gross return and $\mu_i S_j$ is the spread-adjustment (expected liquidation cost per unit of time). Note, that since AM86 assume the firm value, respectively the quoted bid and ask process, to be constant over time, there are no capital gains, the entire return is distributed through the known perpetual payment stream d_j. The investor i selects the assets j for his portfolio which provide him the highest expected spread-adjusted return (for a given price vector). This return is given by

$$r_i^* = \max_{j=0,1,2,...,N} r_{ij.} \tag{2.19}$$

Interestingly, it is true that $r_1^* \leq r_2^* \leq r_3^* \leq ... \leq r_M^*$. This implies that the expected spread-adjusted return on a portfolio increases with the expected holding period and thus the longer her expected holding period the higher returns net of transaction costs an investor will earn. The equilibrium gross return $(r_i^* + \mu_i S_j)$ on an asset j can be found by finding its highest-valued use, and that is in the portfolio with smallest required return. It follows for the equilibrium (ask) prices:

$$V_j^* = \max_{i=1,2,....M} \left\{ d_j/(r_i^* + \mu_i S_j) \right\}. \tag{2.20}$$

AM86's model and its solution has two implications:

1. *Clientele Effect:* In equilibrium, high-spread assets are allocated to portfolios with longer (or the same) expected holding periods.

2. *Spread-Return Relationship:* The observed market (gross) return is an increasing and concave piecewise-linear function of the (relative) spread. (in equilibrium)

Note also, that each investor will only hold a limited number of assets, never all. This may be due to the fact that uncertainty in the firm values are ignored, and thus diversification effects which is substantial to portfolio selection theory cannot occur.

AM86 focus on testing the second implication.[12] The intuition seems clear, the positive association between expected return and spread is compensation for the trading costs. The concavity of this relation is based on the above mentioned clientele effect (1). Since high-spread assets are held by investors with longer (expected) holding periods and transaction costs are amortized over the holding period, the required compensation for the trading costs is relatively smaller. Furthermore, AM86 show that the price of an asset is a decreasing and convex function of the relative transaction costs. This relation will also hold in a more general setting as long as a longer investment horizon mitigates the impact of transaction costs, since transaction costs can be amortized over this longer holding period.

Empirical Tests. AM86 conduct an empirical study for the second implication, the relation between return and spread. They use data of NYSE stocks. The hypothesis is:

$\mathbf{H_0}$: *The expected return is an increasing and concave function of the spread.* (2.21)

The spread variable, S, is the average of the beginning and the end-of-year relative spreads for each year. The relative spread is simply the dollar spread divided by the midpoint. Their spread data covers the years 1960-1979 and the H_0 is tested over the period 1961-1980. AM86 test H_0 by testing the relation between stock returns, relative risk measured by β,[13] and spread. That means AM86 use the spread as the only variable to measure transaction costs, they omit brokerage fees by pointing on the almost perfect correlation between spread and brokerage fees found by Stoll and Whaley (1983).

AM86 employ the portfolio selection methodology introduced by Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), and Black and Scholes (1974) for their cross-sectional analysis. They form portfolios grouped by spread and relative risk. First, they divide the data in twenty overlapping subperiods ($n = 1, 2, ..., 20$) of eleven years. Each of the subperiods are divided in a five-year β estimation period, a five-year portfolio formation period, and a one-year cross-section test period.

[12] The clientele effect is certainly not easily testable.

[13] AM use β to adjust for risk according to the CAPM. Later work by Fama and French (1992, 1993) may suggest to use different variables for the purpose of risk-adjustment.

- E_n *β estimation period:* AM86 use the market model to estimate the β coefficient.

$$R^e_{jt} = \alpha_j + \beta_j R^e_{mt} + \epsilon_{jt}, \qquad t = 1, ..., 60, \tag{2.22}$$

  where R^e_{jt} and R^e_{mt} are the monthly excess returns on stock j and the market index, respectively. The market portfolio is equally-weighted.

- F_n *Portfolio formation period:* The stocks are ranked in seven groups according to their relative spread. Each spread group then is divided in equal subgroups according to their β coefficients from E_n. This gives 49 equal-sized portfolios. As the next step AM86 estimate β for each portfolio:

$$R^e_{pt} = \alpha_p + \beta_p R^e_{mt} + \epsilon_{pt}, \qquad t = 1, ..., 60, \quad p = 1, ..., 49, \tag{2.23}$$

  where R^e_{pt} is the (arithmetic) average excess return of the assets in portfolio p in month t. Then, they determine the portfolio spread S_{pn} by averaging the spreads of the last year in F_n across the stocks in portfolio p. This procedure gives 980 portfolios characterized by (β_{pn},S_{pn}).

- T_n *Cross-section test period:* The relation between R^e_{pn} (average monthly excess return on portfolio p in T_n), β_{pn} and S_{pn} across portfolios.

AM86 document that the β's and the excess returns are increasing with the relative spread. The test methodology employs covariance analysis and pooling of cross-section and time-series data.[14] They introduce two sets of dummy variables, 48 portfolio dummies and year dummies to capture cross-sectional variations as well as differences over time. An important part of H_0 is the fact that the return-spread relation is concave, i.e., the slope declines. Thus, AM86 introduce variables S^i_{pn} ($i = 1, 2, ..., 7$, spread-group index) as $S^i_{pn} = S_{pn}$ if in spread group i and zero otherwise to allow for different slope coefficients across spread groups.

First, they run OLS regressions:

$$R^e_{pn} = \underset{(9.17)}{0.0040 + 0.00947}\, \beta_{pn} + \sum_{n=1}^{19} d_n DY_n + e_{pn}, \tag{2.24}$$

[14] See Judge, et al. (1980, ch. 12-2), Kmenta (1971, ch. 14), and Maddala (1977, ch. 8).

and

$$R^e_{pn} = 0.0036 + \underset{(6.18)}{0.00672}\,\beta_{pn} + \underset{(6.83)}{0.211}\,S_{pn} + \sum_{n=1}^{19} d_n DY_n + e_{pn}, \tag{2.25}$$

where DY_n are the year dummies, one in year n and zero otherwise. The excess returns are increasing in β and in the relative spread.

Next, they estimate their complete model using OLS as well as GLS.[15]

$$R^e_{pn} = a_0 + a_1\beta_{pn} + \sum_{i=1}^{7} b_i \widehat{S}^i_{pn} + \sum_{i=1}^{7}\sum_{j=1}^{7} c_{ij} DP_{ij} + \sum_{n=1}^{19} d_n DY_n + e_{pn}, \tag{2.26}$$

where $\widehat{S}^i_{pn} = S^i_{pn} - \overline{S}^i$ if portfolio (p,n) is in spread group i and zero otherwise ($\overline{S}^i$ is the mean spread for the ith spread group). DP_{ij} are the portfolio dummies, being one if the portfolio is in group (i,j) and zero otherwise. This seemingly complicated setup is necessary to separate the effects of β and the spread group, respectively. To further investigate the effects of β and spread, they regress the coefficient c_{ij} on spread and β group dummies:

$$c_{ij} = \alpha + \sum_{i=1}^{6} \gamma_i DS_i + \sum_{j=1}^{6} \delta_j DB_j + e_{ij}, \tag{2.27}$$

where DS_i $(i = 1, ..., 6)$ are the spread dummies and DB_j $(j = 1, ..., 6)$ are the β dummies.

The results from estimating (2.26) and (2.27) and various other model specifications for subperiods and the entire period all support H_0 and show that the relationship between stock returns and the spread as a measure for transaction costs is increasing and concave.

As next step, they analyze if the results are due to the size effect documented by Banz (1981) and Reinganum (1981), since there is a negative relation between spread and firm size. AM86 include a $SIZE$ variable–market value of the firm's equity in million dollars at the end of the year before the test period–and find that the effect is negligible and highly insignificant. Then, they include $ln(SIZE)$ for the case of non-linear effects and reestimate (2.25) and (2.26). Again, they find that the size effect is insignificant, whereas the risk and the spread effect are still prevailing. Hence, they conclude that the spread effect cannot be explained by the size effect, rather the size effect may be explained by the spread effect, and the firm size is a proxy for liquidity. Thus, the size effect is rather a rational response of the market than an market inefficiency. Similar conclusion are drawn by Stoll and Whaley

[15] Cross-sectional heteroskedasticity as well as cross-sectional correlations call for GLS.

(1983) (see below). AM86, however, go one step further in that they do not fix an investment horizon, the holding periods are endogenously determined in their model. Stoll and Whaley try to explain the size effect with the bid-ask spread, whereas AM86 try to explain expected returns and find that spread-adjusted returns are not only dependent on risk and spread, but also in the investment horizon.

AM86 point on the seasonality in the size effect, which is found to be particularly strong in January, see Schultz (1983). This raises the question if there is also a seasonality in the liquidity. They cannot investigate this issue, since they are lacking the necessary monthly spread data (see below Brennan and Subrahmanyam (1994)).

The main results of the paper are that (i)average returns are an increasing function of the spread, (ii)asset returns net of transaction costs are increasing with the spread and this leads to a (iii)clientele effect, that is, investors with longer expected holding periods tend to choose high-spread stocks, and (iv)the relation between return and spread is concave, due to the clientele effect.

AM86 point out several extensions for their model. Their results suggest that liquidity-increasing financial policy may increase firm value. Hence, the spread could be endogenized in the model rather than taking it as exogenously given. One could also distinguish between total and marginal liquidity of assets, that is the liquidation in a portfolio context compared to the liquidation of an asset itself. In reality, investors have not one liquidation point, rather they have several cash demands over their investment horizon. They also hold diversified portfolios rather than some few assets. This suggests that an investor may sell off only parts of his portfolios, maybe only specific stocks. These effects can enrich the model.

Further extensions and critique can be found. Today, monthly data for spreads are available, thus, the analysis could be improved by utilizing these improved data sets. In addition, a possible seasonality in the liquidity of assets could be investigated. The risk-adjustment could follow the lines of Fama and French (1992, 1993), like in Brennan and Subrahmanyam (1994) or James and Woodward (1995), although the Fama-French factors include a size factor, which could be already a proxy for a spread effect. In addition, the test design would have to change considerably. The AM86 model does not include any return uncertainty in the sense that the future value of the firm or the future bid and ask prices are uncertain. This implies that the investors do not face a traditional portfolio selection

problem, rather they hold only a few number of assets rather than a well-diversified portfolio. Introduction of this kind of uncertainty along with risk-aversion of investors could move the model closer to portfolio selection theory and reality. Together with the mentioned addition of several (partial) liquidation points in time the model could provide valuable insights, although the solution may be considerably more complicated and there may be no closed form solution. They also require 11 years of data for a stock to get included. This requirement introduces a severe survivorship bias, as pointed out by Eleswarapu and Reinganum (1993).

2.2.2. Brennan and Subrahmanyam (1994)

The discussion of the previous paper was thorough, since the model presented in Section 3 is based on the work by Amihud and Mendelson (1986). In addition, the AM86 model is fairly influential on other work and thus deserves a detailed discussion. In the following discussions, I try to abstract from details and focus on the results and implications.

Brennan and Subrahmanyam (1994)–BS–investigate the relation between the expected rate of return on assets and market illiquidity due to adverse information. The logic goes as following: Private information leads to adverse selection costs for uninformed investors; see e.g., Glosten and Milgrom (1985), Kyle (1985), and Easley and O'Hara (1987) as early work in this area. Because of these higher costs, the investors demand higher rates of return on assets which carry severe informational asymmetries.

The BS model is a single representative investor economy. In their empirical analysis, they adjust for risk via the three Fama-French factors.[16] They also adjust for the effects of the quoted bid-ask spread. Note that the bid-ask spread itself is a measure for illiquidity; see Amihud and Mendelson (1986). However, BS find there is still a significant relation between expected rate of return and their measure of illiquidity. To measure illiquidity they basically use the price impact of a trade, which is the inverse of the market depth, λ, as introduced by Kyle (1985).

The single representative investor maximizes utility over the mean and the variance of terminal wealth. There are one riskless asset and N risky securities. To buy x_j shares of a security, he has to pay a price per share of $P_j + \lambda_j x_j$. $\lambda_j > 0$ is the slope of the exogenously

[16] see Fama and French (1992, 1993)

given supply curve the investor faces for security j, or the inverse of the market depth. Note in a competitive market $\lambda_j = 0$. They solve the maximization problem of the representative investor and find the market equilibrium condition

$$E[\widetilde{R}_j] = r_f + h\beta_j + \frac{2\lambda_j n_j}{P_j} r_f, \tag{2.28}$$

where $E[\widetilde{R}_j]$ is one plus the expected return on security j, and $\beta_j \equiv \frac{cov(\widetilde{R}_j, \widetilde{R}_M)}{\sigma^2_M}$,the usual beta. $h > 0$, the market risk premium per unit of beta risk. BS define $C \equiv \frac{\lambda_j n_j}{P_j}$ as the liquidity cost due to adverse selection. It represents the expected compensation for illiquidity of the security. They derive this simple model to illustrate, in their empirical analysis, they replace β by the Fama and French factors, a value-weighted market index, a size portfolio, and a book-to-market ratio portfolio.

BS use the transactions on the ISSM tape for the calendar year 1984 for all NYSE/ AMEX listed securities to estimate λ_j for 1984-1987 and the transactions of 1988 to estimate λ_j for 1988-1991. They form 30 portfolios assigned to five size groups and six λ_j groups. In the analysis the equally-weighted returns for the portfolios are used. The proportional spread is the average across all observations in a year. Note, BS estimate λ_j only in two years due to restriction in their data, this implies that they assume C to be an intertemporal constant.

They employ two different methods to estimate the market depth λ_j. The first one is based on Glosten and Harris (1988):

$$\Delta p_t = \lambda q_t + \psi[D_t - D_{t-1}] + y_t, \tag{2.29}$$

where p_t is the transaction price, q_t the order flow, D_t denotes the sign of the incoming order at time t (buy order: +1, sell order: -1), and y_t is the unobservable error term, the public signal. Note, BS include fixed transaction costs with ψ, thus their model captures fixed costs and adverse selection cost components, but they ignore inventory effects in (2.29).[17]

The second approach is due to Foster and Viswanathan (1993) and Hasbrouck (1991). This approach focuses on unexpected volume as measure of the adverse selection component.

[17]BS argue that empirical findings suggest that the inventory cost component is small and refer to Stoll (1989), George, Kaul, and Nimalendran (1991), and Madhavan and Smidt (1991).

The model consists of a two equation system, one equation for the quantity and one for the price change:

$$\begin{aligned} q_t &= \alpha_q + \sum_{j=1}^{5} \beta_j \Delta p_{t-j} + \sum_{j=1}^{5} \gamma_j q_{t-j} + \tau_t, \\ \Delta p_t &= \alpha_p + \psi[D_t - D_{t-1}] + \lambda \tau_t + \nu_t, \end{aligned} \tag{2.30}$$

where τ_t measures the informativeness of trades. Again, inventory effects are ignored and only adverse selection costs and fixed costs (ψ) are captured. BS estimate λ for all NYSE/AMEX firms available on ISSM tapes for the years 1984 and 1988 with both approaches using OLS. The resulting estimates are used to form the portfolios and to estimate the cost of liquidity, C.

Different from their model in (2.28), BS use the three factor model proposed by Fama and French (1993) as the null hypothesis. The factors are the market excess return, a size factor (the return on a portfolio long in small stocks and short in large stocks), and a book-to-market ratio factor (the return on a portfolio long in high book-to-market ratio and short in low book-to-market ratio stocks). BS test the model for the 30 portfolios using OLS regressions of the excess returns:

$$R_{it} = \beta_{i0} + \beta_{iM} R_{mt} + \beta_{iSMB} SML_t + \beta_{iHML} HML_t + e_{it}, \qquad i = 1, ..., 30. \tag{2.31}$$

If the Fama and French factors are sufficient to explain returns the intercepts of the regressions should be jointly equal to zero:

$$H_0 : \qquad \beta_{i0} = 0, \qquad i = 1, ..., 30. \tag{2.32}$$

BS can reject this hypothesis at the 5% significance level. The Fama and French factors are not sufficient to explain the cross-sectional variations in average returns on the portfolios grouped by firm size and illiquidity. The question is what is missing in (2.31). BS argue that the (or one) missing factor maybe the illiquidity costs of the securities.

To test this assertions, they perform GLS regressions which contain the Fama and French factors as explanatory variables as well as various measures for the costs of illiquidity. To avoid errors in variables problem, they estimate the factor coefficients and the coefficient of the cost of illiquidity measure simultaneously. First, they include dummy variables for the different λ groups next to the Fama-French factors. They find that the dummy coefficients

are monotonically increasing from low to high λ quintiles. Hence, the differences in the portfolio intercepts of the model in (2.31) are at least partially due to differences in costs of illiquidity. They then investigate the functional form of the cost of illiquidity. They include C, the liquidity cost due to adverse selection, into the regression and find the coefficient positive, but only marginally significant. Next, they include $\ln C$, respectively C and C^2 and find high significance. These results are consistent with the cost of illiquidity variable. However, when including the spread as additional variable, the coefficient for the spread is negative and strongly significant (together with C), even when employing C and C^2 in the regression, the spread variable remains negative and significant. This is not consistent with the role of the spread as measure of illiquidity. This hints on some form of misspecification and BS conclude that their single representative investor model is not rich enough to explain fully the effect of illiquidity on returns.

Interestingly and consistent with Eleswarapu and Reinganum (1993), BS find that C, the cost of illiquidity plays a significant role in explaining returns only in January. BS suggest that this may be due to seasonalities in λ. They are restricted in their data to two years and report a February rather than a January seasonal in λ. This implies that the seasonality in the relation between return and cost of illiquidity is not due to a seasonal of λ. However, the question remains open in light of the limited data.

In conclusion, BS provide evidence that the compensation for adverse selection plays a significant role in explaining required rates of return in the equity market. However, the negative coefficient of the spread points to some form of misspecification in their model. There could also arise a conflict between C and λ,since both variables are measures for transaction costs (due to illiquidity) and thus they are certainly correlated. They are not able to explain the seasonality in the compensation for adverse selection costs in their data. They ignore inventory costs and they also do not give any motivation for trading, since it is an equilibrium model and in equilibrium there is no incentive to trade. Another point is that the supply side is assumed to be exogenously, the market making process is not modelled. They use their estimates of λ and thus C for four consecutive years, the assumption of intertemporal constant costs of illiquidity seems to be strong. More important, they employ the Fama-French factors as null hypothesis, but these factors could be proxies for market microstructure rather than for risk. There could be many more factors which should be

included, Fama and French leave the question of economical reasons for their factors mainly open.

2.2.3. James and Woodward (1995)

This paper tries to relate insider trading and the expected returns on stocks. Different from earlier work James and Woodward (1995)–JW–do not focus on the question if insider trading should be prohibited or not, rather they analyze the current "limited informed insider trading allowing system" and its implications. Intuitively, there are two effects of insider trading. First, insider trading releases information to the market maker which only speculators can acquire, thus it reduces informational asymmetry and the expected profits for informed speculators. Due to the lower expected losses to informed speculators, the market makers demand lower compensation. As a result insider trading lowers the firm's cost of capital.[18] Second, insider trading signals that insiders have valuable information. It gives speculators the incentive to acquire information and become informed, thus the informational asymmetry increases and with it the expected profits of informed speculators. The market makers demand higher compensation and the firm's cost of capital increases. These are two offsetting effects and the net effect is unclear.

JW distinguish between opaque and transparent firms. For opaque firms, it is valid that insiders generally posess valuable private information. This implies that speculators tend to always acquire information. The first effect dominates and the firm's cost of capital decreases with insider trading. In contrast, for transparent firms, it is true that insiders generally do not posess private information. Speculators will only invest in acquiring information if insiders reveal that they have private information. Thus, the second effect dominates and the firm's costs of capital will increase with higher insider trading activities. JW consider small firms as opaque and large firms as transparent.

They also point out the difference between dealer markets and a specialist exchange. In a dealer market speculators can trade with each dealer in a specific stock, where on a specialist exchange, they can only trade with one specialist per stock. Thus, JW conclude

[18] Investors do not care who gets the money, they are interested in what rate of return they can expect (net of transaction costs). Thus, if the market maker requires a high compensation, the firm gets less of the money investors are willing to pay when acquiring the stock. Its cost of capital increases.

that the total costs (informed) speculators impose on the market are higher in a dealer market.

In their theoretical model, they employ the following assumptions. Insider trading itself does not matter, i.e., the investment decision of a firm is not affected by information released by insider trading, and it has no direct impact on liquidity (only indirect via speculators). This means that the market maker does not consider insider trading by itself in her optimization problem. Insiders are potentially informed, but not always. Some outsiders are able to acquire information, the speculators, that implies speculators are potentially informed. This introduces three groups, insiders, speculators (both potentially informed), and investors, which rely on public information and are not able to acquire information. Each firm offers an uncertain terminal payoff. As a very interesting point, JW abstract totally from the market making process, there is no bid-ask spread, rather the market maker charges an initial fee at the public offering of the shares at $t = 0$. After the initial offering, they allow for two subsequent trading rounds. The model setup is the following:

t_0 : The firm issues shares which yield an uncertain dividend D in t_5; P_{IPO} to firm and fee c_{Trade} to market maker.

t_1 : With probability K, Insiders learn the value of dividend D in t_5.

t_2 : *First trading round:* Market maker sets price P_2 at which she is willing to buy or sell. Insiders and investors may trade.

t_3 : Direction of trading by insiders becomes public. Speculators may pay c_{Info} to learn what insiders know.

t_4 : *Second trading round:* Market maker sets price P_4 at which she is willing to buy or sell. Investors and speculators may trade.

t_5 : Each share yields a dividend D of either \$1 or \$0.

This model implies that the higher the market maker's losses to informed traders, the higher the fee she charges in the initial offering and the less the firm gets per share. The equilibrium return the firm must offer increases with the market maker's fee.

JW solve this problem employing Bayes law and standard techniques of the literature in the area of informational based market making models (see e.g., Kyle (1985) or Easley and O'Hara (1987)). The solution gets fairly complicated and I refer to the original paper. Their model has the following implications:

1. Insider trading lowers the cost of capital for small firms. (opaque firms)
2. Insider trading raises the cost of capital for large firms. (transparent firms)
3. The cost of capital increases with the probability that investors posess (valuable) private information.
4. Insider trading costs are higher in a dealer market than on a specialist exchange.

Furthermore, firms reduce their investment when they face higher cost of capital, thus welfare overall will decrease with higher cost of capital.

JW test these implications empirically by regressing year t returns on end-of-year $t-1$ empirical measures of the model parameters and controls. To adjust for risk, JW, like Brennan and Subrahmanyam (1994), use the Fama-French factors and not the traditional β. As mentioned they consider small firms as being opaque and large firms as being transparent. As a measure for the probability K, that insiders have private information, they use $\ln(MV)$ and $(R\&D/MV)$, since K increases in both, the size of a firm and its R&D intensity. Moreover, they develop an interesting measure for the probability that insiders engage in informed trading, T_I. JW use the proxy $PITAR$ (**P**robability that **I**nsiders **T**rade before **A**bnormal **R**eturns), that is, they estimate the probability that insiders were trading in the profitable direction in the two month window previous to a two month window with an average absolute abnormal return by employing a logit regression.[19] As further controls, JW use year dummies and exchange dummies (NASDAQ vs. NYSE/AMEX). Their data is restricted to two years (1991, 1992) and they run OLS regressions with White adjustment for the standard errors. JW specify several models which include different explanatory variables and controls and run them on different subsamples.

Their findings seem all to support their hypotheses (1-4 above). And they are able to show that their regression results are not driven by losses to insiders which are ignored

[19]For details see Table 3 in James and Woodward (1995).

in their theoretical model. Furthermore, JW show that the results are indeed economically significant, e.g., for small firms they document an increase in expected return from 18.7% to 36.2% as they move from high to low level $PITAR$. The opposite is true for large firms, NASD firms' expected returns decrease from 21.4% to 7.7% as $PITAR$ decreases.

Thus, the optimal policy regarding insider trading depends on market regulations and firm transparency. JW also show that the nature of the exchange matters if insider trading is possible. This has some interesting implications for regulation policy. If the market regime creates transparent (opaque) firms, then limiting insider trading reduces (raises) the cost of capital and thus improves (lowers) social welfare. A comparison between Germany and the US seems to support their view, the US has fairly strict reporting regulations compared to Germany and also has much stricter insider trading rules (actually, Germany had no insider trading rule at all until 1992). Another implication is, that developed countries should be careful with "exporting" their insider trading rules to developing markets.

The paper by JW offers an interesting way to model the impact of insider trading, the market maker imposes an initial fee, but JW abstract totally from the explicit market making process, they also ignore inventory effects. The measure $PITAR$ which they develop is an appealing alternative of measuring asymmetric information compared to volume measures like the ones in Brennan and Subrahmanyam (1994). The results have important implications for the real markets. However, there are some facts which call for criticism. First, there is a huge step from their model to the empirical analysis and the real world. Their model seems very unrealistic. They ignore inventory costs and order processing costs as well as possible losses to insiders (although they show that these are not driving their empirical results). It should also be questioned if it were not valuable to model the market making process explicitly and why the market maker is not able to acquire information.

Their empirical analysis is based on only two years of data, a better sample may improve the power of their tests. They never motivate the inclusion of a year dummy as control. Furthermore, all controls are insignificant, which may lead to the conclusion that there may be some form of misspecification. The major point of critique is certainly the low power of their tests, all regressions yield an extremely low R^2 and the significance of the coefficients of variables they are focusing on seems rather low. Again, this could point towards some form of misspecification, on the other hand a bigger sample for more than two years may

help improving the power of their analysis.

The next section will move on to purely empirical papers. There are different views on the relation between market micro structure and asset returns, intuitively as well as formally as shown in sections 2.1 and 2.2, especially Constantinides (1986), AMY(1992), and Amihud and Mendelson (1986). Thus, in the end the question is an empirical one.

2.3. Empirical work

2.3.1. Stoll and Whaley (1983)

The single-period, two-parameter CAPM model was questioned by Banz (1981) and Reinganum (1981), who found abnormally large risk-adjusted average returns for small firms. Stoll and Whaley (1983)–SW–suggest transaction costs as a "missing factor" in the single-period, two-parameter CAPM. They find that total market value is inversely related to risk-adjusted returns, an inverse relation between price per share and risk-adjusted returns, and that transaction costs can (at least partially) explain the former anomalies. SW employ arbitrage portfolios in their empirical analysis.

The data sample consists of NYSE common stocks and covers the period from January 1955 until December 1979. They form 10 equally-weighted test portfolios ranked by market value of outstanding shares at the beginning of each year from 1960 until 1979, and estimate the β by employing the market model over five years preceding the respective year under consideration. This yields 10 portfolio time series for the years 1960-1979. The proportional spread of the stocks is calculated by taking the arithmetic average of the relative spreads at the beginning and at the end of the year.

Considering the raw returns on the portfolios (without adjustment for risk), SW find that the mean realized returns decrease as the market values of the portfolio stocks increase. Furthermore, they find that the estimated relative risk coefficients of the portfolios decrease with the market value as well. For the estimation of the relative risk coefficient, they use a value-weighted as well as a equally-weighted index in

$$R_{pt} - R_{ft} = \alpha_p + \beta_p(R_{mt} - R_{ft}) + \varepsilon_{pt}, \qquad t = 1, ..., 240, \tag{2.33}$$

the regression of monthly portfolio excess returns on the monthly market excess return. Both indices yield the decreasing relation between $\widehat{\beta}_p$ and market value of the stocks in the portfolios. SW argue that the equally weighted index is to prefer since the portfolios are equally weighted and the average relative portfolio risk would equal 1 (with the value-weighted index, it is 0.934). They also document a strong relation between price per share and market value of outstanding shares.

To test for a small firm effect, SW construct arbitrage portfolios which have a relative risk coefficient of zero. In forming these arbitrage portfolios, they use the beginning-of-year estimates for β_p, this does not ensure that the arbitrage portfolios necessarily have zero relative risk over the year. To capture this effect, SW run the regression

$$R_{at} = \alpha_a + \beta_a(R_{mt} - R_{ft}) + \varepsilon_{at}. \tag{2.34}$$

The intercepts of these regressions are the estimated abnormal returns realized by the arbitrage portfolio. They find that small firms outperform large firms by about 12% p.a.

Next, they investigate if these abnormal return may be purely statistical biases. The statistical biases are due to infrequent trading. The close price of a security represents the last trade before the closure of the market and does not have to occur at the time of the closure of the market. This leads to positive correlation in the market return series and thus the estimated market return variance is downward biased, as is the covariance between market return and stock return. It follows that the estimators for the relative risk coefficient is downward biased. Roll (1981) documents a strong relation between trading frequency and firm size, and suggests that the size effect is due to this relation and the described statistical bias. SW argue that this bias will be severe for daily data, but it should be less significant for monthly data. They use Dimson's (1979) estimator of the relative risk coefficient (which uses lagged return premia) and compare their results with the results obtained when using the "simple" estimates. They find a statistical bias, but this bias is small and cannot explain the small firm effect.

SW repeat the same experiment, but now they rank the portfolios by share price rather than market value (recall the positive correlation between market value and price per share). They find the same inverse relation as with market values, the mean realized returns decrease

with increasing prices per share. The difference between the highest price per share portfolio and the lowest price per share portfolio is about 10% p.a., a little less than the difference found for the size portfolios.

The above described results seem to imply that investors can earn abnormal returns. However, transaction costs (dealers' bid-ask spreads and brokers' commissions) may prevent investors from earning these abnormal returns. SW find that both the average percentage spread and the percentage commission rate are decreasing with higher market value of the stocks in the portfolios. That is small firms have higher percentage trading costs than large firms.

SW ask if there is a small firm effect net of transaction costs. If there is than investors could indeed earn abnormal returns after transaction costs (what investors are interested in). To investigate this question, they apply the two-parameter CAPM to after-transaction-cost returns:

$$R_{jt}^{\tau} = (1 + R_{jt})(1 - F_{jt})/(1 + F_{jt}) - 1, \qquad t = 1, ..., 240, \tag{2.35}$$

where R_{jt}^{τ} is the after-transaction-cost rate of return on stock j in month t, R_{jt} is the before-transaction cost rate of return, and F_{jt} are the proportional transaction costs which comprise the proportional bid-ask spread as well as the commission rate for the stock. They repeat their methodology and construct arbitrage portfolios. They find that the market value effect is reversed. After transaction costs, the largest firms outperform the smallest firms by about 17% p.a. Furthermore, small firms have significantly (at 5%) negative abnormal returns after transaction costs.

They also repeat their analysis for the low price effect. Including transaction costs in the analysis yields a similar result for the portfolios grouped by prices per share. The effect is reversed compared to the case without transaction costs. The highest priced stock portfolio outperforms the lowest priced stock portfolio by about 28% p.a. and the lowest priced stocks have significantly (5%) negative abnormal returns.

SW point out that the results are sensitive to investment horizons. The usage of monthly returns implicitly implies that investors incur transaction costs each month on all of their stocks. With longer holding periods it is expected that the negative abnormal returns of small firms may disappear and the Banz's (1981) respectively Reinganum's (1981) results

may reappear. Due to reduced number of observations with longer holding periods, SW restrict their analysis to holding periods up to one year and the smallest stock portfolio. They find that the abnormal returns on small firms become positive (though not significant) as the holding period is increased.[20] With a holding period of about 4 months, the portfolio containing the lowest market value stocks just breaks about even, i.e., yields approximately the same return after transaction costs as the market.

In summary, SW show that there is a small firm effect in returns before transaction costs, and they also find a low price effect. However, introducing transaction costs, using one-month returns, they are able to show that the smallest firm portfolio have a significant negative abnormal return after transaction costs. Employing longer periods, this effect diminishes and the abnormal returns on the smallest firms become positive, but insignificant for holding periods between 3 and 12 months. They conclude that for returns before transaction costs, the joint hypothesis of market efficiency and CAPM can be rejected. (Banz (1981) and Reinganum (1981) reject the CAPM as misspecified, they believe in market efficiency.) However, since SW cannot find significant abnormal returns after transaction costs for small stocks for longer holding periods, they find their data to be consistent with the CAPM (applied to after-transaction-cost returns).

The sensitivity of the results on the holding period suggests some kind of clientele effect to achieve an equilibrium, SW do not investigate this issue any further. In addition, for holding periods between 3 and 12 months, there are no significant abnormal returns on small firms. They do not analyze if it is possible to earn significant abnormal returns for holding periods of 2 or more years.[21] This goes in the direction of the first comment: How does the market achieve an equilibrium allocation, and what are the holding periods of actual investors?[22] Furthermore, SW do not address the seasonality issue of the small firm effect: Why are the abnormal returns (before transaction costs) on small firms found mainly in January (even in the first week of January)?[23] In their analysis, SW use the average of the beginning-of-year and the end-of-year proportional spread as estimation of the spread, it may be more appropriate to use monthly spread data. The procedure proposed by Dimson

[20] That is, they outperform the equally weighted market index.

[21] See also Day et al. (1985).

[22] The paper by Amihud and Mendelson (1986), discussed above, deals with this issue on a theoretical basis.

[23] See Eleswarapu and Reinganum (1993).

(1979) for the estimation of the betas has been shown to be incorrect. Fowler and Rorke (1983) suggest a corrected consistent estimator.

2.3.2. Day, Stoll, and Whaley (1985)

In a related study Day et al. (1985)–DSW–reexamine the size effect. As described, investors demand higher risk-adjusted returns on small firms then they do on large firms. Previous studies investigate the size effect by using size portfolios to test the SLM-CAPM. The empirical findings range from a 12-30% p.a. differential in returns on small firms compared to large firms. Possible explanations are that the phenomenon is a purely statistical artefact, that the CAPM does not hold, and that firm size proxies for missing factors (which?). A clear explanation is not provided.

Looking just at the data, it is possible that there is an exchange effect, since 98% of the stocks in the portfolio with the largest stocks are traded on NYSE; see also Reinganum (1990). DSW first consider possible statistical biases in risk-adjusted returns. There can be upward bias in the mean of the stock return due to the method of aggregation of the portfolio returns (cross-sectionally and in the time-series), i.e., geometric vs. arithmetic or rebalanced returns. Infrequent trading leads to positive correlation in portfolio returns, and smaller firms tend to trade more infrequently, thus the problem may be more severe for them. There is also bias in the systematic risk estimates due to thin trading. The closing price can occur at different points in time, adjustments for the estimator are suggested by Fowler and Rorke (1983). Roll (1981) shows that small firms are less frequently traded and thus the β estimates are more downward biased.[24] The weekly seasonal in stock returns is more pronounced for large firms than for small firms (returns on Monday tend to be negative). This bias will always be there as long as daily data is utilized. DSW suggest to use weekly data, which removes the bias due to seasonality and reduces the infrequent trading bias.

Even after adjusting for all the mentioned statistical biases, DSW find a difference between large and small firms of about 18.7% p.a., reduced but still large. So the statistical bias is not able to explain the size effect.

[24]Note, Stoll and Whaley (1983) used monthly data and found no severe biases when employing Dimson's adjustments (although incorrect).

Stoll and Whaley (1983) point out that the holding period affects small firm returns significantly. DSW estimate the holding periods for the portfolios (with three more or less limited measures). They find that average estimates for the holding periods exceed 2 years. Stoll and Whaley (1983) considered only returns up to one year. DSW are able to show that for longer periods, the differential in returns between small and large portfolios seems to stabilize. Furthermore, they document that small firm and middle-sized firm portfolios are held approximately for the same period (3 years), however, the large firm portfolios have much longer holding periods (5 and 8 years). This indicates a clientele effect.

Having shown that statistical bias is not a sufficient explanation, DSW look for economic explanations: what are the missing factors in the SLM-CAPM? They consider the following candidates: size, dividend yield, earnings yield, transaction costs, alternate risk measures, differential information, industry effects, and investor preference for higher-order moments of the return distribution. They run several regressions which include different factors and combinations of factors in the SLM-CAPM, and get the following results. The pure CAPM is slightly supported (surprisingly). Including the market value of the stock they find a negative coefficient which indicates a size effect, however, there are subperiods when the coefficient estimate is positive. The dividend yield has the expected sign for the coefficient, but it is not significant and thus DSW reject the dividend yield as a missing factor. The earnings yield is supported and the combination of size and earnings yield still yields the expected signs for the coefficients and are significant. Thus, there are two distinct effects. DSW rule out debt/equity ratio, they find inconsistent and insignificant coefficient estimates.

As found in Stoll and Whaley (1983), the bid-ask spread is an inverse function of the share price which itself is strongly positively correlated to market value. There exists a share price effect almost as strong as the size effect and the after-transaction-cost returns are sensitive to the length of the holding period. In conclusion, transacion costs could be a missing factor in the CAPM.

DSW use the share price as proxy for the transaction costs (due to lack of data) and they include the price per share in the SLM-CAPM. The results show that price per share performs marginally better than the size variable. To control this finding, they run the regression

$$R_j - R_f = \gamma_0 + \gamma_1 \widehat{\beta}_j + \gamma_2 P_j + \gamma_3 S_j + \varepsilon_j, \tag{2.36}$$

where P_j denotes the price per share and S_j the total market value of the stock. They find the coefficient γ_2 unchanged, but now $\gamma_3 > 0$. This is puzzling, but it clearly indicates that transaction costs play an important role in explaining returns.

DSW conclude that statistical biases are not able to explain the small firm effect. From all the analyzed variables, only firm size and price per share are significant when included in the CAPM. Both, size and price per share, are proxies for transaction costs, the smaller the firm and the lower its price per share, the higher the proportional transcation costs. Surprisingly, when including both variables, DSW find that price performs better in explaining returns then firm size does. The bottom line is that transaction costs seem to be able to explain the small firm effect, although the coefficient for the size variable becomes positive.

DSW leave the question of the seasonality in the small firm effect aside. They do not try to come up with explanations why the phenomenon occurs mainly in January. Furthermore, they do not include transaction costs per se in the regression, they use price per share as proxy. However, firm size has also a correlation with transaction costs. The result that price per share performs better in explaining returns could be due this. When including transaction cost measures like the bid-ask spread directly, the result may change. They regress equally-weighted portfolio returns on a value-weighted market index, although it seems more appealing to regress equally-weighted portfolio returns on equally-weighted indices.[25]

2.3.3. Amihud and Mendelson (1989)

In this paper Amihud and Mendelson (1989)–AM89–test an extension of the CAPM. Merton (1987) assumes that investors have only information about a subset of all assets and they invest only in them. This has the following implications: The expected return on an asset is

1. an increasing function of the relative risk coefficient, β,

2. an increasing function of residual risk (since portfolios held by investors are not perfectly diversified),

[25] see e.g. Stoll and Whaley (1983).

3. an increasing function of the fraction of the market portfolio invested in the specific asset,

4. a decreasing function of the fraction of all investors who buy the asset (which reflects public availability of information about the asset).

AM89 link (4) to their Amihud and Mendelson (1986)–AM86–model. Studies found that a larger number of shareholders narrows the spread. Merton (1987) sees the number of shareholders as measure for availability of information. Since bid-ask spreads also decrease with more publicly available information, the effect (4) is linked to the AM86 spread effect.

AM89 perform a joint test of all four factors above. They point out that test for individual factors have been carried out, but a correct test should test them jointly. As explanatory variables for (3) and (4) they use the market value, respectively the bid-ask spread.[26] The methodology follows Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), and Black and Scholes (1974) which is a pooled cross-sectional and time series estimation. They form 49 test portfolios ranked by size and spread (7x7). The regression equation is

$$R_{pn} = \gamma_0 + \gamma_1 \beta_{pn} + \gamma_2 \sigma_{pn} + \gamma_3 SZ_{pn} + \gamma_4 S_{pn} + \sum_{n=1}^{19} d_n DY_n + \varepsilon_{pn}, \tag{2.37}$$

where $n = 1961, ..., 1980$ and $p = 1, ..., 49$. SZ_{pn} is the average size for stock sin portfolio p, and S_{pn} is the average proportional spread (one observation for each year). R_{pn} is the portfolio excess return (average monthly excess return). They estimate this equation by OLS and GLS.

They find the coefficients for the systematic risk and for the spread to be positive and significant which is consistent with the hypothesis. The coefficient for the residual risk is negative and significant for the OLS estimation (not significant for GLS), which contradicts the hypothesis. The size coefficient is positive but insignificant and negative in certain subperiods, again, this finding is not consistent with the hypothesis.

AM89's contribution is the joint test of all four factors which yields different results from previous studies. They find that their AM86 model is supported by the data, the effect of systematic risk and the bid-ask spread (illiquidity) on returns is shown. They can

[26] They link (4) with liquidity and use their AM86 model to motivate the usage of the bid-ask spread. The bid-ask spread variable is consistent with Merton's variable in (4) and they test implicitely the AM86 model.

reject Merton's (1987) hypothesis that residual risk and size are factors in determining asset returns.[27] The logic is that residual risk as well as size effects can be diversified or will disappear, so they should not be priced, whereas systematic risk and liquidity costs cannot be diversified and will not dissipate, thus they are priced.

2.3.4. Reinganum (1990)

In this empirical study Reinganum (1990) tries to contrast NYSE and NASDAQ securities with respect to their liquidity premia. He compares liquidity premia in a specialist system with liquidity premia in a multiple dealer system. He estimates differences in liquidity premia inferred from monthly stock returns, i.e., he focuses on the long-run. In contrast, earlier work tried to estimate liquidity from bid-ask spreads in transaction data. Since firm size is a proxy for liquidity he controls for it, thus, only differences in institutional arrangements should matter. Since largest firms tend to trade on NYSE, the analysis is restricted to smaller firms, due to insufficient data for NASDAQ firms of the same size. However, the choice of largest firms to list on the NYSE indicated that for larger firms the NYSE seems to provide superior liquidity services whereas for smaller firms NASDAQ seems preferable. The analysis builds on the Amihud and Mendelson (1986) model and results.

For non-risk-adjusted returns, Reinganum finds that the liquidity premia for smaller firms are higher on NYSE stocks than on NASDAQ stocks. The adjustment for risk is in favor of NYSE stocks, however, the result remains, liquidity premia for smaller firms seem to be higher on NYSE than on NASDAQ. His results imply that no market system dominates the other, rather NASDAQ provides liquidity cheaper for smaller firms whereas the difference diminishes for larger stock, and for the largest stock the NYSE seems preferable. They also suggest that market microstructure indeed affects asset pricing on NYSE and on NASDAQ.

Aside from firm size there are other variables which are related with liquidity. Trading volume is negatively related to the spread, that is, lower volume implies higher average returns.[28] This explains the earlier results, since trading volume is higher on NASDAQ than compared to the NYSE. Roll (1984) derived a serial covariation estimate for the effective bid-

[27] This is of course in contrast to Fama and French (1992, 1993) who show that size is a significant factor in determining returns.

[28] according to Amihud and Mendelson (1986).

ask spread, $BAS = 2\sqrt{-Scov}$, where $Scov$ is the first-order serial covariance of price changes. If one believes that the documented differences in liquidity premia are due to liquidity differences, one expects Roll's measure to be greater for the NYSE than for NASDAQ. Reinganum finds this result in the data.

Reinganum also considers several risk measures which are possible related to spreads, however, he does find that the risk measures are very similar on the NYSE and on NASDAQ. The same is true for variance ratio test, which try to estimate the depth of the market, there is no significant difference between NYSE and NASDAQ. Reinganum also dismisses price reversals as explanation for the differences in (gross) liquidity premia. However, the findings for trading volume and Roll's covariance measure are consistent with the finding that NASDAQ provides higher liquidity than the NYSE.

Applying a Fama-MacBeth (1973) methodology, Reinganum controls for the above mentioned liquidity-related variables and analysis if the market-microstructure effect on asset returns is still persistent. He restricts his analysis to the second smallest firm portfolio and estimates the model

$$\begin{aligned} R_{it} &= \beta_{0t} + \beta_{1t} EXCH_{it} + \beta_{2t} Roll_{it} + \beta_{3t} acbeta_{it} + \beta_{4t} size_{it} + \beta_{5t} share_{it} \\ &\quad + \beta_{6t} price_{it} + \beta_{7t} vratio_{it} + \beta_{8t} retlag_{it} + \varepsilon_{it}, \end{aligned} \tag{2.38}$$

where the variables are an exchange dummy, Roll's implicit spread, the aggregated-coefficient beta, stock-market capitalization, number of shares outstanding, price per share, variance ratio, and stock return during the prior 12-month period (in this order). The exchange dummy represents a proxy for the market microstructure environment and he calls the coefficient of this variable the adjusted monthly differential liquidity premium–ADLP. Reinganum employs a two-step procedure, he first filters the return from any effects of variables other than the exchange dummy, that is, he calculates the residuals of 2.38 without the exchange dummy. Then he regresses these residuals on the exchange dummy.

The results suggest that differences in average returns persist after controlling for risk and some other liquidity-related variables. However, the adjusted differences are smaller than the not adjusted ones. He concludes that the exchange dummy proxies for some missing factors or indicates misspecified variables.

Summarized, Reinganum finds that the average returns on small firms are larger for

NYSE stocks than for NASDAQ stocks. For larger firms, the difference diminishes. After controlling for various risk and liquidity measures, the differences are still there. There are two explanations, the exchange dummy is a proxy for risk or for liquidity. Reinganum tends to believe in the liquidity explanation. He points out that market microstructure matters in asset pricing, and that it matters not because of the institutional differences of the NYSE and NASDAQ per se, rather, the institutional differences should be seen as proxy for fundamental underlying economic differences.

Reinganum's whole argumentation depends crucially on the reversion of the conclusion of the Amihud and Mendelson (1986) model. That is that a lower return indicates higher liquidity, rather than higher liquidity implies a higher return (everything else equal). I am not completely convinced if this step is that innocent. It certainly depends on the quality of the control for other relevant factors.

2.3.5. Eleswarapu and Reinganum (1993)

As previous mentioned, most research did not focus on seasonalities in liquidity premium, which is important since the small firm effect has a strong January seasonal and so far there is no satisfactory explanation for this behavior. Eleswarapu and Reinganum (1993)–ER–try to fill this gap and investigate the seasonal behavior of the liquidity premium. They build their analysis on Amihud and Mendelson's (1986) model–AM86.

As noted above, AM86 do not explore possible monthly seasonalities in the relation between expected return and the bid-ask spread. ER try to investigate the relation between expected return and the bid-ask spread in January and in non-January months. In addition, the results accomplished by AM86 are sensitive to the rather selective portfolio selection technique.

ER use data for NYSE stocks for the time period 1961-1990. They use the average of the beginning-of-year and the end-of-year relative spreads for their spread variable. Like AM86, they form 49 equally-weighted portfolios according to spread and systematic risk, β, (7x7). The portfolios show that low spread stocks tend to have also low β, and spreads and market value are inversely related. The portfolios are formed in the same way like in AM86 and ER find that the average returns increase with the proportional spread in January. However, in

non-January months, they do not find such a relation. They point out problems with this portfolio formation technique. Additionally, AM86 assume that the market risk premium is constant over time, ER suggest to use the methodology introduced by Fama and MacBeth (1973).

Using the Fama-MacBeth methodology, they regress portfolio returns on β, the proportional spread and a size variable. For the subperiod from 1961-1990, they find the spread coefficient to be positive, but not significant in all months. In January, the spread coefficient and the β coefficient are significantly positive, whereas size is not priced. However, for non-January months, the spread coefficient is negative, but insignificant. In the subperiod 1981-1990, they find for January that only the spread coefficient is significant and positive, size and β are not significant. Considering non-January months and considering all months, the spread coefficient, i.e., the liquidity premium, is negative but not significant.

ER criticize the survivorship bias in the portfolio selection technique of AM86, which requires 11 years of data for each stock to be included. They modify the selection technique and require only 3 years of data for each stock to be included. They form again 7x7 groups based on spread and β. The resulting portfolios include smaller size firms with larger bid-ask spreads. The number of firms increases dramatically by 45%. They also use unconditional β's in their regression of the average portfolio returns on β, average spread, and size. For the period 1961-1990, the liquidity premium is only present in January, but now, the size effect is present in January months. Looking on all months, the size effect is even the only significant variable, which is clearly in contrast to AM86 results. The β-risk premium is not significant in presence of the size and the spread variables, even not in January. In non-January months, non of the variables is reliably priced. This is consistent with the results of Fama and French (1992). In the subperiod 1981-1990, considering all months, the liquidity premium is negative, but insignificant, the liquidity premium is still negative but now significant in the non-January months. In January, the liquidity premium is positive and significant. Hence, the AM86 model is confirmed for January, but not for non-January months.

ER offer evidence for a strong seasonal component, they find a positive liquidity premium only in January for 1961-1990. The effect of the proportional bid-ask spread on asset pricing is not significant different from zero in non-January month (and even negative). In

contrast to AM86, they find the size effect to be significant after controlling for the effect of the bid-ask spread. The portfolio selection technique employed by AM86 is too restrictive. It systematically excludes smaller firms and introduces a strong survivorship bias. The AM86 results are biased against finding a size effect. ER conclude that the finding that the positive relation between bid-ask spread and average returns appears only in January is consistent with studies that also β is only priced in January. It indicates that the liquidity premium may be a determinant of asset pricing. However, it remains a puzzle why this should only happen in January and not in the other months.

3. Modifications and Extensions of the Amihud and Mendelson (1986) Model

As indicated earlier, the model in Amihud and Mendelson (1986)–AM86–attracted attention and in this section, I present directions for possible extensions or modifications of their model. First, I develop a simple model.

3.1. A Simple Model

At the beginning, it should be pointed out, that the model discussed below–although different in several points–is inspired by the AM86 model and thus similar to it.[29] As pointed out in the discussion of AM86 in Section 2.2.1., their model excludes any uncertainty in asset returns (before transaction costs), the perpetual cash flows as well as the firm value are given constants.

There are M investors in the economy. They are price-taking and risk-neutral. An investor of type $i = 1, 2, ..., M$, with initial wealth W_i buys securities at the posted ask prices at time 0 and holds them for the period (certain) T_i. At time T_i the investor liquidates his position at the posted bid prices. The investment horizon increases with i, i.e., $T_1 \leq T_2 \leq ... \leq T_M$. The economy has $N + 1$ assets, denoted by j. Asset 0 is a zero-spread and riskless asset, whereas assets $1, 2, ..., N$ are risky and have relative spreads S_j, with

[29] In the following, I will not explicitly refer every time when the model contains ideas from the work by Amihud and Mendelson (1986), rather I encourage the reader to compare with their work.

$S_1 \leq S_2 \leq ... \leq S_N < 1$. The riskless asset has unlimited supply and there is one unit of each risky asset available. All assets are perfectly divisible. Trading is achieved through competitive market makers. They quote bid and ask prices for the assets at which they are willing to buy and sell. The market makers are compensated with the (competitive) spreads S_j, which equals the trading costs. Following the notation of AM86, the ask prices are denoted by V_j and the bid prices are given by $V_j(1-S_j)$. Note that this means that V_j is not equal to the value of the asset, it is rather the value of the asset plus the cost of buying.

The assets do not pay dividends, all return is in form of capital gains. The value of the riskless asset follows the following process:

$$dV_0(t) = \mu_0 V_0(t)dt, \tag{3.1}$$

where μ_0 is the drift rate and represents the continuously compounded instantaneously rate of return on the riskless asset 0. $V_0(t)$ denotes the value of the riskless asset at time t. Since asset 0 has a spread of zero, this is also equal to the bid respectively the ask price for asset 0. Equation 3.1 implies that

$$V_0(\tau) = V_0(0)e^{\mu_0 \tau}.$$

For the risky assets which have a positive proportional spread the process for the ask prices of asset j follows a geometric Brownian motion:[30]

$$dV_j(t) = \mu_j V_j(t)dt + \sigma_j V_j(t)dz(t), \tag{3.2}$$

where $V_j(t)$ denotes the ask price of asset j at time t, $dz(t)$ is a standard Brownian motion, μ_j is the drift rate and σ_j is the variance of the diffusion process. Equation 3.2 implies that the ask price of asset j is lognormally distributed with the time 0 expectation for the time τ value to be:

$$E_0[V_j(\tau)] = V_j(0)e^{\mu_j \tau}, \tag{3.3}$$

where E_0 denoted the expectation formed at time 0. Since I assume risk-neutral investors, only the expectation matters and the investor will discount this expectation with the riskfree

[30]Note again, since I follow AM86's notation, V_j is not equal to the underlying asset value. Of course, it would be possible (and maybe more realistic) to model the value and then add and subtract half of the spread to get ask respectively bid prices. However, the conclusions would not change.

rate μ_0. That is any investor tries to maximize the expected present value of his portfolio. The expected payoff at time T_i on the portfolio of an investor of type i (discounted at the riskfree rate of return) is

$$e^{-\mu_0 T_i} E_0 \left(\sum_{j=0}^{N} x_{ij} V_j(T_i)(1 - S_j) \right), \tag{3.4}$$

where x_{ij} denotes the holding of investor i in asset j, with equation 3.3 this yields

$$e^{-\mu_0 T_i} \sum_{j=0}^{N} \left[x_{ij} V_j(0) e^{\mu_j T_i} (1 - S_j) \right]. \tag{3.5}$$

Thus, the optimization problem for a type-i investor facing given vectors is:

$$\begin{aligned} \max_{x_{ij}} \quad & \sum_{j=0}^{N} \left[x_{ij} V_j(0) e^{\mu_j T_i} (1 - S_j) \right], \\ s.t. \quad & \sum_{j=0}^{N} x_{ij} V_j(0) \leq W_i, \\ & x_{ij} \geq 0, \qquad \forall j = 0, 1, 2, ..., N. \end{aligned} \tag{3.6}$$

The conditions are the budget constraint and they prohibit short positions. To achieve an equilibrium allocation, all investors have to solve their respective optimization problem and the market has to clear, that is:

$$\sum_{i=0}^{M} x_{ij} = 1 \qquad \forall j = 1, 2, ..., N. \tag{3.7}$$

I refer to AM86 for the proof that there exists an equilibrium for this problem. In the following, I focus on the optimization problem of the investors. Consider the objective function in equation 3.6. At the end investors do not care about gross returns, that is returns before transaction costs. They take transaction costs into account. Analogous to AM86, I can define the expected return per unit of time net of transaction costs on asset j for an investor i as:

$$r_{ij} = \mu_j + \frac{\ln(1 - S_j)}{T_i}. \tag{3.8}$$

Note, the second term is negative and thus the drift term of the diffusion process is adjusted downwards for the spread. The drift rate can be viewed as the expected gross return before transaction costs. It is obvious that this adjusted return depends on the characteristics of

the asset as well as of the investor. It is clear, since the transaction costs amortize over T_i, that for a given asset j, the adjusted return is non-decreasing in i.

The investors are price-takers and thus for given ask prices, an investor i will choose the assets j which provide him the highest expected returns net of transaction costs, that is:

$$r_i^* = \max_{j=0,1,2,...,N} r_{ij}. \tag{3.9}$$

with $r_1^* \leq r_2^* \leq ... \leq r_M^*$ since as mentioned, r_{ij} is non-decreasing in i for given j. See also AM86, this implies that the adjusted return on a portfolio increases with the investment horizon. Investors with longer holding periods earn higher (expected) returns after transaction costs. Hence, the investor i demands a gross return of $r_i^* - \frac{\ln(1-S_j)}{T_i}$ on asset j. That is the incurred transaction costs are added back. Note, the gross returns are in the end the returns observed in the market. The equilibrium gross return is determined by the market and depends on the highest valued use of asset j, that is by the investor who is willing to buy for the highest price and thus minimal required return:[31]

$$\mu_j^* = \min_{i=1,2,...,M} \left\{ r_i^* - \frac{\ln(1-S_j)}{T_i} \right\}. \tag{3.10}$$

Now, the logic works similar to AM86 and the model yields the same conclusions. The following propositions can also be found in AM86 (besides adjustments due to the different model). Note, $\mu_j^* - \mu_0^*$ is the premium on asset j required by the market as compensation for the transaction costs.

Proposition 1: –Clientele Effect– *Assets with higher spreads are allocated in equilibrium to portfolios with (the same or) longer holding periods.*

Proof. Take two assets, in equilibrium asset j is in portfolio i and asset k is in portfolio $i+1$. With 3.9 it follows that $r_{ij} \geq r_{ik}$ and $r_{i+1,k} \geq r_{i+1,j}$. Using 3.8 and 3.10, $\mu_j^* + \frac{\ln(1-S_j)}{T_i} \geq \mu_k^* + \frac{\ln(1-S_k)}{T_i}$ and $\mu_k^* + \frac{\ln(1-S_k)}{T_{i+1}} \geq \mu_j^* + \frac{\ln(1-S_j)}{T_{i+1}}$. That is $\left(\frac{1}{T_i} - \frac{1}{T_i}\right)(\ln(1-S_j) - \ln(1-S_k)) \geq 0$.

[31] Note, that the equilibrium gross return μ_j^* is of course determined by the prices at time 0. However, since I do abstract from the actual valuation process, this step is not easily made without further assumptions. This is a weakness of this model which uses diffusion processes to model the asset prices. Possible assumptions would be, to define a point in time when the firm liquidates for an uncertain payoff, or to define an uncertain perpetual payment stream. However, I will not investigate this possibilities any further.

It is immediately clear, that if $T_{i+1} > T_i$ then it is valid that $S_k \geq S_j$. This result extends to non-consecutive portfolios.

Proposition 2: –Spread-Return Relation– *In equilibrium, the observed market (gross) return is an increasing and concave piecewise-linear function of the relative spread.*

Proof. Define $f_i(S) = r_i^* - \frac{\ln(1-S)}{T_i}$. From 3.10 it is clear that the market return on an asset with proportional spread S is given by $\min_{i=1,2,..,M} f_i(S)$. Now, all f_i are linear, monotonically increasing and concave in S. The minimum operator preserves monotonicity and concavity and the minimum of a finite collection of linear functions is piecewise-linear.

As mentioned in AM86 and in Section 2.2.1., the increasing relation between spread and return is due to the higher compensation for the incurred transaction costs required by investors. The concavity of the relation is an effect of the clientele effect, since investors with longer holding periods tend to hold the assets with higher spread, the negative effect of the spread gets mitigated, since the transaction costs can amortize over a longer period. The fact that the relation is piecewise linear is an outcome of the portfolio selection by the investors. Usually, an individual investor does not hold all assets in this model, she only invests in some assets and requires the same return from all of these assets. See also AM86, especially Figure 1, page 230.

It is possible to solve an exemplary economy–under imposition of more assumptions–to illustrate the relation. However, I want to refer to AM86 for such an example. In the next sections I present further comments and empirical implications on this simple model and AM86.

3.2. Comments and Thoughts

One problem with the presented model is as already mentioned that the prices at time 0 do not explicitly appear in equation3.10. To convert the required drift rate into prices, we had to specify the payment stream of the firm somehow. Either we could assume that the firm will be liquidated at a specific time in the future and the liquidation value has certain expected value or a perpetual payment like in AM86 has to be introduced. However, AM86 have to fix this payment stream at a specific value to solve their numerical example and

determine the time 0 prices. In my model, it would be possible to assume all prices at time 0 to be equal to one and then to determine the required rate of return on the assets. This is of course not even close to reality and imposes the problem how to think about a steady state.

In order for a solution of the presented model to represent an equilibrium,a steady-state, I have to assume that at any moment in time the same situation with the same number of investors can be found in the market. Otherwise, the solution of a specific model is only a solution for one moment in time. This implies that if an investor liquidates and leaves the market, a new investor of the same type enters at the same moment. The problem that the prices will have changed remains.[32] AM86 solve this problem by introducing a stochastic arrival of investors. The investors arrive according to a Poisson process. This feature could easily be introduced in the model presented in this paper and the market clearing condition would change accordingly. However, it would not change the implications of the model.

In the presented model, the investors know the length of their holding period with certainty. AM86 introduce a random variable with exponential distribution for this. However, at the end this uncertainty does not explicitly enter the solution, since investors in their model maximize the expected discounted liquidation revenues. This implies that the introduction of the uncertainty about the investment horizon is not necessary to generate the result. The same is of course true for the return uncertainty on the risky assets in my model, since the investors are assumed to be risk-neutral. Actually, as long as we assume independence between holding period length and return, we could have introduced the uncertain holding period in the model.

However, it would be interesting to investigate such a model with return uncertainty and/or uncertainty in investment horizon in a framework with risk-averse investors or a single risk-averse representative investor. This would bring back the traditional portfolio theory where benefits from diversification have to be traded-off against transaction costs. This leads to work presented in Section 1.1. Unfortunately, such problems have not been solved in a very general set-up with more than two assets. As mentioned in Section 2.2.1. the restriction on risk-neutral investors leads to the fact that investors will not hold diversified

[32]Note, that introducing a liquidation point for the firm cannot solve the problem, since the time period until liquidation will have changed. Thus, again, the allocation found for a specific moment does not have to be an equilibrium for another point in time (a steady-state).

portfolios and only "specialize" in some few assets. Such behavior seems not rational and is not observed in the market, where investors tend to hold diversified portfolios.

Related to the last comment is that the model does not allow for intertemporal portfolio adjustments. Even in the present set-up with risk-neutral investors, it could be possible that an investor could yield higher returns on his portfolio when he is allowed to rebalance his portfolio (especially if prices move in certain directions and he has to reformulate his expectations about future asset values). Such a setup would also cover the case where arbitrageurs could come in and try to replicate a position for a long holding period with roll-over positions with shorter holding periods. Neither AM86 nor the presented model does allow for such behavior. In both cases, the individual investor behaves like in a single-period model.

Another possible extension of the model is to introduce fix transaction costs. Again, such costs would be especially interesting in an intertemporal framework, where rebalancing is allowed and where investors may be risk-averse. Furthermore, it could be interesting to model the market making process more explicitly and endogenize the spread into the model. AM86 suggest that this could also include that firms try to change their financial policy to affect the spread. Moreover, the introduction of random cash needs for investors, which could motivate why investors want to trade–in contrast to equilibrium models in which everybody is "happy" and trading is not motivated–may provide insights.

3.3. Empirical Issues

AM86 only test their second proposition, since the first proposition to test a clientele effect is hard to test. There are some implications–besides the positive relationship between return and spread–which could be tested. The clientele effect, that is that assets with higher spreads tend to be held by investors with longer holding periods. Also, securities which are held by investors with longer holding periods tend to have larger expected gross returns. Furthermore, the returns investors with longer holding periods earn net of transaction costs are higher (see equation 3.9). However, tests for those kinds of implications require data about the holding periods of investors (and their transaction costs). Day et al. (1985)–DSW–offer three different ways to estimate average length of investors' holding periods, relating number of shares outstanding and number of shares traded in several ways. These measures

are not completely satisfactory as pointed out by DSW. To test the clientele effect it would be also better to have more detailed information than the average holding period, since such estimates can be biased due to estimation errors as well as since some few investors or interdealer trading may drive the results. Such more detailed data may be received from market makers, dealers, and brokers, but this would carry substantial costs and the confidentiality may be a problem. Of course, from such sources one could also receive information about other costs besides the bid-ask spread, like commissions and brokerage fees.

AM86 and many others use the proportional spread as a measure for illiquidity. Another measure is offered by Brennan and Subrahmanyam (1994), namely the inverse of the market depth, λ. Moreover, as pointed out earlier, AM86 adjust for risk, using the traditional "β-approach" in a Fama and MacBeth (1973) framework. In more recent work authors tend to use the Fama and French (1992, 1993) factors to adjust for risk. Although, the discussion of the economical justification for the Fama-French factors is still ongoing, it should be considered if the three Fama-French factors may be the more appropriate way to adjust for risk. Another issue is that now there is monthly spread data available, thus, the analysis of a relation between monthly returns and spreads should utilize this data.

4. Concluding Remarks

This paper gives an extensive review over existing literature in the area of market microstructure and its effect on asset pricing. The first group of papers offer different theoretical models related to the question on hand. Surprisingly, the authors reach different results. For example, Constantinides (1986) states that transaction costs have only second-order effects on asset pricing whereas Amihud and Mendelson, and Yu (1992), or AM86 claim that transaction costs have a significant effect on asset returns. The results are as different as the models. Some papers employ single-period frameworks (e.g., AM86 or Chen, Kim, and Kon (1975)), whereas others focus on intertemporal portfolio selection problems in a continuous framework (Constantinides (1986), or Amihud, Mendelson, and Yu (1992)).

However, the empirical evidence offered in several papers in Section 2.2. and Section 2.3. seems to support the opinion that transaction costs (the spread) have a significant effect

on asset pricing. Although, Eleswarapu and Reinganum (1993)–ER–shed some doubt on the analysis in AM86, in that they are able to show a severe survivorship bias in the portfolio selection of AM86. After mitigating this bias, they find that there is still a size effect present even when adjusting for differences in spreads. That implies previous studies in AM86 or Stoll and Whaley (1983) which state that the size effect can be explained by transaction costs may be flawed. In addition, the spread effect is only significant in January months. The seasonality in returns remains a puzzle and ER are not able to explain their findings. Consistent with ER, Brennan and Subrahmanyam (1994) find that their measure of trading costs, C, is only significant in January months and they cannot explain this finding with seasonalities in the illiquidity of assets, there is no January seasonality.

This discussion indicates that the question of the seasonality in returns as well as in the transaction costs effects remains an open question which deserves further consideration. Another interesting topic could be the development of an empirical test for the clientele effect documented in the AM86 model.

References

Amihud, Y., and H. Mendelson, 1986. Asset Pricing and the Bid-Ask Spread. *Journal of Financial Economics* **17**, 223–249.

Amihud, Y., and H. Mendelson, 1989. The Effect of Beta, Bid-Ask Spread, Residual Risk, and Size on Stock Returns. *Journal of Finance* **44**(2), 479–486.

Amihud, Y., H. Mendelson, and G. Yu, 1992. Income Uncertainty and Transaction Costs. Working paper, New York University. 31 pages.

Banz, R. W., 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* **9**, 3–18.

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

Black, F., and M. Scholes, 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* **1**, 1–22.

Brennan, M. J., and A. Subrahmanyam, 1994. Market Microstructure and Asset Pricing: On the Compensation for Adverse Selection in Stock Returns. Working paper, University of California at Los Angeles. 22 pages plus tables.

Chen, A. H., E. H. Kim, and S. J. Kon, 1975. Cash Demand, Liquidation Costs and Capital Market Equilibrium under Uncertainty. *Journal of Financial Economics* **2**, 293–308.

Constantinides, G. M., 1976. Comment on Chen, Kim and Kon. *Journal of Financial Economics* **3**, 295–296.

Constantinides, G. M., 1986. Capital Market Equilibrium with Transaction Costs. *Journal of Political Economy* **94**(4), 842–862.

Day, T. E., H. R. Stoll, and R. E. Whaley, 1985. *Taxes, Financial Policy, and Small Business*, pp. 103–154. Lexington Books, D.C. Heath and Company/Lexington, Massachusetts/Toronto.

Dimson, E., 1979. Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics* **7**, 197–226.

Easley, D., and M. O'Hara, 1987. Price, Trade Size, and Information in Securities Markets. *Journal of Financial Economics* **19**(1), 69–90.

Eleswarapu, V. R., and M. R. Reinganum, 1993. The seasonal behavior of the liquidity premium in asset pricing. *Journal of Financial Economics* **34**, 373–386.

Fama, E. F., and K. R. French, 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance* **47**(2), 427–465.

Fama, E. F., and K. R. French, 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3–65.

Fama, E. F., and J. MacBeth, 1973. Risk, return and equilibrium: Empiricla tests. *Journal of Political Economy* **81**, 607–636.

Foster, F., and S. Viswanathan, 1993. Variations in trading volume, return volatility, and trading costs: Evidence on recent price formation models. *Journal of Finance* **48**, 187–211.

Fowler, D. J., and C. H. Rorke, 1983. Risk Measurement When Shares Are Subject to Infrequent Trading: Comment. *Journal of Financial Economics* **12**, 279–283.

Gale, D., 1960. *The theory of linear economic models*. McGraw-Hill, New York.

George, T., G. Kaul, and M. Nimalendran, 1991. Estimating the compnents of the bid-ask spread: A new approach. *Review of Financial Studies* **4**, 623–656.

Glosten, L. R., and L. Harris, 1988. Estimating the components of the bid-ask spread. *Journal of Financial Economics* **21**, 123–142.

Glosten, L. R., and P. R. Milgrom, 1985. Bid, Ask and Transaction Prices in a Specialist Market with Heterogenously Informed Investors. *Journal of Financial Economics* **14**(1), 71–100.

Hasbrouck, J., 1991. Measuring the information content of stock trades. *Journal of Finance* **46**, 179–207.

James, K. R., and S. Woodward, 1995. Insider Trading and the Cost of Capital. Working paper, Office of Economic Analysis, US Securities and Exchange Commission. 31 pages plus tables.

Judge, G. G., et al., 1980. *The theory and practices of econometrics*. Wiley, New York.

Kmenta, J., 1971. *Elements of econometrics*. Macmillan, New York.

Kyle, A. S., 1985. Continuous Auction and Insider Trading. *Econometrica* **53**(6), 1313–1335.

Maddala, G. S., 1977. *Econometrics*. MacGraw-Hill, New York.

Madhavan, A., and S. Smidt, 1991. A Bayesian model of intraday specialist pricing. *Journal of Financial Economics* **30**, 99–134.

Magill, M. J. P., and G. M. Constantinides, 1976. Portfolio Selection with Transaction Costs. *Journal of Economic Theory* **13**, 245–263.

Merton, R., 1969. Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economic Statistics* **60**, 247–257.

Merton, R. C., 1971. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* **3**, 373–413.

Merton, R. C., 1987. A Simple Model of Capital Market Equilibrium with Incomplete Information. *Journal of Finance* **42**, 483–510.

Reinganum, M. R., 1981. Misspecification of capital asset pricing: Empirical anomalies based on earning yields and market values. *Journal of Financial Economics* **9**, 19–46.

Reinganum, M. R., 1990. Market microstructure and asset pricing. *Journal of Financial Economics* **28**, 127–147.

Roll, R., 1981. A possible explanation for the small firm effect. *Journal of Finance* **36**, 879–888.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* **39**, 1127–1139.

Schultz, P., 1983. Transaction Costs and the Small Firm Effect: A comment. *Journal of Financial Economics* **12**, 81–88.

Stoll, H. R., 1989. Inferring the components fo the bid-ask spread: Theory and empirical test. *Journal of Finance* **44**, 115–134.

Stoll, H. R., and R. E. Whaley, 1983. Transaction Costs and the Small Firm Effect. *Journal of Financial Economics* **12**, 57–79.

E L I N K E I N O E L Ä M Ä N T U T K I M U S L A I T O S (ETLA)
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
LÖNNROTINKATU 4 B, FIN-00120 HELSINKI

Puh./Tel. (09) 609 900 Telefax (09) 601753
Int. 358-9-609 900 Int. 358-9-601 753
http://www.etla.fi

KESKUSTELUAIHEITA - DISCUSSION PAPERS ISSN 0781-6847

No 664 GRIGORI DUDAREV - MICHAEL ZVEREV, Energy Sector in Russia. Economic and Business Outlook. 15.01.1999. 49 p.

No 665 JYRKI ALI-YRKKÖ - PEKKA YLÄ-ANTTILA, Omistus kansainvälistyy - johtamis- ja valvontajärjestelmät muuttuvat. 29.01.1999. 32 s.

No 666 MIKKO MÄKINEN - MIKA PAJARINEN - SIRKKU KIVISAARI - SAMI KORTELAINEN, Hyvinvointiklusterin vientimenestys ja teollinen toiminta 1990-luvulla. 08.02.1999. 67 s.

No 667 OLAVI RANTALA, Tuotannon ja työllisyyden alueellisen ennustamisen menetelmät. 19.02.1999. 43. s.

No 668 JARI HYVÄRINEN, Globalisaatio, taloudellinen kasvu ja syvenevä alueellistuminen. 02.03.1999. 68 s.

No 669 JUKKA LASSILA, An Overlapping-Generations Simulation Model for the Lithuanian Economy. 02.03.1999. 21 p.

No 670 JUKKA LASSILA, Pension Policies in Lithuania - A Dynamic General Equilibrium Analysis. 02.03.1999. 44 p.

No 671 HENRI PARKKINEN, Black-Scholes-malli ja regressiopohjainen lähestymistapa stokastisen volatiliteetin estimointiin - Katsaus suomalaisten FOX-indeksioptioiden hinnoitteluun. 15.03.1999. 88 s.

No 672 JUHA SORJONEN, An Econometric Investigation between Volatility and Trading Volume of the Helsinki and New York Exchanges: A Firm Level Approach. 26.03.1999. 99 p.

No 673 ANTTON LOUNASHEIMO, The Impact of Human Capital on Economic Growth. 30.03.1999. 35 p.

No 674 PASI SORJONEN, Ex-Dividend Day Behaviour of Stock Prices in Finland in 1989-90 and 1993-97. 30.03.1999. 29 p.

No 675 PASI SORJONEN, Ex-Dividend Day Stock Returns and Tick Rules. 30.03.1999. 21 p.

No 676 PASI SORJONEN, Ex-Dividend Day Stock Price Behaviour, Taxes and Discrere Prices; A Simulation Experiment. 30.03.1999. 28 p.

No 677 JUHA HONKATUKIA, Kioton mekanismien käytön rajoittamisen vaikutukset Suomeen. 08.04.1999. 41 s.

No 678 ANSSI PARTANEN - INKERI HIRVENSALO, North and Westbound Foreign Trade Potential of the Baltic Rim. 28.04.1999. 17 p.

No 679 GRIGORI DUDAREV, The Role of Technology in Shaping the Energy Future in Russia. 06.05.1999. 48 p.

No 680 REIJA LILJA - EIJA SAVAJA, En översikt av systemet för arbetslöshetsskydd i Finland. 06.05.1999. 21 s.

No 681 REIJA LILJA - EIJA SAVAJA, Olika sätt att söka arbete, attityder och motivation hos arbetssökande i Finland. 06.05.1999. 73 s.

No 682 JARMO ERONEN, Cluster Analysis and Russian Forest Industry Complex. 24.06.1999. 16 p.

No 683 SEPPO HONKAPOHJA - ERKKI KOSKELA, The Economic Crisis of the 1990s in Finland. 09.08.1999. 53 p.

No 684 STEPHEN KING - ROHAN PITCHFORD, Private or Public? A Taxonomy of Optimal Ownership and Management Regimes. 12.08.1999. 33 p.

No 685 HANNU HERNESNIEMI - MIKKO HONGISTO - LASSI LINNANEN - TORSTI LOIKKANEN - PÄIVI LUOMA, Kioto-sopimus ja yritykset. Esitutkimus strategioista. 07.09.1999. 68 s.

No 686 PETRI ROUVINEN, R&D Spillovers among Finnish Manufacturing Firms: A Cost Function Estimation with Random Coefficients. 08.09.1999. 51 p.

No 687 ANNE ERONEN, Classification of Intangibles - Some Comments. 04.10.1999. 13 p.

No 688 HANNU PIEKKOLA, Rent Sharing and Efficiency Wages. 06.10.1999. 25 p.

No 689 MIKA PAJARINEN, Foreign Firms and Their R&D in Finland. 11.10.1999. 33 p.

No 690 PETRI ROUVINEN, Characteristics of Product and Process Innovators among Finnish Manufacturing Firms. 11.10.1999. 29 p.

No 691 HANS GERHARD HEIDLE, Market Microstructure and Asset Pricing: A Survey. 25.10.1999. 57 pages.

Elinkeinoelämän Tutkimuslaitoksen julkaisemat "Keskusteluaiheet" ovat raportteja alustavista tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista. Tässä sarjassa julkaistuja monisteita on mahdollista ostaa Taloustieto Oy:stä kopiointi- ja toimituskuluja vastaavaan hintaan.
Papers in this series are reports on preliminary research results and on studies in progress. They are sold by Taloustieto Oy for a nominal fee covering copying and postage costs.

d:\ratapalo\DP-julk.sam/25.10.1999