EMPIRICAL RESEARCH ON MARKET EFFICIENCY WITH RESPECT TO ACCOUNTING INFORMATION: A CRITICAL SURVEY OF THE LITERATURE

1. INTRODUCTION

The efficient markets hypothesis (henceforth EMH), which in its most stringent form claims that security prices always fully reflect all available information, has been a dominant paradigm in finance for more than three decades. During that period, literally thousands of articles have been devoted to empirical examinations of the validity of the EMH[1]. A subset of information that has thereby received particularly much attention are accounting data. This is probably partly due to the fact that financial statements constitute a source of information with respect to which a remarkable amount of evidence has been documented that seems to be anomalous to the EMH. Although the validity of some of these anomalies is definitely questionable, they do suggest, as a whole, that the stock market may have a problem with processing accounting information efficiently. In my opinion, this property of financial statement data also accounts at least partly for the central place that they have taken in the ongoing debate between EMH proponents and opponents.

Despite its magnitude, the impressive body of evidence that appears to be inconsistent with the predictions of the theory of efficient markets has not even come near to convincing the academic community that the EMH ought to be substituted by an alternative paradigm. Apart from the EMH's appealing theoretical foundations and its tradition-based status, the reluctance to let go of it as the dominant paradigm is in my opinion also due to the doubtful validity of many of the allegedly anomalous results.

---

[1] For an overview of the vast empirical EMH literature, see Van Uytbergen (2002).

Validity aspects should obviously be of concern to anyone who embarks upon scientific research. Yet the need to see to the validity of one's findings is probably more urgent in the area of market efficiency testing than in other circumstances. After all, the concept of validity basically refers to the best available approximation to the truth (falsity) of propositions. Consequently, the main overall threat to the validity of inferences arises from the fact that obtained results may be explainable in terms of competing propositions (the maintained proposition). Since the infamous joint-hypothesis problem makes that that duality of explainability is nearly inherent in empirical studies of the EMH, the latter are clearly very prone to that threat. In order to be able to make a significant contribution to the extant literature in this field, it is therefore crucial to devote considerable attention to the set-up of a research design that takes the susceptibility of market efficiency tests to validity threats into account.

It is herein that the motivation for this research paper takes root. No one benefits from more "empirical evidence" that merely adds to the controversy as to whether stock markets are efficient with respect to accounting information or not. Clear-cut and unambiguous research results are undoubtedly called for. Stated otherwise, the development of a research design that is likely to produce as valid results as possible is absolutely critical in this context. The aim of this research paper is to infer the features of such a research design by critically reviewing the methodological strengths and weaknesses of the existing empirical literature on market efficiency with respect to accounting information.

To structure my critical discussion of the literature on cross-sectional accounting-based anomalies, I rely on the framework developed in the classical work by Cook and Campbell (1979). They have introduced a subdivision of the general concept of validity into four types of validity: **statistical conclusion validity**, **internal validity**, **construct validity** and **external validity**. On top of that, Cook and Campbell have drawn up lists of the main threats to each of these validity types. I shall discuss the more relevant of those threats and relate them to the research designs that have been used to date in accounting-based tests of market efficiency. As such, I intend to give a considerable number of detailed methodological suggestions for future research in this area.

The structure of this paper is as follows: in section two, I give a brief overview of the main cross-sectional anomalies with respect to accounting information that have been revealed to date. Next, section three contains an introduction to the terminology used by Cook and Campbell (1979) as well as a description of their overall validity framework. In sections four through seven, I discuss the major threats to the four dimensions of validity identified by Cook and Campbell. The paper ends with some concluding comments.

2. THE ACCOUNTING-BASED ANOMALIES

As far as accounting-based anomalies are concerned, I suggest that an additional distinction be made. Specifically, I shall distinguish between **direct** and **indirect accounting-based anomalies**. The former are the ones that directly relate currently observable accounting(-based) information to subsequent stock returns, while the latter make use of an intermediary variable or event. To some, this distinction may seem somewhat artificial, but I introduce it here because I shall to some extent rely on it to critically review the literature on accounting-based EMH tests.

2.1. DIRECT ACCOUNTING-BASED ANOMALIES

The value-versus-growth anomaly

The major direct accounting-based anomaly is undoubtedly the so-called **value-versus-growth** (or **glamour**) **anomaly**. The term value stocks is used to refer to companies trading at relatively low multiples of certain accounting variables, while growth stocks are the ones with high prices relative to accounting measures of value. There is overwhelming evidence that the former type of stocks systematically outperforms the latter on a risk-adjusted basis. Although there is no theoretical reason to expect this to be true, Graham and Dodd (1934) already argued that value strategies are profitable. Over the years, several variables have been used to distinguish between value and growth stocks. The most important ones are the **price-to-earnings ratio** (henceforth P/E ratio; e.g. Nicholson (1960), Basu (1977),

Reinganum (1981) and Jaffe, Keim and Westerfield (1989)), the **price-to-book ratio** (henceforth P/B ratio; e.g. Rosenberg, Reid and Lanstein (1985), De Bondt and Thaler (1987), Keim (1988) and Fama and French (1992)), the **price-to-cash-flow ratio** (henceforth P/CF ratio; e.g. Chan, Hamao and Lakonishok (1991)) and the **price-to-sales ratio** (henceforth P/S ratio; e.g. Senchack and Martin (1987) and Jacobs and Levy (1988))[2].

The leverage anomaly

Another direct accounting-based anomaly is the **leverage anomaly** described by Bhandari (1988). Bhandari showed that the debt-to-equity ratio (henceforth, D/E ratio) is positively related to expected stock returns. According to him, the observed relationship could not be fully explained by traditional risk measures. In addition, the results of a number of sensitivity tests led Bhandari to the conclusion that the leverage variable did not just have explanatory power for subsequent returns due to inappropriate risk adjustment. Thus, without providing a fully fledged explanation as to its source, Bhandari implicitly suggested that what he discovered ought to be considered evidence of some market inefficiency.

The Holthausen and Larcker anomaly

The last direct accounting-based anomaly is the **Holthausen and Larcker anomaly**. The article by Holthausen and Larcker (1992) was primarily written in reaction to earlier evidence documented in Ou and Penman (1989a)[3]. Holthausen and Larcker have presented empirical evidence that questions the robustness of Ou and Penman's results and suggests that the profitability of the latter's trading strategy may have been sample-specific. In response, Holthausen and Larcker have developed an investment strategy of their own that is based on a research design largely identical to Ou and Penman's. Holthausen and Larcker too, start out with a large set of financial descriptors to which (stepwise) LOGIT analysis is applied in order to select the accounting-based variables with forecasting abilities. However, the two studies differ fundamentally with respect to the nature of their dependent

---

[2] It should be noted that many studies have not used the variables as defined here, but rather their inverses. In this paper, I use the variables and their inverses interchangeably. In view of the qualitative similarity, this should not pose any problems, although it is obviously important that one be wary of the change in sign with respect to the relation with subsequent returns.

[3] See section 2.2.

variables: whereas Ou and Penman constructed a measure to predict the sign of one-year-ahead earnings changes, Holthausen and Larcker directly focus on the sign of future abnormal returns. The investment strategy proposed by Holthausen and Larcker then consists of going long (short) in the stocks with the highest (lowest) estimated probability of earning positive abnormal returns during the subsequent period. Using data for the major American stock markets for the 1978-1988 period, they find that their strategy clearly outperforms Ou and Penman's and earns abnormal returns in a more consistent manner.

2.2. INDIRECT ACCOUNTING-BASED ANOMALIES

The acquisition probability anomaly

As I mentioned in the introduction of section two, indirect accounting-based anomalies differ from the direct anomalies in that they are not based on a direct supposed relationship between the currently observable accounting variable(s) and subsequent abnormal stock returns, but rather on the accounting-based predictability of some event or variable that is alleged to be associated with future abnormal returns. Most of these indirect predictability studies have focused on forecasting earnings. One strand of research forms a notable exception, though, namely the **acquisition probability** literature. One of the most representative studies in this particular area is Wansley, Roenfeldt and Cooley (1983). Their approach was founded on earlier evidence (e.g. Mandelker (1974), Franks, Broyles and Hecht (1977) and Elgers and Clark (1980)) of abnormal returns accruing to shareholders of acquired firms in the period prior to merger. Following Simkowitz and Monroe (1971) and Stevens (1973) who reported reasonable success in predicting mergers from publicly available information, Wansley, Roenfeldt and Cooley have evidenced the possibility to earn abnormal returns by investing in firms with financial profiles indicating a high potential for merger. In spite of the interesting rationale underlying the acquisition probability anomaly, it has received relatively little academic attention. The reason probably lies in the fact that some of the variables included in the merger prediction models show great resemblance to variables upon which other anomalies are based. Many academics have, therefore, considered the acquisition probability anomaly a distorted manifestation of more fundamental anomalies like the size effect and the value-to-growth effect.

Post-earnings announcement drift

The most thoroughly studied anomaly among the indirect accounting-based anomalies relating to the prediction of future earnings is the **post-earnings announcement drift** (henceforth PEAD), surely. PEAD refers to the empirical regularity that companies which announce positive (negative) unexpected earnings tend to earn higher (lower) *subsequent* returns than the market on a risk-adjusted basis. The first formal documentation of a drift in stock returns following earnings announcements is attributed to Ball and Brown (1968). It is probably quite telling and indicative of PEAD being inherent in stock price formation that evidence of this anomaly was already found in the article that is commonly viewed as the starting point of capital markets research in accounting. It is important to note, though, that Ball and Brown's evidence of PEAD pertained to *annual* earnings announcements. This particular finding has rarely if ever been corroborated since. In contrast, the PEAD pattern has proved to be very robust as far as *quarterly* earnings announcements are concerned (e.g. Jones and Litzenberger (1970), Joy, Litzenberger and McEnally (1977), Latané and Jones (1977 and 1979), Watts (1978) and Rendleman, Jones and Latané (1982)).

Some readers may find it surprising and possibly even disturbing that I classify PEAD as an *indirect* accounting-based anomaly. Admittedly, the term was initially conceived to refer to a *direct* relationship between unexpected earnings and subsequent abnormal returns. Early efforts to rationalize PEAD (e.g. Ball (1978) and Foster, Olsen and Shevlin (1984)) mainly focused on possible specification errors and problems in estimating (abnormal) returns, but failed to satisfactorily explain the anomalous pattern. However, during the past 15 years or so, our understanding of the causes of PEAD has improved considerably, changing the perceived nature of the anomaly in the process. The articles by Rendleman, Jones and Latané (1987), Bernard and Thomas (1989 and 1990) and Freeman and Tse (1989) have been seminal in this respect. They have suggested the so-called *naïve expectations hypothesis* as the main cause underlying PEAD. This hypothesis claims that investors, being unaware that firms' seasonally-differenced quarterly earnings are serially correlated, mistakenly use a seasonal random walk model in setting their expectations with respect to future quarterly earnings. As a result,

investors supposedly make or imply inferior forecasts of future earnings changes, making it possible that the market reaction to subsequent earnings announcements be partly predicted. Subsequent research (e.g. Bhushan (1994), Ball and Bartov (1996), Soffer and Lys (1999) and Bartov, Radhakrishnan and Krinsky (2000)) has further refined our understanding of the descriptive validity of the naïve expectations hypothesis. In spite of these extensions and refinements, the essential claim of the naïve expectations hypothesis has basically been confirmed in all instances: investors do not fully use past quarterly earnings information when setting their expectations regarding future quarterly earnings and that is what causes PEAD. In fact, the evidence consistent with the predictions of the naïve expectations hypothesis has been so overwhelming that even the most fervent proponents of the EMH (e.g. Fama (1998)) have found it hard to reject the hypothesis as the most likely explanation for PEAD. That is why I think it is fair to say that our increased understanding of PEAD has turned it from a direct into an indirect accounting-based anomaly that uses future earnings as the intermediary variable(s).

The Ou and Penman anomaly

The **Ou and Penman anomaly** is another indirect accounting-based anomaly. As mentioned earlier, there are quite a lot of similarities between the investment strategy developed in Ou and Penman (1989a) and the Holthausen and Larcker anomaly. Both trading strategies are founded on summary financial statement measures calculated from LOGIT models. The latter are thereby estimated to select on purely statistical grounds the accounting variables with forecasting abilities. Contrary to the major similarities between Ou and Penman's work on the one hand and Holthausen and Larcker's (1992) on the other, the two studies differ fundamentally with respect to the nature of the future event of which the probability is estimated: whereas Holthausen and Larcker focus directly on the sign of abnormal returns, Ou and Penman concentrate on the direction of earnings changes. Concretely, they construct zero-investment portfolios consisting of long positions in companies with high estimated probabilities of experiencing positive one-year-ahead earnings changes and short positions in firms with low estimated probabilities. Ou and Penman report that, after controlling for risk, their zero-investment portfolios earn mean 24-month buy-and-hold returns of about seven

percent. In a follow-up study, Ou and Penman (1989b) have claimed that their summary financial statement measure actually distinguishes transitory components in annual earnings. In combination, these findings have led Ou and Penman to the conclusion that investors recognize the future earnings-related information that is present in financial statements with a lag and hence that stock markets are not efficient with respect to accounting information.

The accrual-based anomaly

Building on the arguments brought forward by some financial analysts that investors insufficiently take the differential implications of the components of earnings into account, Sloan (1996) has unveiled what has become known as the **accrual-based anomaly**. Specifically, Sloan has examined whether stock prices fully reflect the information about future earnings that is contained in the accrual and cash flow components of current earnings. Apparently, investors tend to overestimate the time-series persistence of the accrual component of earnings and to underestimate that of the cash flow component. Moreover, the misweightings seem to be also economically significant, as evidenced by the large mean returns (i.e. about ten percent for a twelve-month holding period) earned by the hedge portfolios constructed by Sloan. In addition, Xie (2001) has recently refined Sloan's findings by demonstrating that investors particularly tend to overprice *discretionary* accruals (i.e. accruals induced by earnings management). Thus, it has been suggested that the strong and consequent profitability of accrual-based trading strategies stems from investors' incapability to "see through" the accounting process underlying the reported figures.

The Abarbanell and Bushee anomaly

The last indirect accounting-based anomaly I discuss is the **Abarbanell and Bushee anomaly**. Abarbanell and Bushee (1998) have examined whether the application of true fundamental analysis can yield significant abnormal returns. Their approach is similar to Ou and Penman's (1989a) in that their trading strategy is also based on multivariate accounting-based prediction of one-year-ahead earnings. However, unlike Ou and Penman, Abarbanell and Bushee's selection of the variables upon which their fundamental strategy is based, is directly motivated by economic

arguments. Abarbanell and Bushee find that their portfolios earn an impressive average twelve-month abnormal return of about 13%.

3. THE COOK AND CAMPBELL FRAMEWORK OF VALIDITY

Cook and Campbell (1979) focus on *causal* relationships. While the everyday meaning of the terms "cause" and "effect" is quite clear, their scientific interpretation is not so straightforward. As a matter of fact, Cook and Campbell dedicate the entire first chapter of their book to giving an overview of how the concept of causality has been defined under some of the major epistemologies. Personally, I am most comfortable with the critical-realist perspective that Cook and Campbell bring forward themselves, and which may be seen as a mishmash of positivist, essentialist, falsificationist, activity and evolutionary influences. Cook and Campbell's position on causation can be outlined as follows: a causal relationship implies that the manipulation of a certain "cause" construct will usually result in the manipulation of an "effect" construct. The term "usually" in this description refers to the probabilistic nature of causal laws, which results from the fact that Cook and Campbell acknowledge that causal assertions are also meaningful at the *molar*[4] level. The ultimate *micromediation* need not be known. The fact that molar causes will not invariably lead to a particular effect is one consequence. Some necessary intermediary conditions may not be satisfied in all instances. Hence the probabilistic nature of (molar) causal assertions. Another important implication of accepting that causal laws may be specified at a molar level is that effects necessarily follow causes in time.

---

[4] Cook and Campbell (1979) define the terms "molar" and "micromediation" as follows (p. 32): *"(...) molar refers to causal laws stated in terms of large and often complex objects. Micromediation refers to the specification of causal connections at a level of smaller particles than make up the molar objects and on a finer time scale."* Cook and Campbell repeatedly refer to Collingwood's (1940) illustration of a light bulb being turned on to distinguish between these different levels of specification. The fact that flicking a particular switch *causes* the bulb to burn is clearly an example of a molar causal law. Stating that the flicking of the switch *causes* a circuit to be closed, which makes it possible for the electric current to flow into the bulb, where it *causes* the filament to heat up, which, in turn, *causes* the light is of course a causal law at a more micromediational level. Whether it is the ultimate micromediation is debatable, though; it might be from an electrician's point of view, but probably not from that of a physicist. This neatly illustrates how there is no such thing as *the* cause of a particular effect. Multiple specifications, dependent upon the context and the observer's background, are virtually always conceivable.

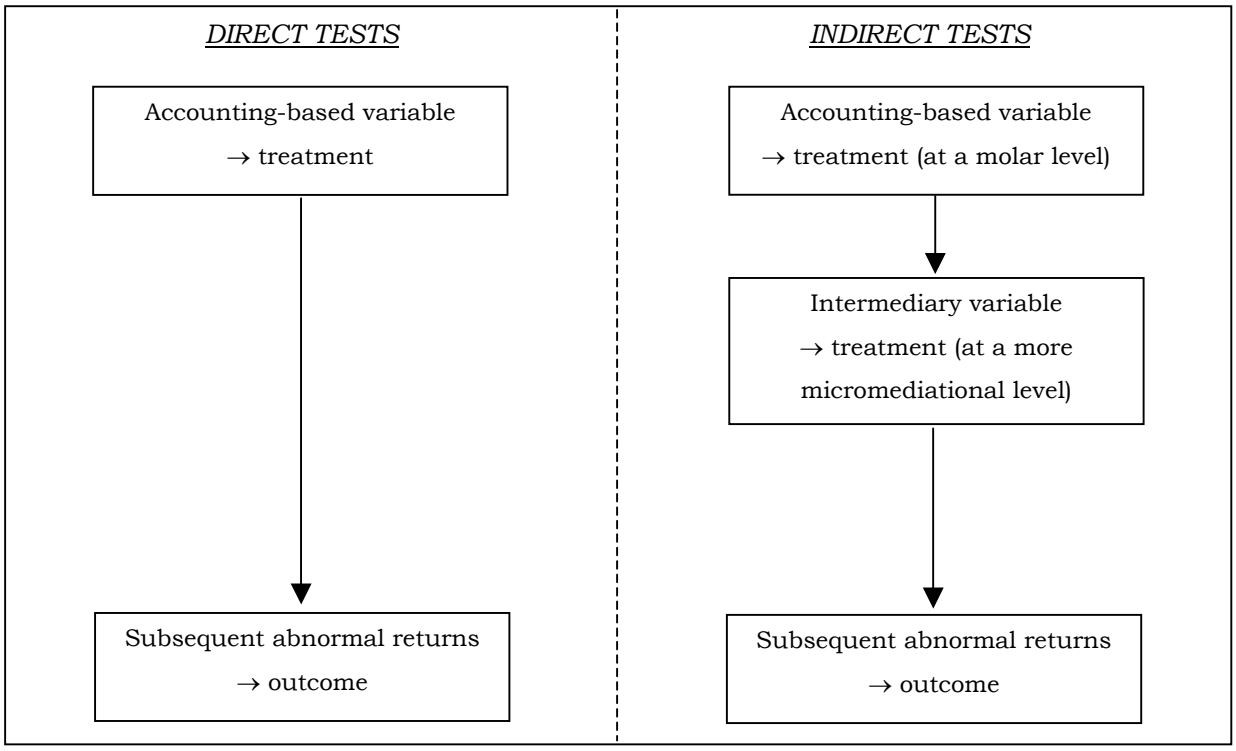
Now how does this translate to the language of research design? First of all, I shall, analogous to Cook and Campbell (1979), in the context of empirical research refer to possible causes as *treatments* or *independent variables*, while possible effects will be termed *outcomes* or *dependent variables*. Following Cook and Campbell's definition of causation described above, this then implies that the existence (absence) of a causal connection may only be inferred from an empirical research set-up if the existence (absence) of a relationship between a putative treatment and a putative outcome can *validly* be established on the basis of that set-up. Thus, I come to the issue of validity. However, before actually looking into Cook and Campbell's validity framework, I briefly discuss how their position on causation is to be interpreted in the context of accounting-based empirical tests of market efficiency.

The definition of the EMH has generally been taken to imply that the information available at a given point in time is fully utilized to determine equilibrium expected returns. This has led to the inference of the implied and testable fair game property of subsequent abnormal returns with respect to the available information set. I find it interesting to highlight how the EMH is quite peculiar in this respect. Most theories tend to predict the *existence* of a relationship between constructs, while the EMH actually entails the prediction that such a relationship *does not exist*. As far as the specific case of accounting information is concerned, the EMH thus pretends that it is impossible to find a systematic relationship between the values of some pre-specified accounting-based variable (i.e. the treatment) and subsequent abnormal stock returns (i.e. the outcome). The alternative hypothesis claims that the information contained in the accounting-based variable is not fully used to determine equilibrium expected returns, *causing* the stocks concerned to earn predictable subsequent abnormal returns.

What I have just described, may be regarded as the very general framework for empirical tests of market efficiency with respect to accounting information. However, keeping in mind the distinction between direct and indirect accounting-based anomalies that I have introduced in section two, I find it important to briefly consider how these two approaches fit into that framework. A graphical illustration is presented in figure 1. As their name indicates, direct tests of market efficiency

with respect to financial statement information test for a direct relationship between an accounting-based variable and subsequent abnormal returns. In the case of indirect tests, that direct relationship is basically broken down into (at least) two parts. Concretely, the accounting-based variable and the subsequent abnormal returns are linked to each other through some intermediary variable. In the Cook and Campbell (1979) terminology, this intermediary variable is in fact a treatment at a more micromediational level. It is readily apparent that whereas the direct tests almost inherently remain silent about the actual sources of any potential inefficiencies, the indirect tests are at least somewhat more detailed as far as the specification of the alternative hypothesis is concerned. I would like to stress that this difference between the two approaches is in my view certainly not purely academic. Why should become clear as the paper progresses.

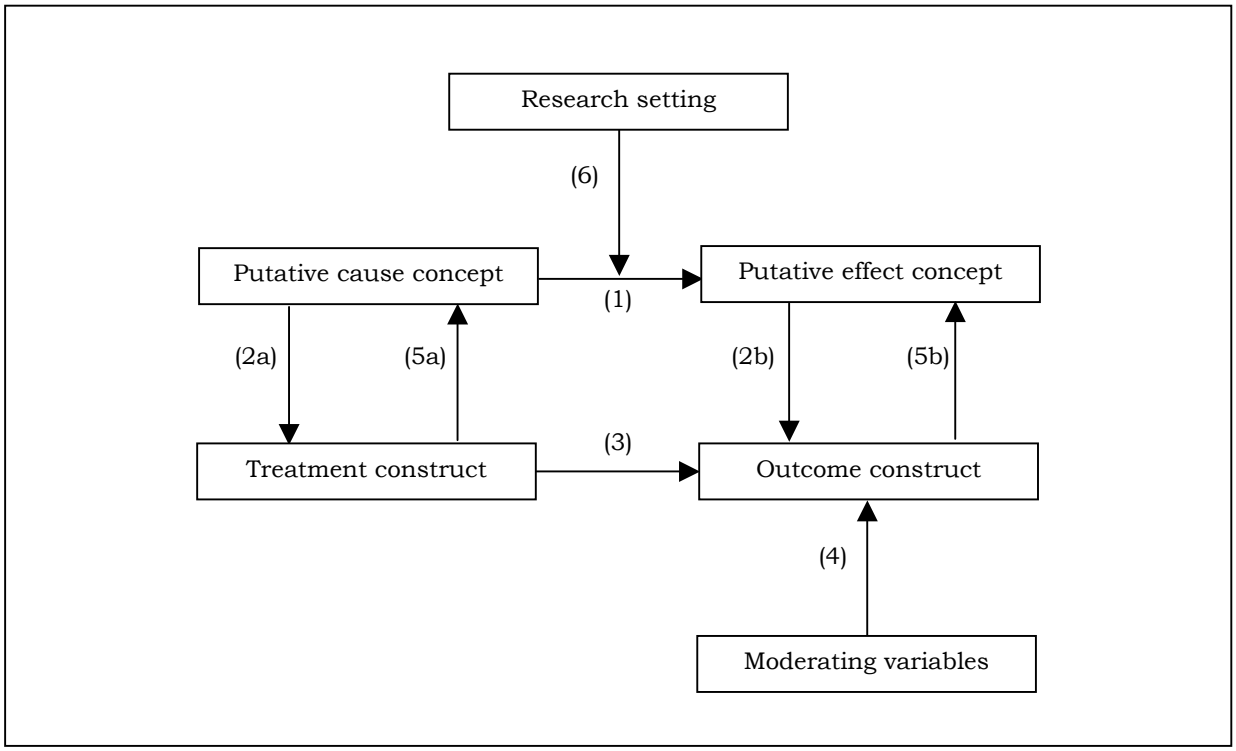
FIGURE 1: THE RELATIONSHIPS IN DIRECT AND INDIRECT TESTS OF MARKET EFFICIENCY WITH RESPECT TO ACCOUNTING INFORMATION



Having given an overview of the terminology employed by Cook and Campbell (1979) and having briefly discussed how their position on causation may be transposed to the context of accounting-based EMH testing, I can now turn to their general validity framework. As an aid in describing the latter, I make use of figure 2, which

is a strongly adapted version of a figure taken from Libby (1976). Link 1 in the figure is the causal relationship that is supposed to be under scrutiny. So the question the research design is supposed to generate a valid answer to is whether link 1 actually does or does not exist. A first problem researchers thereby have to face is that the putative cause and effect are normally merely concepts, i.e. they are not observable. In order to make those concepts operational, it is therefore required to define treatment and outcome constructs that serve as measures for the alleged cause and effect, respectively. This operationalization is represented by links 2a and 2b in figure 2. The aim then is to set up a research design that allows as valid as possible inferences to be made about link 1 on the basis of evidence pertaining to link 3 (the relationship between the treatment and outcome constructs, that is). As mentioned earlier, Cook and Campbell discern four types of validity that have to be taken into consideration to achieve that goal.

FIGURE 2: THE COOK AND CAMPBELL (1979) FRAMEWORK OF RESEARCH DESIGN VALIDITY



Source: Based upon Libby, R. (1976), "Discussion of Cognitive Changes Induced by Accounting Changes: Experimental Evidence on the Functional Fixation Hypothesis", *Journal of Accounting Research*, p. 20.

Normally, the execution of a research project begins with examining statistically whether link 3 exists or not. As such, the researcher should ascertain that the

combination of the statistical testing procedures used and the data at hand allow that meaningful conclusions be drawn in that respect. Cook and Campbell (1979) label this the need for *statistical conclusion validity*. Once it has been validly established that there is (no) statistical covariation between the treatment and outcome constructs, it is crucial to investigate what the obtained result originates from: is it a reflection of the actual relationship between the treatment and the outcome, or is it driven by the impact of moderating variables on the outcome construct (i.e. link 4)? *Internal validity* is concerned with the extent to which the influence of such moderating variables is adequately controlled for. Having determined the nature of link 3, the researcher's next step consists of using those findings to draw inferences about link 1. Obviously, a necessary condition thereby is that the treatment and outcome constructs can be considered sufficiently accurate representations of the theoretical cause and effect concepts, respectively. This condition is termed *construct validity* and is denoted by links 5a and 5b in figure 2. Basically, construct validity is about whether the results obtained for the operational constructs used can be generalized to the higher-order theoretical concepts of interest. The fourth and last type of validity, i.e. *external validity*, deals with overall generalizability. Concretely, it refers to the extent to which the conclusions drawn about the causal relationship under investigation (i.e. link 1) are specific to the particular research setting (i.e. link 6).

For each of the four types of validity, I shall now discuss the threats identified by Cook and Campbell (1979) that are in my opinion most pertinently present in the context of accounting-based testing of the EMH. As mentioned before, it is my intention to deduce the features of a proper research design by critically examining how the existing literature has dealt with those threats or has failed to do so.

4. STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity has to do with the extent to which a particular research design enables one to draw meaningful conclusions with respect to the existence or the absence of covariation between the treatment and outcome constructs used. A study's statistical conclusion validity is obviously very closely

connected with the specific quantitative procedures that are employed in it. Given that many different procedures have been used over the years, it is far from evident to come up with a more or less general evaluation of the performance of the accounting-based market efficiency literature as far as this type of validity is concerned. Nevertheless, I shall try to do so.

Quite helpful in this respect is the fact that the various ways that have been developed to conduct stock selection tests may all ultimately be regarded as adaptations of two possible basic approaches to the issue. The first potential approach consists of implementing *a sorting algorithm*. At a certain point in time, firms are hereby ranked on the basis of the treatment variable of interest and grouped into portfolios (in most cases deciles or quintiles). Next, the subsequent performance of the (extreme) portfolios is examined. Another way to proceed involves the estimation of *a cross-sectional regression*, whereby the explanatory power of the treatment construct for future abnormal returns is examined.

4.1. LOW RELIABILITY OF MEASURES AND LOW STATISTICAL POWER

A question that might arise naturally is why one would want to resort to a sorting procedure anyway. After all, sorting basically transforms the values of the independent variable into ranks. As such, not all the information contained in the treatment variable is then used in the analysis. Yet most stock selection tests have relied on some kind of sorting algorithm to draw inferences regarding market efficiency. Apart from the fact that the explicit calculation of the returns earned by a portfolio of selected stocks may be considered desirable as far as a research design's external validity[5] is concerned, the decision to adopt a sorting approach has mostly been rationalized on the basis of the data's alleged proneness to a threat to statistical conclusion validity that Cook and Campbell (1979) have labelled **low reliability of measures**. Indeed, financial statement figures are likely to contain quite a lot of noise. In a regression context, such measurement error in the independent variable has two consequences: firstly, it introduces a downward bias in the ordinary least squares (OLS) estimate of the slope coefficient (e.g. Theil (1971)) and, secondly, it leads to an increase in the standard error of that estimate.

In turn, both of these consequences have a downward impact on the t-statistic that is commonly used to assess whether the slope coefficient is statistically significantly different from zero.

This immediately brings me to a second potential threat to statistical conclusion validity, namely that of **low statistical power**. After all, a downward bias in the t-statistic implies that a bias in favour of the null hypothesis is introduced and, hence, that the significance test used is rendered less powerful. One has to wonder, though, whether the initial assertion that regression analysis is more susceptible to the bias resulting from noise in accounting data, is appropriate to begin with. According to Lys and Sabino (1992), it is not. They have formally demonstrated that the measurement error-induced bias is not mitigated by switching to sorting-based tests. On the contrary, Lys and Sabino have shown that as a result of the reduced variability in the independent variable, sorting techniques are uniformly less powerful than regression. Concretely, they claim that when the grouping variable is normally distributed, regression is up to 45% (dependent upon the sample size and the degree of correlation between the dependent and the independent variable) more powerful than a corresponding sorting strategy for which the power is maximized. On top of that, such maximization of power has not been achieved in any of the many EMH studies that have relied on some kind of sorting procedure. Lys and Sabino claim that for the degrees of correlation between abnormal returns and grouping variables that are typically present in market efficiency tests, the optimal proportion of observations in the extreme portfolios is about 27%. Clearly, this proportion is considerably larger than the proportion of observations that is present in the extreme groups when one focuses on quintiles or deciles, as EMH researchers have generally done. Lys and Sabino contend, for example, that regression may be no less than 85% more powerful than a decile-based sorting algorithm. These differences in power are obviously not negligible. In fact, they constitute an important enough advantage of regression analysis to certainly continue to use the latter in EMH research.

---

[5] See section seven.

But what about sorting then? Do sorting-based tests have no use whatsoever beyond regression analyses and are the results obtained in market efficiency studies that have solely relied on grouping techniques worthless? I do not think so. As far as the existing anomalies literature is concerned, the threat of low statistical power is not particularly relevant. After all, the various anomalies derive from rejections of the null hypothesis of market efficiency. Since the power of a test is equal to the probability of correctly rejecting the null, rejections obtained on the basis of tests that had relatively low power actually constitute stronger evidence against the case of market efficiency. On the other hand this does not imply, of course, that future studies should try to keep the power of their empirical tests low. Given a certain sample size and a pre-specified significance level, it is merely logical to try to maximize statistical test power so as to minimize the probability of committing a Type II error. So in this respect the preference for regression analysis stands up.

Sorting-based tests have got, however, a major advantage over regression tests in that they do not impose a specific functional relation between the grouping variable and the dependent variable. In my opinion, this advantage is everything but trivial, especially in view of the general lack of a well-elaborated theory that relates currently observable information to future abnormal returns, i.e. a theory of market inefficiency. On the whole, I am therefore inclined to suggest that regression and sorting tests be used in combination in EMH studies. Later on in this paper, in section seven to be precise, I shall argue that this combination also has its merits in terms of a research design's external validity. As far as statistical conclusion validity is concerned, it offers the advantage that possible inefficiencies that might remain unidentified if one were to use just one of the two approaches are more likely to be uncovered. This thanks to the fact that a regression test enables the researcher to subject the null hypothesis of market efficiency to more powerful scrutiny, while the use of a sorting algorithm allows for more flexibility with respect to the nature of the functional relation between the independent and dependent variables. The latter consideration does not, however, free authors from seeing to the power of the sorting-based tests they employ. In my view, the findings of Lys and Sabino (1992) regarding the optimal proportion of observations to be taken up

in the extreme portfolios should therefore definitely receive more attention in future stock selection studies.

4.2. VIOLATION OF ASSUMPTIONS OF THE STATISTICAL TESTS USED

Probably the most important threat to statistical conclusion validity is the one described by Cook and Campbell (1979) as the **violation of assumptions of the statistical tests used**. Obviously, statistical tests are only useful if the assumptions that underlie them are satisfied. As Cook and Campbell rightfully indicate, not all assumptions are equally critical of course. However, in the specific context of market efficiency research statistics-related difficulties in drawing inferences have proved to be so pertinent that a separate stream of literature on the measurement of long-horizon (i.e. one year or more) stock performance has developed. Before looking into those difficulties and some of the solutions that have been suggested, I need to define a few concepts to which I shall repeatedly refer in the discussion that follows.

More specifically, I have to describe the distinction between *cumulative abnormal returns* (henceforth CAR's) and *buy-and-hold abnormal returns* (henceforth BHAR's). The theoretical notion of "abnormal return" may easily be defined as the difference between the actual return on an investment and the expected return on that investment. In practice, academics have, however, been faced with the lack of a truly satisfying model to estimate expected returns. As a result, a well nigh endless series of abnormal return constructs has come about. The extent to which the most important ones of those constructs may be considered more or less appropriate operationalizations of the abnormal return concept is rather a matter of construct validity. That discussion will, therefore, be deferred to section six. Here I focus solely on statistical issues and in that respect especially the method used for the accumulation of returns is of interest.

The accumulation problem is rooted in the fact that long-term returns are to be seen as aggregates of returns over shorter time intervals[6]. As such, monthly

[6] Continuously compounded returns constitute a notable exception of course, although they are in fact nothing but an accumulation of returns over infinitely short time intervals. However, continuously compounded return

returns have served as original data in most long-horizon market efficiency studies. Roughly speaking, two methods of accumulation can be discerned. The first one consists of simply taking the sum of the monthly abnormal returns. So, letting AR_{it} represent the abnormal return earned on security i during month t, whereby $AR_{it} = R_{it} - \mathrm{E}(R_{it})$, the CAR for that security over a period of τ months ($CAR_{i\tau}$) may be formally described as

$$CAR_{i\tau} = \sum_{t=1}^{\tau} AR_{it} \tag{1}$$

In contrast, the $BHAR$ of security i over a τ-month period ($BHAR_{i\tau}$) is defined as the difference between the compounded monthly returns on that security and its compounded expected monthly returns:

$$BHAR_{i\tau} = \prod_{t=1}^{\tau}\left[1 + R_{it}\right] - \prod_{t=1}^{\tau}\left[1 + \mathrm{E}(R_{it})\right] \tag{2}$$

The most pronounced difference between these two calculation methods is probably the discrepancy at the level of the assumptions that underlie them: whereas CAR's involve the implicit assumption of rebalancing after each sub-period (i.e. here, after each month), BHAR's are based on the assumption that no more trading takes place after the initial investment position has been taken at the beginning of the first sub-period.

Having defined the main approaches to calculating long-horizon returns, I can now turn to a discussion of the most important statistical difficulties that have been identified in this context. First of all, I concentrate on the difficulties encountered in sorting-based tests. Remember that the null hypothesis of market efficiency claims that it is impossible to use just the information that is available at a given point in time to compose a portfolio of stocks that will subsequently earn predictably positive or negative abnormal returns. As such, authors' answers to the research question of stock market efficiency have traditionally been dependent upon the value obtained for some measure used to assess the statistical significance of the mean abnormal return earned by the selected stocks. In other words, virtually

---

calculation has hardly received any attention in the empirical market efficiency literature. Barber and Lyon (1997) and Kothari and Warner (1997) are among the few who have investigated the properties of continuously compounded returns and they conclude that the latter produce negatively biased estimates of long-run abnormal

all of the statistical problems encountered in efficient markets research may ultimately be reduced to troublesome estimation of the extent to which the selected stocks earn a mean abnormal return that is statistically significantly different from zero, the latter being the value predicted under the null hypothesis of course. The most conventional measure to assess the statistical significance of abnormal returns is the following parametric test statistic:

$$t = \overline{ARi\tau} \Big/ \left[\sigma(ARi\tau) \Big/ \sqrt{n}\right], \tag{3}$$

where $ARi\tau$ may be either the CAR or the BHAR for security i over a period of τ months, $\overline{ARi\tau}$ is the cross-sectional sample mean of the τ-month abnormal returns of the selected stocks, $\sigma(ARi\tau)$ is the cross-sectional sample standard deviation of those abnormal returns and n is the number of selected stocks. If the sample of n stocks is drawn randomly from a normal distribution, the test statistic calculated in equation 3 will follow a Student's t-distribution under the null hypothesis so that the probability of obtaining its sample value may easily be computed. The problems that have to be overcome in order to be able to draw meaningful conclusions on the basis of this conventional t-statistic are legion, however.

First of all, one is faced with the requirement of normality as to the distribution of the abnormal returns in order for the test statistic to follow a t-distribution under the null hypothesis. It is, however, a well-established fact that long-term abnormal returns are clearly non-normally distributed. Kothari and Warner (1997)[7], for example, show that the distributions of both CAR's and BHAR's are fat-tailed compared to normal distributions. Furthermore, they demonstrate that the distribution of BHAR's is heavily skewed to the right, while the distribution of CAR's shows slight left-sided skewness. In itself, this non-normality need not constitute an insurmountable difficulty. After all, the Central Limit Theorem would normally guarantee that for large sample sizes the distribution of the mean abnormal return converges to normality anyway. Unfortunately, the Central Limit Theorem only

---

returns. Consequently, in the context of efficient markets research inferences drawn on the basis of continuously compounded returns are prone to being spurious.

[7] It should be noted that Kothari and Warner (1997) use a modified BHAR measure, i.e. $BHARi\tau = \prod_{t=1}^{\tau}(1+ARit) - 1$. Given that this modification hardly affects the statistical properties of the BHAR's, I do not devote special attention to it in this section. As constructs and as far as the implied underlying investment strategies are

applies to independent and identically distributed drawings from finite variance distributions. Since, as I shall discuss in more detail later in this section, abnormal returns are generally not independent in the cross-section, convergence of the mean abnormal return distribution to normality is not guaranteed. Consequently, the conventional test statistic defined in equation 3 may not follow a t-distribution under the null hypothesis.

On top of that, serious problems might arise as far as the estimation of the separate components of the test statistic is concerned. Basically, the test statistic consists of two components: the mean abnormal return in the numerator and its standard error in the denominator. Important biases have been found to manifest themselves in the traditional methods used to estimate both of these components. First, I focus on the mean abnormal return parameter. The expected value of this parameter should normally be zero under the null hypothesis. Simulations by Kothari and Warner (1997) and Barber and Lyon (1997) have shown, however, that for longer time horizons this expected value may deviate significantly from zero for a random sample of stocks[8]. Here are some of the potential reasons underlying the observed biases:

- many abnormal return measures require that firm-specific parameters be estimated. Usually, this estimation is conducted on the basis of data pertaining to a period that immediately precedes the actual test period. Companies for which such estimation period data are incomplete or simply unavailable (e.g. firms that were newly listed during the estimation period or at the beginning of the test period) are then automatically excluded from the set of feasible investment opportunities. Nevertheless, those newly listed companies are often taken into consideration as far as the calculation of the expected return benchmarks is concerned. This would not be problematic, if

---

concerned, the differences between the modified BHAR measure and the BHAR measure defined in equation 2 are not trivial, though. Therefore, I shall treat them separately in section six on construct validity.

[8] It should be noted that these simulation-based results themselves are quite dependent upon the specific methodology that is being used as well as upon the time period that is under review. The contradictory findings with respect to BHAR's of Kothari and Warner (1997) on the one hand and Barber and Lyon (1997) on the other hand are an excellent example of this lack of stability: for a simulated distribution of 1,000 generated random samples of 200 firms using data for the 1963-1994 period, Barber and Lyon find the mean three-year BHAR to be negatively biased. Using a somewhat less extensive dataset (250 random samples of 200 firms during the 1980-1989 period) and a slightly modified BHAR measure (see the preceding footnote), Kothari and Warner conclude that the expected value of the mean three-year BHAR is actually significantly positive. Consequently, cautious interpretation of such simulation-based evidence is clearly called for.

it were not for the regularity that newly listed stocks have been shown to perform worse than the market (e.g. Ritter (1991)). As a result, the estimate of the mean abnormal return associated with a particular investment strategy may be biased upward due to this so-called *data requirement bias*;

- a very closely related bias is the *new listing bias* (Barber and Lyon (1997)). Even if no pre-test period data are required, computation of expected return benchmarks including firms that list for the first time during the test period may still create an upward bias in the estimated mean abnormal return;
- specific to BHAR's is the so-called *rebalancing bias* (Barber and Lyon (1997)). This bias is rooted in the fact that in many studies the proxy for the long-horizon expected return is obtained through the compounding of the returns earned by some benchmark portfolio that is equally weighted for each sub-period, i.e. month. The equal weighting for each month implicitly entails the assumption of monthly rebalancing. Due to the presence of negative autocorrelation in monthly returns (as evidenced by Canina, Thaler and Womack (1998), among others), this implicit rebalancing may cause the benchmark returns to be inflated. After all, the rebalancing naturally involves the purchase (sale) of stocks that have performed worse (better) than the benchmark as a whole during the last sub-period. As such, the estimated long-horizon benchmark returns are actually derived on the basis of the implicit assumption that a trading strategy is adopted which may be expected to take advantage of any negative autocorrelation in monthly returns[9]. If no similar trading strategy is assumed for the portfolio of selected stocks, i.e. if no monthly rebalancing is implied, the estimated mean abnormal return for that portfolio will be biased downward.

Although the importance of the above biases with respect to the mean abnormal return parameter should not be underestimated, it is clear that they may be avoided quite easily. Both the data requirement bias and the new listing bias can be removed by simply ensuring that the benchmarks only contain firms that are eligible for selection at the beginning of the test period. The rebalancing bias can of

[9] As Barber and Lyon (1997) emphasize, the negative autocorrelation in monthly returns does not necessarily constitute evidence of a market inefficiency itself. Although the autocorrelations in short-horizon returns have generally been statistically significant, their economic significance has remained questionable to say the least.

course be avoided by compounding the returns on a benchmark portfolio that is not re-weighted after each month. Unfortunately, the story does not end there. In fact, the usefulness of the conventional test statistic is threatened more by the biases to which the estimation of its denominator, i.e. the standard error, is potentially susceptible:

- as Kothari and Warner (1997) show, also the estimate of the standard deviation may be prone to a *data requirement bias*. Sometimes pre-test period returns are used to estimate the cross-sectional standard deviation of abnormal returns. Such an estimate is conditional in that the firms included for its calculation are all required to have survived the estimation period preceding the test period. As it appears, that estimation period variance is considerably lower than the test period variance which is unconditional upon further survival. So, in that case the true abnormal return standard deviation during the test period is underestimated and an upward bias in the test statistic (in absolute value) is introduced. The most obvious way to tackle this issue is of course to use test period returns in the standard deviation estimation procedure, although Kothari and Warner's empirical analysis fails to document an improvement in the specification of the test statistic when this alternative procedure is followed;
- quite closely related is the *prediction error variability bias*. Although the discussion of the main possible constructs for abnormal returns is deferred until section six, explanation of this bias is impossible without briefly describing one of the potential abnormal return proxies. A fairly popular technique consists of setting abnormal returns equal to the prediction errors from a model designed to capture the stochastic process of stock returns (e.g. the market model), whereby the firm-specific model parameters are in most cases estimated on the basis of pre-test period data. In some instances, the residuals from the regressions used to estimate those firm-specific model parameters serve as input to calculate the estimated abnormal return standard deviation. Prediction errors have, however, been shown to be more variable than fitted regression residuals (e.g. Maddala (1988)). As a result, the true abnormal return standard deviation is underestimated if regression residuals from a pre-test period are used in its estimation. Kothari and Warner (1997) suggest lengthening of the estimation period as a way to

mitigate this bias. They claim that a larger set of pre-test period data leads to more accurate estimates of the model parameters and, in turn, to a more accurate image of abnormal return variability. I do not deny that this claim is correct, but it should be noted that lengthening of the estimation period is likely to exacerbate the data requirement bias discussed in the preceding point. Besides, less laborious solutions are available (e.g. estimation of the abnormal return standard deviation on the basis of the test period prediction errors or usage of an abnormal return construct that does not rely on prediction errors);

- earlier in this section I have already described the asymmetrical nature of abnormal return distributions. Especially long-run BHAR's tend to be heavily right-skewed. The particular shape of the distribution induces a so-called *skewness bias* in sample-based estimates of its standard deviation. Concretely, the bias stems from the fact that the means and standard deviations of samples taken from right-skewed distributions are positively correlated with one another. Barber and Lyon (1997, pp. 347-348) provide a very clear description of the intuition behind the impact of this correlation, which may be summarized as follows: conditional upon observing a positive (negative) sample mean, it is more (less) likely that the particular sample contains one or more of the extreme positive observations. As a consequence, the extreme positive observations are probably overrepresented (underrepresented) in the sample under consideration, which, in turn, leads to an inflated (deflated) estimate of the distribution's true standard deviation;
- finally, an important source of misspecification of the test statistic lies in the *cross-correlation bias*. Abstracting from the potential biases described above, a prerequisite for the test statistic being correctly specified is the independence of observations. It is, however, a well-known fact (e.g. Bernard (1987)) that the frequently used expected return benchmarks fail to capture all common variation in stock returns. Particularly intra-industry cross-sectional correlations may be quite considerable for longer-horizon abnormal returns[10]. Due to this cross-sectional dependence in abnormal returns the standard error of the mean abnormal return parameter is generally

[10] Bernard (1987) makes mention of a mean intra-industry correlation of 30% for annual CAR's during the 1955-1984 period.

underestimated. After all, if n in the denominator of the standard error formula is simply set equal to the sample size, the number of independent observations is likely to be overestimated. As a consequence, the test statistic will be biased upward (in absolute value).

It should be obvious that these last two sources of bias are the hardest nuts to crack. As a matter of fact, it has proved to be virtually impossible to avoid them without being prepared to sacrifice the conventional procedure for the assessment of statistical significance based upon the comparison of a cross-section-based test statistic's computed value to a theoretical probability distribution. To be clear, the core of the issue lies in the fact that biases in its constituent parts (i.e. the estimated standard error as far as the skewness and cross-correlation biases are concerned) cause the test statistic to be misspecified. This misspecification reveals itself in discrepancies between actual rejection frequencies and theoretical rejection frequencies, whereby their relative magnitude is of course dependent upon the relative importance of the various biases in the context under consideration. Stated otherwise, the estimated p-values associated with a particular test statistic value are also biased due to the fact that the test statistic does not follow a t-distribution under the null hypothesis. As a result, any pre-specified level of significance becomes, in fact, meaningless.

Barber and Lyon (1997) and Kothari and Warner (1997) have demonstrated that misspecification of the conventional test statistic is not typical of a particular expected return benchmark. Apparently, the frequencies with which the null hypothesis is rejected deviate considerably from their theoretical levels for basically all benchmarks that have commonly been used in efficient markets research. Admittedly, Barber and Lyon contend that a procedure whereby sample firms are matched to control firms of similar sizes and B/M ratios yields well-specified test statistics in most sampling situations, but I feel that the importance of this finding should be put into perspective. Firstly, the quite fragile robustness of Barber and Lyon's results in general renders the generalizability of their evidence with respect to the control firm approach somewhat dubious. And secondly, as I shall discuss in much more detail in section six, an expected return benchmark based on size and B/M factors may not be fully appropriate to test the EMH.

That is why I am inclined to concur with Kothari and Warner (1997) who claim that the conventional t-statistic should probably be substituted by some alternative approach if one is to avoid drawing spurious inferences regarding the statistical significance of the abnormal returns associated with a particular investment strategy. One such alternative involves the use of *nonparametric significance tests*. Although the latter have the advantage of imposing less strict requirements on the distribution from which the observations are supposed to originate, I do not think that tests like the ones used by Spiess and Affleck-Graves (1995), for example, provide a satisfactory solution to the issue of troublesome significance assessment. After all, one must not forget that a zero median under the null hypothesis and independence of the observations are recurring assumptions in most nonparametric procedures. These assumptions are clearly threatened by the skewness and cross-correlation biases, respectively. As a matter of fact, Brown and Warner (1980) have explicitly demonstrated that there is considerable misspecification in sign tests when monthly data are used.

Another alternative approach has been suggested by *Fama and MacBeth* (1973). Under this approach, the form of the conventional t-statistic is maintained, but it is operationalized in a different manner. Specifically, the time-series standard error is substituted for the cross-sectional standard error. In other words, the series of mean abnormal portfolio returns for the various sub-periods (e.g. years) of the test period is used to derive a standard deviation estimate from. Over the years, the Fama-MacBeth methodology has definitely become the most popular approach to significance testing in sorting-based examinations of market efficiency with respect to accounting information. Bernard and Thomas (1990), Sloan (1996), Abarbanell and Bushee (1998) and Fama and French (1998), for example, all make use of it. However, while the Fama-MacBeth approach has the potential to considerably mitigate the problems of cross-correlation and skewness, it may suffer from a major drawback as far as its practical applicability is concerned. Calculation of a meaningful time-series standard error requires that a sufficiently large number of time-series observations be available. Given that the observations are long-horizon abnormal returns here, this requirement may be quite hard to meet.

That is why I am inclined to give preference to a relatively more advanced technique that has received increasing attention in EMH research in recent years, namely *bootstrapping*. Bootstrapping procedures may be applied to either the test statistic or the observations themselves. Basically, bootstrapping-based inference comes down to the generation of a simulated distribution of the test statistic or parameter of interest under the null hypothesis. The simulation of the empirical distribution is achieved by drawing a pre-specified number (e.g. 1,000) of re-samples from the original sample and calculating the test statistic or parameter's value for each of these re-samples. A most interesting article in this context is Lyon, Barber and Tsai (1999). They examine the specification of significance tests based upon (a) a bootstrapped test statistic[11] and (b) bootstrapped abnormal returns. For nearly all sampling situations, Lyon, Barber and Tsai are unable to reject the hypothesis that their tests are correctly specified. Albeit that simulated evidence regarding the specification of test statistics should generally be interpreted with caution, the results of Lyon, Barber and Tsai constitute in fact nothing but empirical confirmation of what might be expected on theoretical grounds. After all, by construction, the technique of bootstrapping ought to lead to the development of a distribution that incorporates the typical features of the test statistic or parameter of interest in the specific research context under consideration. As such, the skewness bias present in conventional significance tests of long-term abnormal returns should be seriously mitigated, if not removed, because the generated empirical distribution will normally account for that skewness. Although the cross-correlation issue is not directly addressed by adopting a bootstrapping procedure, the latter is highly likely to considerably reduce this bias as well. In normal circumstances, the mean cross-sectional correlation of the various re-samples taken in the process of bootstrapping will approximate the average cross-sectional correlation in the sample as a whole. As such, the cross-correlation among observations should automatically be corrected for, except for the case in which the correlation among the abnormal returns of the selected stocks deviates substantially from the average cross-sectional sample correlation. Lyon, Barber

---

[11] To be precise, Lyon, Barber and Tsai (1999) apply the bootstrapping procedure to a *skewness-adjusted* test statistic. The adjustment is taken from Johnson (1978) and leads to the following transformed t-statistic: $t_S = \sqrt{n}\left(S + \frac{1}{3}\cdot\hat{\gamma}S^2 + \frac{1}{6n}\cdot\hat{\gamma}\right)$, where $S = \overline{AR_{it}} \Big/ \sigma(AR_{it})$ and $\hat{\gamma} = \frac{\sum_{i=1}^{n}\left(AR_{it} - \overline{AR_{it}}\right)^3}{n\sigma(AR_{it})^3}$.

and Tsai demonstrate empirically, however, that apart from when there is extreme industry clustering in the portfolio of selected stocks, this problem is unlikely to pose itself in practice.

All this should certainly not be taken to imply that bootstrapping techniques are flawless. One must, for example, not overlook the fact that the drawing of the re-samples requires availability of a sufficient number of observations in the cross-section, which may not always be the case. Also, bootstrapping is merely a very general procedure that can be implemented in a large number of ways. For instance, Ikenberry, Lakonishok and Vermaelen (1995), Frankel and Lee (1998), Lyon, Barber and Tsai (1999) and Desai and Jain (2000) all use slightly different bootstrapping techniques in their investigations of market efficiency. I shall not look into the details of the various potential specifications here, but it is obvious that simply the choice of the expected return benchmark may have a serious impact. Stated otherwise, bootstrapping cannot in itself correct for any bad-model problems[12]. Sensitivity analyses as to the influence of the use of a particular bootstrapping procedure are therefore certainly called for.

Nevertheless, as far as the assessment of the statistical significance of long-run abnormal returns is concerned, the theoretical considerations described above as well as the existing simulation-based empirical evidence should make the supremacy of bootstrapping techniques in this context quite clear. The fact that they dominate more traditional approaches to estimating the p-value associated with a particular mean abnormal return is well-documented. Consequently, I find it merely logical to strongly recommend their use in empirical tests of market efficiency.

So far, I have focused on violations of the statistical assumptions in sorting-based EMH research. It should be obvious, however, that regression-based tests are not free from methodological difficulties either. In order for the traditional OLS coefficient estimates to satisfy the so-called "B.L.U.E."[13] property, quite some assumptions have to be fulfilled and at least a few of them will commonly be

[12] See section six.
[13] Best Linear Unbiased Estimator.

violated in cross-sectional (or pooled) regressions of abnormal returns on firm-specific variables. One of those assumptions is the multivariate *normality* of the disturbance vector. Especially if BHAR's are used to measure the dependent variable, the skewness of their distribution may cause the regression residuals to be non-normally distributed as well. If the probable lack of independence among the observations makes that the Central Limit Theorem does not apply, that is. Luckily, the OLS method is fairly robust to violations of the normality assumption. The conventional statistical significance tests may no longer produce the most powerful results for the given non-normal distribution, but no insurmountable problems should generally arise.

Much more troublesome are violations of the assumption of *homoscedasticity*. Disturbances are said to be homoscedastic if their variance remains constant across all independent variable values. The homoscedastic nature of the disturbances is a crucial assumption underlying the application of OLS. If it is violated, coefficient estimates remain unbiased and consistent, but they are no longer efficient (i.e. their variance is not minimal anymore) and the t-statistic that is commonly used to assess the significance of regression coefficients will yield incorrect results. This last consequence results from the fact that the general OLS formula for calculating the variances of the coefficient estimates critically relies upon the assumption of homogeneous residual variances. Otherwise the formula tends to produce downwardly biased estimates of the true parameter variances, so that the t-statistic is generally overestimated (in absolute value).

Heteroscedasticity is very common in cross-sectional regressions due to the latter's proneness to so-called scale effects. As such, the use of (abnormal) returns as the dependent variable is positive in that the beginning-of-period price in the denominator automatically serves as a scaling factor that is likely to mitigate the degree of heteroscedasticity. On the other hand, the potential problem of heteroscedasticity may be worsened by the difficulties inherent in measuring abnormal returns (i.e. the joint-hypothesis issue[14]). Due to those difficulties most abnormal return proxies contain measurement error. To be more precise, a portion

[14] For more detailed discussions I refer to sections five and, in particular, section six.

of the alleged abnormal returns might very well reward some ex ante risk factor that the expected return benchmark used has failed to fully capture. Any firm-specific independent variable that happens to proxy for the risk factor concerned will then not only be associated with the "abnormal" returns, but probably also with their variability. After all, risk is bound to reveal itself in terms of return variability. Consequently, the problems involved in filtering out true abnormal returns from raw returns increase the probability of regression residuals being heteroscedastic. Fortunately, there are quite a few ways to deal with the issue of non-constant disturbance variance.

Short of transforming the variables, heteroscedasticity may be adjusted for by using a so-called *Weighted Least Squares* (WLS) *procedure* (e.g. Gujarati (1995)). WLS involves weighting each observation by the inverse of the standard deviation of the disturbance. The logic behind it is that observations with a larger disturbance variance (as proxied for by the regression residuals) are more imprecise and should therefore be down-weighted when it comes to estimating the regression coefficients. Application of WLS yields estimators that are B.L.U.E. However, WLS may prove to be relatively impractical in that it basically requires the true disturbance variance structure to be known. Given that this is rarely the case, one generally has to resort to estimating that disturbance variance structure on the basis of the available dataset. If the latter is not rather extensive in terms of the number of time periods available, such estimation can prove to be quite tricky.

That probably explains why most authors have settled for a less complete but somewhat more practical adjustment procedure. *White* (1980) has provided such a procedure in the form of formulae for obtaining so-called heteroscedasticity-consistent standard errors. Substitution of the latter for the traditional standard error estimates makes it possible to draw correct inferences regarding statistical significance on the basis of the conventional test statistics. White's procedure has the major advantage that it may also be applied if the true disturbance variance structure is unknown. Its main minus lies in the fact that it does not fully correct for the presence of heteroscedasticity in that the coefficient estimates themselves are left undisturbed, so that they remain inefficient. Nevertheless, the White-

adjustment provides a relatively easy and neat solution to the issue of heteroscedasticity, which is far too important to be ignored.

Unfortunately, as far as cross-sectional regressions in market efficiency tests are concerned, the story does normally not end with adjusting for heteroscedasticity. Another critical assumption underlying OLS is also likely to be violated. Homoscedasticity is in fact just one part of the bipartite assumption of *spherical disturbances*. As such, OLS does not only require that all diagonal elements of the residual variance-covariance matrix are equal, but also that all of its off-diagonal elements are zero. In other words, there should be no *cross-correlation* among the disturbances. As it happens, I have already described earlier in this section that long-run abnormal returns tend to be correlated in the cross-section. If then the independent variable taken up in the regression-based test of market efficiency does not capture the common factor that is responsible for the cross-sectional correlation among abnormal returns (and to the extent that it is truly firm-specific it will not), the regression residuals will be cross-correlated as well. The statistical consequences of cross-correlated disturbances are very similar to those of heteroscedasticity: estimators' unbiasedness and consistency are unaffected, but they are no longer efficient and hypothesis testing becomes troublesome. As such, cross-correlation among regression residuals calls for corrective action just as much as the presence of heteroscedasticity does. There is, however, no clear-cut solution to the cross-correlation issue, which probably explains why the literature on the specification of regression-based EMH tests has given it considerably more attention than violations of other assumptions.

Again, using the time-series standard error as suggested by *Fama and MacBeth* (1973) and extended by Sefcik and Thompson (1986) constitutes one possible remedy. The remark made earlier regarding the potentially doubtful practical feasibility of this approach remains valid, though. In addition, the Fama-MacBeth procedure fails to tackle the inefficiency of the coefficient estimators. This makes it fairly easy to find an alternative approach that is, at least theoretically, superior.

The *Generalized Least Squares* (GLS) *method* is such an alternative (see for example Johnston (1984), pp. 287-342). Under GLS, the regression model to be estimated is

pre-multiplied by a transformation matrix. The latter is thereby chosen in such a manner that the assumption of spherical disturbances is satisfied for the transformed model. In turn, this makes it possible to apply OLS to obtain coefficients that are B.L.U.E. GLS may be perceived as a more general form of the WLS method described in the preceding paragraph. It should therefore not come as a surprise that the practical impediments to using WLS apply just as well in the case of GLS. Concretely, GLS also requires the true disturbance variance structure to be known with a view to computing the transformation matrix. Since such knowledge is next to never available, it is usually necessary to fall back on an estimated version of the residual variance-covariance matrix with the accompanying weaknesses. For instance, as indicated by Marais (1986) and Bernard (1987), sampling error in the estimated disturbance variance structure may render the GLS coefficients inefficient and may introduce bias in the commonly used test statistics pertaining to those coefficients. Much more importantly, though, is the fact that estimation of the residual variance-covariance matrix will very often simply be impossible in empirical tests of market efficiency. One must realize that for a sample of n firms, the residual variance-covariance matrix contains no less than n^2 elements. The matrix' symmetry makes $\frac{n(n-1)}{2}$ elements basically redundant, but that still leaves $n^2 - \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$ elements that have to be estimated. This implies that the number of time periods available has to exceed about half the number of cross-sectional observations. Clearly, in studies that focus on long-horizon returns this requirement may be impossible to meet even for extremely small sample sizes in the cross-section. But even if the return horizon is shortened, say to one month, application of GLS requires the number of sample firms to be small. After all, being able to obtain *an* estimate of the residual variance-covariance matrix is one thing, being able to obtain a *satisfactory* estimate is another. It is those practical obstacles to applying the theoretically superior method of GLS that have encouraged academics to look for more feasible approaches without thereby ignoring the cross-correlation issue.

The approach that I personally prefer is the *Constant Correlation Model* (CCM) *regression* suggested by Chandra and Balachandran (1992). CCM methodology was

initially developed for selecting efficient portfolios (e.g. Elton and Gruber (1973) and Elton, Gruber and Padberg (1976) and (1977)). It may be regarded as the golden mean, so to speak, between OLS and GLS in that it produces coefficient estimators that take the cross-sectional correlations partially into account. Concretely, CCM departs from the clustering of similar firms. In principle, one is free to choose the factor based upon which the grouping is performed, but Chandra and Balachandran suggest industry membership. In that case, CCM comes down to imposing a specific structure upon the intra-industry and inter-industry correlations, respectively: all correlations between firms of the same industry are set equal to the mean correlation between firms of that industry. Analogously, correlations between firms belonging to different industries are set equal to the mean correlation for all pairs of firms from those two industries. By adopting the assumption of intra-industry homogeneity, the number of parameters to be estimated in the residual variance-covariance matrix is reduced to $n + \frac{m(m+1)}{2}$, where m represents the number of industries. As long as m is kept quite low, this reduction will generally be substantial enough to not only make estimation feasible, but also to obtain relatively reliable estimates. In fact, Chandra and Balachandran demonstrate on the basis of simulations that CCM's small sample properties are actually better than those of GLS.

These considerations must not make one blind to the potential shortcomings of the CCM procedure. Most important thereby is CCM's reliance on the assumption of within-group homogeneity. If a particular sample requires that the number of groups be kept quite large in order to ensure such homogeneity, the main advantage of CCM (i.e. the reduction in the number of parameters to be estimated) can be seriously eroded. An obvious trade-off exists between the degree of within-group homogeneity on the one hand and the feasibility of the estimation on the other hand. In cases where the dissimilarities across firms are just too marked, CCM may hardly be more practical than GLS. Ultimately, the usefulness of the CCM method is therefore largely dependent upon the specific research context. If the context allows that CCM be applied, I would definitely recommend it, though. If the context does not allow it, and bearing in mind the minuses of the alternatives

discussed above, researchers have in fact no choice but to fall back on what I would describe as the "approach of last resort".

In instances where it is impossible to satisfyingly account for the cross-sectional correlations, one's only option consists of trying to eliminate those correlations. Adapting the expected return benchmark in such a manner that abnormal returns are purged from any remaining common factors is thereby the most obvious way to go. I cannot stress enough, however, that this approach should only be used if all alternatives fail. After all, aside from the fact that it can prove to be quite a challenge to identify additional factors that actually eliminate the cross-correlation, such factors may introduce a bias against rejecting the null hypothesis by capturing a portion of the treatment effect[15]. Suppose, for example, that one wants to examine the value-versus-growth anomaly and does so by regressing abnormal returns on lagged book-to-market ratios (henceforth B/M ratios). Furthermore, assume that firms with relatively high B/M ratios tend to be clustered in a small number of industries and firms with relatively low B/M ratios in a few other industries. If, in such a case, the researcher tries to avoid the cross-correlation issue by taking up an industry factor in the calculation of abnormal returns, the probability of detecting a B/M effect will be considerably reduced. After all, the thus estimated abnormal returns would only reflect within-industry variation. To the extent that the industry factor added constitutes or at least proxies for some ex ante risk factor that ought to be priced and that the original expected return benchmark failed to capture adequately, there is no problem. If this is not the case, however, abnormal returns will no longer be appropriately measured and construct validity will be threatened. I shall discuss the latter possibility in more detail in section six.

4.3. FISHING AND THE ERROR RATE PROBLEM

After this very extensive discussion of the various assumptions underlying statistical tests that are likely to be violated in empirical EMH research, the reader may be under the impression that all threats to statistical conclusion validity have surely been dealt with by now. However, Cook and Campbell (1979) identify one

more threat that is certainly relevant in the context of market efficiency studies, namely the so-called **fishing and error rate problem**. The latter has to do with the increasing likelihood of falsely concluding that the treatment and outcome constructs covary (i.e. committing a Type I error) when multiple conceptions of tests of the null hypothesis are possible and it is not recognized that a certain proportion of the conceptions will yield statistically significant results by chance. Normally, the probability that a true null hypothesis is rejected by a correctly specified test is equal to the pre-determined significance level α. If that null hypothesis is subjected to several tests, the probability that it is falsely rejected by at least one of them obviously rises above α. Using that one rejection as evidence to demonstrate that the null hypothesis is incorrect does then no longer constitute a statistically significant result at significance level α of course.

The relevance of the fishing problem for market efficiency research can hardly be overestimated. The (semi-strong version of the) EMH is so general in its predictions (i.e. security prices reflect *all* (publicly) available information) that the number of ways in which it may be empirically tested on the basis of a given sample is nearly inexhaustible. In principle, *any* subset of the information available at a given point in time may be examined for its relationship with subsequent abnormal returns. It is crucial to realize that for a particular dataset a portion of those information subsets will appear to be significant predictors of abnormal returns by chance. Clearly, this need not imply that the information subsets concerned should be considered to be at the centre of some systematic market inefficiencies. Proponents of the EMH (e.g. Ball (1992) and Fama (1998)) have seized upon this threat to statistical conclusion validity to dispose of most of the anomalous findings by deeming them the fruits of a drawn-out data mining exercise. According to them, the lack of a comprehensive "theory of inefficiency" has incited researchers to start fishing haphazardly in the enormous pond of information subsets with respect to which the EMH may be examined. In combination with a publication bias in favour of results inconsistent with the EMH's predictions, this is supposed to have led to a very distorted image of the actual degree of stock markets' informational efficiency.

---

[15] See Salamon (1985) for a detailed discussion of this issue.

I concur only partly with this line of reasoning. It is true that it is very hard, if not impossible, for the average reader of an article to assess whether the revelation of a particular anomaly has resulted from a multitude of efforts to discover evidence that is inconsistent with market efficiency. I also agree that the inherently more "shocking" nature of anomalous findings has created a publication bias in their favour. When it comes to the claim that basically all of the allegedly anomalous evidence may be attributed to fishing, however, I beg to differ. By now, some anomalies can fall back on quite a history of research. The value-versus-growth anomaly and especially PEAD are examples of accounting-based anomalies that find themselves in this case. They have survived an impressive array of thorough sensitivity analyses and have been found to exist in different time periods and across various stock markets. Without wanting to imply that both of these anomalies constitute actual rejections of the EMH, I find it extremely hard to believe that the statistical covariations upon which they are based ought to be deemed chance results.

Admittedly, there are also accounting-based anomalies that are exceptionally prone to the threat of fishing. Textbook examples are of course the Ou and Penman (1989a) and Holthausen and Larcker (1992) anomalies, which are susceptible to fishing by construction. The authors of these anomalies explicitly use data mining approaches in their quest for accounting variables with predictive qualities with respect to abnormal returns. In my opinion, a pure data mining procedure is simply incapable of producing meaningful evidence as far as market efficiency is concerned. Any apparent relationship between a variable selected on the basis of the available data and subsequent abnormal returns must not be considered significant anymore at any reasonable level of statistical significance. This may be avoided by reducing the significance levels in the individual tests, but the required reduction is likely to be so large that the tests will probably be rendered insufficiently powerful to still detect any treatment effect. Given that neither Ou and Penman nor Holthausen and Larcker acknowledge that a portion of their initial set of accounting variables is bound to show statistically significant association with subsequent abnormal returns by chance, their results may very well be spurious and sample-specific. In fact, the instability from one sub-period to another in terms of the variables that are ultimately taken up in the computation of

Ou and Penman's Pr-measure, for example, is a quite painful illustration of the impact of random effects on their findings. Furthermore, I do not know of any empirical evidence that has corroborated the initial results underlying either of the two anomalies. Therefore, I feel that their extreme proneness to the threat of fishing makes that both the Ou and Penman anomaly and the Holthausen and Larcker anomaly are definitely among the weaker challenges to market efficiency with respect to accounting information.

Having discussed how the specific set-up of a study may seriously aggravate the fishing threat, I now have to look into the possible ways to protect one's research design from it. Clearly, a very effective potential protection lies in the ex ante development of an alternative security pricing theory. If empirical results are found to be inconsistent with the EMH's predictions, but consistent with those generated by a worthy theoretical alternative, the probability of them being obtained by chance may be seriously reduced. One must not be naïve, though; specifying a general theory of market inefficiency is far from evident. Consequently, looking for solutions that do not rely on the specification of a comprehensive alternative pricing theory shows perhaps more sense of reality. As such, I feel that the threat of fishing can also be mitigated considerably by focusing on a small number of potentially predictive variables that are defined beforehand.

Admittedly, this approach does obviously not offer watertight protection to the fishing threat. Researchers who are not too particular when it comes to ethics may always start out with a large set of different information items, while reporting only about the ones that appear to "work" ex post. These considerations make that I am inclined to conclude that the proneness of a study's findings to the threat of fishing might best be assessed on the basis of the study's performance with respect to another type of validity, namely external validity. After all, to the extent that empirical results may be replicated for different research settings and time periods, they are less likely to be obtained by chance. Therefore, I recommend that the fishing threat be avoided by designing a research set-up in such a manner that sufficient attention is given to the external validity of the empirical evidence that the set-up concerned will yield. How this may be achieved in practice, will be discussed in detail in section seven.

5. INTERNAL VALIDITY

As mentioned earlier, internal validity is concerned with whether the existence of a causal relationship between the treatment and outcome constructs may be inferred from their statistical covariation. In other words, it raises the issue if such statistical covariation should be interpreted as evidence of the treatment causing the outcome or whether it actually stems from the impact of moderating variables. It is important to acknowledge the particular nature of accounting-based EMH tests in this respect. Basically, the accounting variable(s) serving as treatment construct cannot be considered to be the actual cause of any abnormal returns occurring, for example, several months later. It is, after all, extremely unlikely that financial statement items constitute the immediate reason for abnormal returns that occur so long after those financial statement items have been disclosed. This would in fact require that the accounting numbers lead to direct revisions in market expectations long after their disclosure date. Although this possibility cannot be ruled out with certainty, it is in my opinion much more logical for abnormal returns to be brought about by more contemporaneous events, disclosures or information items. Therefore, it is my view that any systematic relationship between accounting variables and medium- to long-term subsequent abnormal returns is naturally indirect and stemming from investors' failure to fully appreciate those accounting figures' usefulness with respect to predicting the realization of some future value driver. As other information about the value driver is released or when it is actually realized, the market will adjust its expectations and (partly) predictable abnormal returns will occur in that case.

It is important to emphasize that the almost necessarily indirect nature of the relationship between accounting figures and subsequent abnormal returns does not damage the usefulness of examining its existence: the EMH simply predicts that *no* such relationship exists, neither direct nor indirect. Any *valid* evidence to the contrary therefore constitutes a rejection of market efficiency. Furthermore, within the Cook and Campbell (1979) framework, the indirectness of any such causality

does not in itself inhibit that it is investigated by means of what I have labelled direct tests. In fact, direct and indirect tests may be regarded as examinations of similar underlying causal laws, whereby the latter are simply specified at a more micromediational level in the case of indirect tests. Given that, as indicated in section three, Cook and Campbell acknowledge the usefulness of molar causal laws, the use of direct (and hence more molar) tests does not pose any particular problems within their framework. I feel it is abundantly clear, though, that with a view to investigating an indirect causality, a research design's internal validity benefits considerably if the tests themselves are carried out in an indirect manner. Stated differently, it is my opinion that the formulation of a meaningful alternative theory for the EMH requires the specification of some intermediary event or disclosure that elicits the predictable abnormal returns and as such links the latter with the accounting information with respect to which market efficiency is being examined. In turn, a proper evaluation of the descriptiveness of such an alternative theory in my view asks for an indirect testing procedure that allows investigation of the supposedly predictable abnormal returns' relation to predictable "surprises" (that is, from the market's point of view) with respect to the intermediary event. Thus, although direct accounting-based tests of the EMH should not be deemed completely useless, the superiority of the indirect approach is in my opinion quite evident. As a matter of fact, indirect tests constitute probably the only sensible approach if one is to gain a true understanding of any potential inefficiencies.

5.1. SELECTION

As far as the specific threats identified by Cook and Campbell (1979) are concerned, the most important one is probably that of what they call **selection**. This threat refers to the possibility that outcomes may appear to be treatment-initiated, while they actually originate from initial differences among the research units. Cook and Campbell suggest that *randomisation*[16] is the most efficient method to deal with this problem. Unfortunately, randomisation clearly requires an experimental research

[16] In the case of a randomised research design, all research units have equal probabilities of receiving a particular treatment. Cook and Campbell (1979) describe the consequences of such a procedure as follows (p. 5): *"Given a sufficient number of units relative to the variability between units, the random selection procedure will make the average unit in any one treatment group comparable to the average unit in any other treatment group before the treatments are applied. (...). Random assignment is the great ceteris paribus – that is, other things being equal – of causal inference".*

set-up and is, therefore, infeasible in the context of accounting-based EMH tests which necessarily rely on passive observation.

In this particular context, the threat of selection pertains to the issue that systematic abnormal returns apparently accruing to accounting-based investment strategies may in fact stem from other differences between firms for which the accounting variable acts as a proxy. Basically, such a research result would of course still constitute falsifying evidence for the EMH. Two extremely important considerations are, however, in order. Firstly, it is crucial to see that accounting-based evidence inconsistent with the theory of market efficiency need not be indicative of investors actually failing to fully anticipate the implications of accounting information. Such a far-reaching conclusion cannot be drawn solely on the basis of statistical covariation between an accounting variable and subsequent abnormal returns. This brings me back to the point made in the preceding paragraph. If one is to draw inferences about potential flaws in the way in which investors process accounting information, it is vital to examine whether the abnormal returns "earned" are related to biases in the market's implicit assessments of the implications of that accounting information. Again, I plead in defence of the indirect testing approach. Secondly, it has to be taken into consideration that the contention that apparently abnormally profitable accounting-based investment strategies constitute evidence of market inefficiency is critically contingent upon abnormal returns being measured appropriately. If this is not the case, the supposedly "abnormal" returns may in reality just be fair compensations for some risk factor that has not been adequately controlled for. As a consequence of the generally alarmingly low degree of (effect) construct validity[17] of research designs in this area, this condition is far from trivial. In fact, it is the source of the already often cited joint-hypothesis issue that has been prevalent throughout the entire empirical EMH literature[18]. The pervasiveness of the joint-hypothesis problem is an indication of the impossibility to satisfactorily eliminate it. It is my opinion, however, that certain measures can be taken to at least mitigate the threat of risk-related selection. Since the problem is ultimately a matter of construct validity, I shall discuss those measures in the following section.

---

[17] See section six.
[18] See chapter 2.

It is important, though, to acknowledge how internal validity and construct validity are interwoven in this respect: the almost inherently low degree of construct validity creates the possibility that the variables used to measure abnormal returns are partly risk-determined. In turn, this possibility increases the likelihood that any statistical covariation between the accounting variable and the alleged abnormal returns does not evidence a causal relationship between these two concepts, but simply arises as a consequence of the accounting variable acting as a risk proxy. As such, the generally doubtful construct validity of empirical market efficiency studies at the same time makes them prone to the threat of selection which is harmful to their internal validity. Thus, the importance of carefully seeing to the construct validity of accounting-based tests of the EMH is doubled in a way. Therefore, I shall dedicate quite some attention to that matter in the following section.

5.2. MORTALITY

Before doing so, I would like to highlight one more threat to internal validity identified by Cook and Campbell (1979), namely **mortality**. This threat originates from the possibility that outcomes appearing as being treatment-initiated may in reality be induced by differential dropout rates in the various treatment groups. The bias that may be introduced in this manner has been labelled *survivorship bias*. Clearly, the potential relevance of this bias for empirical EMH research can hardly be overestimated. In the preceding paragraph, it was already described how the problems involved in measuring abnormal returns may mistakenly lead a researcher to the conclusion that a particular investment strategy yields extraordinary profits. However, survivorship bias makes that similarly wrong inferences can be drawn even if abnormal returns are measured appropriately.

Suppose for example that a certain accounting variable is positively related to equity risk. If that variable is then used to select stocks, it is very conceivable that relatively more of the selected firms as opposed to the rest of the market will subsequently delist or perhaps even go bankrupt. If in that case a researcher uses a sample composed only of firms that are still listed at the end of the research

period, he will obviously seriously overestimate the feasible profitability of the investment strategy under consideration. It is therefore crucial to construct one's sample solely on the basis of information that was available at the beginning of the research period[19]. Unfortunately, this is easier said than done.

After all, with a view to having access to large amounts of easily treatable data, most academics who are active in this field of interest have made use of commercial databases. In itself, using such databases need not be a problem of course. It is crucial, though, to always recognize their limitations and shortcomings. Apparently, far from everyone seems to acknowledge that the backfilling procedures adopted by most database vendors fail to safeguard against the threat of mortality. On the contrary, it is quite conceivable that such procedures actually introduce additional biases. Clearly, it is extremely difficult to assess the extent to which the various accounting-based anomalies that have been revealed are attributable to survivorship bias. In light of the almost universal use of the combined CRSP[20] and Compustat files, encouraging evidence has been presented by Chan, Jegadeesh and Lakonishok (1995), though. In response to claims of the anomalous results being spurious brought forward by Kothari, Shanken and Sloan (1995), Chan, Jegadeesh and Lakonishok have demonstrated that neither the backfilling procedure used in the development of the Compustat database, nor the matching of the CRSP and Compustat files create a severe survivorship bias. Furthermore, as far as the particular case of the value-versus-growth anomaly is concerned, Chan, Jegadeesh and Lakonishok have shown that the anomaly did not disappear using a sample of large firms free from survivorship bias.

Encouraging as these findings may be, they are not conclusive. In fact, one can only really protect a research design against the mortality threat by either hand-collecting data from the past for firms omitted from the database being used or by implementing the investment strategy concerned to a sample of currently listed

---

[19] The aim of requiring that a firm's probability of being taken up in the sample should be independent of any information that was not available until during or after the research period is to avoid so-called hindsight bias. A meaningful test of market efficiency (i.e. do stock prices fully reflect all available information at a certain point in time?) should obviously only make use of information that was indeed available. Since survivorship bias is caused by the implicit use of more recent information about subsequent delistings, it can be regarded as a special case of hindsight bias.

[20] The Center for Research in Security Prices at the University of Chicago.

firms and awaiting its future performance. Obviously, the time-consuming nature of these approaches is likely to be prohibitive in practice. Therefore, I am inclined to recommend a second best solution if one is uncertain about the degree of survivorship bias that is present in a particular database. This solution consists of using a contemporaneous version of the database for each moment in the past at which the investment strategy under consideration is supposed to have been implemented. For example, if one is interested in a strategy's mean twelve-month performance during the 1995-2000 period, the procedure would consist of using six different versions of the database, one for each year in the overall research period. The threat of survivorship bias is completely ruled out in that manner. As in most cases, there is, however, another side to the coin. Apart from the question whether such "older" database versions are actually available to the researcher, it is important to see that they are likely to offer a coverage of firms that is by far not as extensive as more recent versions. This then raises the issue of whether the researcher is prevented from constructing a sample that may be deemed representative of the target population. If so, the research design's *external* validity[21] may be seriously threatened. As a consequence, some might argue that this procedure merely shifts the problem without resolving it. I beg to differ for at least two reasons. First of all, I concur with Cook and Campbell when they claim that internal validity ought to be given priority over external validity. Although the latter's importance should not be underestimated, one cannot pass over the fact that it is merely concerned with the generalizability of results. Internal validity, on the other hand, relates directly to the obtained results themselves. As such, it is generally taken to be situated at the core of a research design's validity. Consequently, the conversion of a potentially substantial threat to internal validity into a possible threat to external validity may still be considered advantageous. Secondly, the procedure involving the use of several database versions need not be adopted independently. In fact, it is perhaps more appropriate to use this procedure as a sensitivity check with respect to the findings obtained through the use of a backfilled recent database version. Thus, the researcher may be enabled to infer to which extent results are driven by survivorship bias. Both internal and external validity can be preserved in this manner.

---

[21] See section seven.

6. CONSTRUCT VALIDITY

As I have indicated in section three, the general impossibility to observe the theoretical cause and effect concepts used in the formulation of a hypothesis makes that empirical researchers have to rely on constructs to make those concepts operational. Construct validity is thereby concerned with the extent to which the chosen treatment and outcome constructs may be considered adequate operationalizations of the theoretical concepts they are supposed to represent. In many instances, researchers will have an array of potential measures to choose from. Clearly, the aim is then to select those constructs that neither fail to incorporate all the dimensions of the target concepts (i.e. construct underrepresentation), nor contain dimensions that are irrelevant to the target concepts (i.e. surplus construct irrelevancies). It will become clear in the remainder of this section that I concur with Cook and Campbell (1979) when they assert that assessment of a research design's construct validity often largely comes down to examination of the convergence across different possible measures of the same underlying concepts. I shall also argue, however, that simply investigating the convergence across operationalizations may not suffice in the specific case of empirical EMH tests. The latter's inherent proneness to threats to construct validity is so large that I find it almost imperative to take additional steps with a view to safeguarding them against the potential impact of those threats on the research results.

6.1. INADEQUATE PREOPERATIONAL EXPLICATION OF CONSTRUCTS

A first threat identified by Cook and Campbell (1979) lies in the **inadequate preoperational explication of constructs**. In order to be able to select appropriate measures it is necessary to specify beforehand which requirements the constructs have to meet. In turn, this requires detailed delineation of the theoretical concepts that need to be operationalized. As I have already described in section three, in market efficiency tests with respect to accounting information, the latter constitutes *the concept of cause*. Although the focus on accounting information entails a

considerable limitation compared to the concept of "all available information" that is found in the general definition of the EMH, it is abundantly clear that one is still dealing with a very broad concept. Even if interim reports are ignored and the concept of accounting information is restricted to annual accounts data, one remains faced with a sheer inexhaustible array of potentially predictive variables to draw upon. In principle, any accounting item can be used either individually or in combination with other items as an operationalization of the accounting information concept. I feel, though, that the importance of the notion of "value-relevant" information should not be overlooked.

Not all financial statement items are equally relevant as far as equity valuation is concerned. As a matter of fact, a large number of accounting items and ratios is likely to have little to no value relevance. Therefore, I consider it merely logical to use only treatment constructs for which the information content can be specified in advance. This specification is to be interpreted as ex ante explication of the clues that the treatment constructs concerned provide about subsequent realisations of value drivers. Apart from the fact that it is simply not meaningful to test market efficiency with respect to variables that do not contain any value-relevant information to begin with, the use of treatment constructs for which the usefulness in the context of valuation has been established ensures a close connection with the concept of accounting *information* in that particular context. In addition, to the extent that the prior explication of the relevance of the treatment operations is not entirely empiricism-driven, the requirement to specify treatments' information content beforehand provides a natural selection among potentially predictive variables that may be considered very helpful to protect a research design from the fishing threat to statistical conclusion validity discussed in section 4.3. Note that the a priori specification of treatment constructs' informativeness with respect to subsequent value driver realisations is in fact only compatible with the indirect approach to testing market efficiency.

The effect concept that has to be operationalized in accounting-based EMH tests is that of "abnormal returns". Defining the theoretical notion is child's play: an abnormal return is nothing but the difference between a security's actual return during a given period and its normal return for that period. The concept's practical

complexity in terms of the development of a suitable construct is hard to overestimate, though. In fact, the troublesome measurement of abnormal returns is at the root of the issue that has been plaguing EMH research for several decades now, namely the joint-hypothesis problem. Clearly, the joint-hypothesis problem results from the need to operationalize *normal* returns. I shall discuss this tricky matter directly. Before doing so, there is another issue that deserves attention in the process of deciding which abnormal return measure to use. It is necessary to determine one's investment horizon as well as the features of the investment strategy that one is (implicitly) adopting, as these elements should be reflected in the outcome construct used. Quite crucial in this respect is the distinction between CAR's and BHAR's that was discussed in section 4.2. As argued by Ritter (1991), CAR's and BHAR's are intrinsically relatively disparate constructs in that they can be used to answer different questions. Barber and Lyon (1997) neatly summarize Ritter's point as follows (p. 344):

> *"Consider the case of a 12-month CAR and an annual BHAR. Dividing the 12-month CAR by 12 yields a mean monthly abnormal return. Thus, a test of the null hypothesis that the 12-month CAR is zero is equivalent to a test of the null hypothesis that the mean monthly abnormal return of sample firms during the event year is equal to zero; it is not a test of the null hypothesis that the mean annual abnormal return is equal to zero. To test the latter hypothesis, a researcher needs to use the annual BHAR."*

For researchers interested in the longer-term performance of stocks, BHAR's offer an additional advantage. Contrary to CAR's, they do not imply rebalancing after each sub-period. As such, BHAR's are associated with somewhat more realistic investment strategies that entail a considerably lower number of transactions. The potentially important negative impact of transaction costs on paper returns is seriously mitigated in that manner. Thus, at first sight, the dominance of BHAR's as far as the measurement of long-term security performance is concerned seems pretty straightforward. However, BHAR's too have been claimed to suffer from a few possibly significant drawbacks that may challenge this dominance.

First, Fama (1998) has contended that the justification for using particular asset pricing models in tests of market efficiency comes at least partly from the results obtained in empirical tests of the descriptive validity of those models. Researchers have, however, typically used shorter-term (i.e. monthly) return data in such asset pricing model tests. In fact, there is hardly any empirical evidence to support the use of asset pricing models in long-run return-based examinations of the EMH. Therefore, Fama recommends using CAR's, as they are necessarily calculated on the basis of (accumulated) short-term returns. Although Fama's initial point regarding the lack of empirical evidence with respect to the descriptive validity of asset pricing models in a long-term context is correct, it does not strike me as a particularly solid argument against using BHAR's in tests of market efficiency. After all, this counterargument only applies if one defines a BHAR as the difference between an actual return and some *direct* estimate of a long-run "normal" buy-and-hold return. Thus, simply using the specification given in equation 2 circumvents this issue. Second, BHAR's typically have distributional properties that make them somewhat less manageable than CAR's as far as the assessment of their statistical significance is concerned. BHAR's are clearly more prone to the skewness bias discussed in section 4.2., for example. And finally and perhaps most importantly, there may be a major practical impediment to the use of long-term BHAR's in tests of market efficiency. Some quite appealing abnormal return constructs (e.g. Jensen's alphas) are derived from time-series regressions in which raw or excess[22] security returns pertaining to the test period are used as dependent variable values. A logical consequence of focusing on returns that are associated with a longer time period (e.g. annual buy-and-hold returns) is that the number of time-series observations available for a given sample period is reduced. In turn, this reduction can make it very difficult, if not impossible, to obtain reliable estimates for the kind of abnormal return measures I have just described.

In conclusion, I feel that the superiority of BHAR's as a concept is patently obvious when it comes to assessing the long-term performance of investment strategies. In cases where it is possible, I would therefore definitely recommend that BHAR's be

[22] An excess security return is equal to the difference between the raw return of that security and the return on a risk-free asset.

used in long-horizon tests of market efficiency. After all, it is not too difficult to sidestep Fama's argument concerning the lack of empirical evidence to justify the use of certain asset pricing models in long-term examinations of the EMH, while there are possibilities to considerably mitigate the statistical difficulties associated with BHAR's[23]. Nevertheless, one should not ignore the problematic practicability of calculating BHAR's for those expected return benchmarks whereby abnormal return estimates are supposed to emerge directly from regressions based on test period data. One way to avoid these estimation problems lies in using the modified BHAR measure suggested by Kothari and Warner (1997). Using the same symbols as in equations 1 and 2, Kothari and Warner propose the following accumulation method to estimate a security's long-term abnormal performance[24]:

$$COMPAR_{i\tau} = \prod_{t=1}^{\tau}(1 + AR_{it}) - 1, \qquad (4)$$

where $COMPAR_{i\tau}$ represents the compounded abnormal return of security i over a τ-month period. I deliberately do not use the term "buy-and-hold" in the labelling of this abnormal return measure. First of all, for the practical purpose of being able to distinguish the measure defined in equation 4 from the one defined in equation 2. And secondly, for the more fundamental reason that, as I shall explain directly, the formula in equation 4 does not correspond with a pure buy-and-hold investment strategy. Basically, COMPAR's combine features of both CAR's and BHAR's. Similar to CAR's, they have the advantage of being based on direct estimates of sub-period abnormal returns instead of on separate estimations of sub-period actual and normal returns. The compounding, on the other hand, makes that COMPAR's are quite closely connected to BHAR's. With respect to the calculation of abnormal return constructs in the spirit of Jensen's alphas, COMPAR's thus basically unite the strong points of the two basic return accumulation methods. This should not make one blind, though, to the somewhat awkward investment strategy that is implied by the computation of COMPAR's. Contrary to the strategy underlying BHAR's and analogous to that underlying CAR's, COMPAR's measure the long-term return to an investment strategy that requires rebalancing. However, whereas the rebalancing after each sub-period is complete in the case of CAR's (i.e. *all* returns earned during a sub-period are

[23] See section 4.2.

removed from the investment position), the strategy implied by COMPAR's does not fully restore the situation of the beginning of the sub-period. After all, the compounding of the abnormal returns implies that after each sub-period only the *normal* returns are removed from the investment position, while the remainder is assumed to be reinvested. Due to this rebalancing implicit in COMPAR's, I feel that they are intrinsically slightly inferior to BHAR's as far as the measurement of stocks' long-term abnormal performance is concerned, albeit that COMPAR's certainly constitute a worthy alternative for BHAR's when the latter are incompatible with the selected expected return benchmark.

Although the importance of the distinction between BHAR's and CAR's should definitely not be underestimated, one cannot pass over the fact that the matter is ultimately rather trivial when compared to the other problem that has to be dealt with in operationalizing the concept of abnormal returns. This "other problem" is of course the joint-hypothesis issue. As stated before, a security's abnormal return simply equals the difference between its actual return during a particular period and its normal return for that period. Abstracting from the choice that has to be made with respect to the way in which returns are accumulated, actual returns are of course fairly easy to determine ex post. Naturally, the real challenge (or should I say pitfall) lies in the estimation of stocks' normal returns, i.e. in the development of appropriate expected return benchmarks. Given that the benchmark selection issue is at the heart of the ongoing and apparently dead-end debate between advocates and opponents of the EMH, it is basically impossible to give it too much attention in a methodology-oriented discussion of market efficiency research.

Now how should one begin to tackle the issue of estimating normal security returns? First of all, it is crucial to come to a clear understanding of the theoretical concept of normal or expected returns. If one's aim is to identify possible anomalies in the cross-section of average returns, it is important that normal return estimates be derived from a formal asset pricing model that is consistent with the conditions for market equilibrium. This implies that, as Fama (1998) emphasizes, a purely stochastic model like the market model is in fact fairly useless in this respect.

---

[24] See also footnote 7.

The most basic equilibrium condition may formally be described as follows[25]:

$$P_{j,t} = \mathrm{E_t}\left[a \frac{u'(c_{t+1})}{u'(c_t)} x_{j,t+1} \right], \tag{5}$$

where $P_{j,t}$ represents the equilibrium price of asset j at time t, $\mathrm{E_t}$ denotes expectation at time t, a is the subjective discount factor, $u'(c_t)$ is the marginal utility of consumption at time t and $x_{j,t+1}$ is the asset's payoff at time $t+1$. The intuition behind equation 5 may be clarified by the following rearrangement:

$$P_{j,t} \cdot u'(c_t) = \mathrm{E_t}\left[a u'(c_{t+1}) x_{j,t+1} \right] \tag{6}$$

Equation 6 clearly demonstrates the logic underlying the equilibrium condition described by equation 5. Specifically, it shows how the basic pricing equation (i.e. equation 5) follows directly from the equality of the loss in utility associated with the purchase of an additional unit of asset j and the accompanying decrease in consumption at time t on the one hand (i.e. the left-hand side of equation 6) and the expected discounted gain in utility associated with the extra payoff and consumption at time $t+1$ on the other hand (i.e. the right-hand side of equation 6). Evidently, such an equality between marginal loss and marginal gain is required to reach a state of equilibrium.

Expression 5 may be rewritten more simply as follows:

$$P_{j,t} = \mathrm{E_t}\left[m_{t+1} \cdot x_{j,t+1} \right], \tag{7}$$

whereby $m_{t+1} \equiv a \frac{u'(c_{t+1})}{u'(c_t)}$ is called the *stochastic discount factor* or *marginal rate of substitution*. Equation 7 provides a very elegant and extremely general representation of how future asset payoffs map into current asset prices. It is obvious, though, that this representation has no immediate testable implications. If one aims at describing the cross-section of average returns, additional assumptions are clearly called for.

Before doing so, it is useful to make the transitions from a price-oriented specification to a return-oriented specification and from the discount factor

---

[25] For reasons of parsimony, equation 5 describes the first-order condition in the two-period case. For the relatively straightforward extension to the multi-period case, see, for example, Cochrane (2001).

representation to the more traditional beta representation. These transitions can be achieved without any difficulties, nor do they entail a loss in generality. Given that asset j's return between time t and time $t+1$, i.e. $R_{j,t+1}$, is equal to $x_{j,t+1}/P_{j,t}$, equation 7 implies

$$1 = \mathrm{E_t}(m_{t+1} \cdot R_{j,t+1})$$

Since $\mathrm{E_t}(m_{t+1} \cdot R_{j,t+1}) = \mathrm{E_t}(m_{t+1}) \cdot \mathrm{E_t}(R_{j,t+1}) + \mathrm{cov}(m_{t+1}, R_{j,t+1})$, one obtains

$$\mathrm{E_t}(R_{j,t+1}) = \frac{1}{\mathrm{E_t}(m_{t+1})} - \frac{\mathrm{cov}(m_{t+1}, R_{j,t+1})}{\mathrm{E_t}(m_{t+1})}$$

Multiplying and dividing the second term on the right-hand side by $\mathrm{var}(m_{t+1})$ gives

$$\mathrm{E_t}(R_{j,t+1}) = \frac{1}{\mathrm{E_t}(m_{t+1})} + \left[\frac{\mathrm{cov}(m_{t+1}, R_{j,t+1})}{\mathrm{var}(m_{t+1})}\right] \cdot \left[-\frac{\mathrm{var}(m_{t+1})}{\mathrm{E_t}(m_{t+1})}\right]$$

Defining γ as $\frac{1}{\mathrm{E_t}(m_{t+1})}$, $\beta_{j,m}$ as $\frac{\mathrm{cov}(m_{t+1}, R_{j,t+1})}{\mathrm{var}(m_{t+1})}$ and λ_m as $\left(-\frac{\mathrm{var}(m_{t+1})}{\mathrm{E_t}(m_{t+1})}\right)$, leads to

$$\mathrm{E_t}(R_{j,t+1}) = \gamma + \beta_{j,m} \cdot \lambda_m \qquad (8)$$

Equation 8 is the common single-beta representation of expected returns. Thus, the basic pricing equation 7 is equivalent to stating that mean asset returns are linear in the regression betas of those asset returns on the stochastic discount factor. While the latter conjecture may *appear* somewhat more workable as far as describing the cross-section of average returns is concerned, it does of course not have any immediate practical relevance either. Concretely, it is impossible to derive any empirically testable implications without a more detailed specification of the stochastic discount factor. In turn, such specification requires extra assumptions that place additional structure on investors' consumption-portfolio decisions. It is the variation in those extra assumptions that has led to the parallel existence of multiple asset pricing models.

Roughly speaking, four types of asset pricing models can be discerned that have commonly been used in tests of market efficiency:

1. Consumption-based asset pricing models (CBAPM)
2. Capital asset pricing models (CAPM)
3. Intertemporal capital asset pricing models (ICAPM)

4. Arbitrage pricing models (APM)

As emphasized by Cochrane (2001), it is important to bear in mind that these four types of models are ultimately all derived from the consumption-based equilibrium condition described by equation 5. Consequently, they may all be regarded as different specifications of the same basic theory. This does not keep them from generating quite diverse empirical predictions in some cases, though; a diversity that is obviously brought on by the variation in underlying assumptions as well as by the need to rely on proxy measures in implementing the models; a diversity, also, that renders the choice of expected return model often everything but trivial.

As far as the decision on which model to use in EMH tests is concerned, one particular consideration is in my opinion absolutely essential. In his review article of the empirical literature on market efficiency, Fama (1991) gives the impression of considering the various types of asset pricing models more or less equally useful to examine the EMH. Personally, though, I fully agree with Loughran and Ritter (2000) who argue that meaningful empirical testing of the EMH requires the use of a *normative* expected return model, i.e. a model that has been developed using deductive reasoning. After all, the EMH itself is a normative hypothesis: it predicts consequences of rational aggregate market behaviour. Admittedly, in view of the fact that the EMH may only be tested jointly with some expected return model, one necessarily has to rely on the assumption that the model used is valid in order to be able to make any statements regarding the degree of market efficiency. It is crucial to see, however, that the descriptive validity of the expected return model ought to be assessed from a normative point of view and not from a positive point of view. In other words, the model should provide an appropriate description of what the cross-section of expected stock returns *should* look like in equilibrium and not of what it actually looks like in reality. Therefore, so-called "horse race" comparisons of asset pricing models must never guide the model choice in empirical EMH research.

What should guide the choice then? Assessing the normative validity of a model is obviously a lot easier said than done. By definition, an asset pricing model's degree of normative validity cannot be examined empirically without assuming that markets are efficient: clearly, we are at the core of the joint-hypothesis problem!

Due to the latter, one has no other choice but to assess models' normative validities on the basis of somewhat more subjective considerations. Quite useful in this respect is in my opinion what I would describe as a model's *theoretical value*, i.e. the combination of the extent to which a model follows naturally from the basic equilibrium condition and the restrictiveness of its underlying assumptions. It should not be forgotten, however, that the ultimate aim is to *apply* a particular asset pricing model with a view to estimating abnormal returns (indirectly, that is). As a result, a second decision-relevant criterion should be considered, namely a model's *practical usefulness*. Concretely, the latter refers to the extent to which the available data suffice to yield satisfying measures of the model's variables so as to make it possible to appropriately estimate its parameter values. Similar to the theoretical value criterion, assessment of a model's practical usefulness is fairly subjective. In addition, it is likely to be conditional upon the specific research context. Consequently, the choice to use a certain asset pricing model in a test of market efficiency is in the end a largely subjective matter. Nevertheless, as far as the usefulness of the four general types of models enumerated above is concerned, I feel that it is possible to make at least a few objective comments.

First of all, there appears to be a trade-off between a model's theoretical value and its practical usefulness. As such, CBAPM (e.g. Rubinstein (1976), Lucas (1978) and Breeden (1979)) are at one end of the spectre: closely connected to the basic equilibrium condition as they are, they clearly dominate the alternative models from a theoretical point of view. Implementing them adequately has proved to be quite troublesome, though. The main object of criticism in this respect has been the inevitable use of measures of *aggregate* consumption. It is indeed questionable whether such aggregates are representative of the consumption paths of the individual agents who determine asset prices. The existence of heterogeneous constraints (e.g. Campbell and Mankiw (1989), Mankiw and Zeldes (1991) and Brav, Constantinides and Geczy (1999)) and the presence of costs of consumption adjustment (e.g. Grossman and Laroque (1990) and Marshall and Parekh (1999)) are just two examples of potential reasons as to why aggregation might fail. In spite of Campbell and Cochrane's (1999 and 2000) extremely interesting contention that standard CBAPM may be seriously improved upon by introducing habit formation

into investors' utility functions, I feel that the aggregation issue remains a largely insurmountable impediment to the practical application of CBAPM.

At the other end of the spectre are APM (e.g. Ross (1976) and Fama and French (1993)). Given that the latter derive predictions about expected asset returns from analyses of the factor structure of realised returns, their practical usefulness is beyond dispute. However, the positive nature of APM makes that they have virtually no theoretical value whatsoever. The factors included in APM are selected inductively and on the basis of empirical grounds. APM are in fact designed to capture the existing patterns in asset returns. As such, they may very well include factors that are related to market inefficiencies rather than to priced risk. Stated otherwise, to the extent that the uncovered empirical patterns in stock returns are persistent, APM are basically biased toward being descriptive of reality. Therefore, I totally concur with Loughran and Ritter's (2000) contention that this feature makes APM of little to no use in tests of market efficiency. It should be stressed that this fundamental criticism also pertains to the renowned Fama and French three-factor model, for example. After all, Fama and French's model is just as much a positive model that treats patterns that should still be regarded as being anomalous as if they have been incorporated in a fully fledged paradigm of the security price formation process whereby the latter is implicitly considered to be consistent with market efficiency. As a consequence, I feel that in the context of EMH research a model like the Fama and French three-factor model only has its place in additional analyses in that it can be used to verify whether a particular regularity in security returns is similar to other known patterns. Nonetheless, it has become almost standard practice in recent years (e.g. Sloan (1996), Abarbanell and Bushee (1998), Ali, Hwang and Trombley (2000), Brown and Han (2000), Collins and Hribar (2000) and Thomas (2000)) to conduct market efficiency tests using fairly arbitrary expected return models that adjust for anomalies like the size effect and the B/M effect. Totally unjustly, in my opinion: as Loughran and Ritter indicate, one does then not examine whether markets are efficient, but whether it is possible to identify regularities that are distinct from known anomalous patterns in security returns. At best, this is a research question that deserves some attention in the second stage of an empirical study of the EMH. As far as the basic test of market efficiency itself is concerned, I repeat the point made earlier: the hypothesis of

efficient markets *must* be examined in conjunction with a *normative* expected return model.

This leaves CAPM and ICAPM. Judging by theoretical value and practical usefulness, these two types of models are situated between CBAPM and APM. They are theoretically slightly inferior to CBAPM in that they rely on additional assumptions that make it possible to specify the stochastic discount factor in function of non-consumption data. At the same time, the absence of the need for consumption data obviously avoids the potential difficulties relating to the use of aggregate consumption measures. On the other hand, the deductive reasoning underlying both the CAPM and ICAPM makes that they are markedly the better of APM from a theoretical point of view. Finding appropriate proxies for the factors taken up in the models will, however, generally prove harder for CAPM and ICAPM than for APM. Thus, CAPM and ICAPM may in fact be perceived as models that offer a compromise between the more extreme CBAPM and APM. Contrary to the latter two types of models, CAPM and ICAPM need not be ruled out automatically due to either lack of practical usefulness or lack of theoretical value. As such, CAPM and ICAPM sort of provide the best of two worlds.

So, which type of model should ultimately be used: a CAPM or an ICAPM? In my opinion, this question cannot be answered in a general fashion. Using the two criteria of evaluation discussed above, neither of the two model types strictly dominates the other. Stated differently, the trade-off between theoretical value and practical usefulness continues to apply. The CAPM is theoretically neater as it specifies the factor that the stochastic discount factor is supposed to be a function of, namely the return on total wealth. According to the ICAPM, the marginal rate of substitution is not only determined by the return on total wealth, but also by a number of additional state variables that proxy for shifts in the set of investment opportunities. The theory underlying the ICAPM remains silent, though, about the precise nature of those state variables, thus giving the implementation of the model somewhat of an ad hoc touch[26]. On the other hand, the ICAPM may have a slight

[26] Admittedly, work along the lines of Lettau and Ludvigson (2001), who attempt to provide some theoretical rationalization for selecting the additional state variables, may be considered quite promising to help reduce in the near future the importance of the ad hoc aspects currently involved in empirical implementations of the ICAPM.

edge over the CAPM in terms of practical usefulness. The single-factor nature of the CAPM makes that it is quite vulnerable to measurement error in that particular factor. Since it is a well-established fact that measuring the return on total wealth involves a substantial number of difficulties, this disadvantage of the CAPM should not be underestimated. Some authors (e.g. Roll (1977)) have even claimed that it renders the CAPM empirically non-testable and so fairly useless in practice. Although I feel that this claim is somewhat exaggerated, I do believe that extension of the proxy measure for return on total wealth beyond the commonly used return on a broad stock market index is definitely worth considering. Jagannathan and Wang's (1996) inclusion of a proxy measure for the return on human capital is a most interesting example of such an extension. The main point here, though, is that the additional state variables taken up in an ICAPM may be regarded or at least serve as a natural extension of the proxy used for the return on total wealth. Stated otherwise, the additional variables may capture some of the measurement error in the return on total wealth. Thus, the predictions of the CAPM as it is typically implemented may sometimes be improved upon by using an ICAPM instead.

In conclusion, I think it is fair to say that the most important element to take home from the above discussion is the need to use normative expected return models in tests of market efficiency. On the basis thereof APM, including the Fama and French three-factor model as well as all models containing arbitrary adjustments for size and B/M effects, were in fact ruled out. CBAPM, on their part, were said to be very appealing from a theoretical point of view, but their practical applicability was claimed to be seriously hampered by the general lack of non-aggregated consumption data. Thus, the need arises to rely on models that deductively substitute consumption out of the specification of the stochastic discount factor. The CAPM and the ICAPM are model types that meet this need. Therefore, I feel that using either of these two types of models in empirical EMH research is definitely justifiable. Whether one ultimately chooses to rely on a CAPM or an ICAPM is a largely subjective decision, however, which is bound to be dependent upon the specific context as well. It should be stressed, though, that obviously nothing precludes researchers from applying both types of models and examining the sensitivity of their findings to the operationalization of the effect construct.

6.2. MONO-OPERATION BIAS

This brings me immediately to a second potential threat to construct validity, namely **mono-operation bias**. With respect to this threat, Cook and Campbell (1979) state the following (p. 65):

> *"Since single operations both underrepresent constructs and contain irrelevancies, construct validity will be lower in single exemplar research than in research where each construct is multiply operationalized in order to triangulate on the referent. There is rarely an adequate excuse for single operations of effect constructs, since it is not costly to gather additional data from alternative measures of the targets."*

Basically, I totally agree with Cook and Campbell's (1979) contentions. However, as far as market efficiency research is concerned, I am prepared to go even further and claim that there is *never* an adequate excuse for relying on just one abnormal return construct. After all, it should be abundantly clear by now from the many references to the joint-hypothesis problem earlier in this paper as well as from the more detailed discussion in the preceding subsection that effect construct validity is in fact the weak point of empirical EMH research in general. Remember in particular how I argued in the previous paragraph that the selection of the expected return model ultimately remains a largely subjective matter. Clearly, subjective considerations are not ideal counsellors with a view to obtaining scientifically sound results. As a consequence, I feel that researchers should aim to enhance the (effect) construct validity of their research designs by keeping the degree of subjectivity involved in their findings as low as possible. Naturally, the most obvious approach to help achieve this consists of examining if and how findings are affected by the way in which the concept of abnormal returns is operationalized. Such sensitivity analyses should be extensive: in my opinion, using one CAPM-based estimate of abnormal returns and one ICAPM-based estimate does not suffice, for example. After all, the subjectivity does not end with the choice of a certain model type, since each type may be implemented in numerous different ways. That is basically why I

have been referring to them as "model types" instead of as just "models". Consequently, I feel that the impact of at least some of the potential differences in model type implementations should also be investigated. For instance, as far as the CAPM is concerned, a few examples of possibly relevant issues are: does one use the return on a stock market index to proxy for the return on total wealth or is the return on human capital also accounted for in some way? Does one use conditional or unconditional market β estimates? Are the model parameters estimated on the basis of separate estimation period data and if so, what are the characteristics of that estimation period (length, frequency of observations, etc.)? Obviously, this enumeration of questions is far from exhaustive. The mere handful of exemplary issues given should nevertheless be sufficient to see that many of the choices to be made in terms of implementing a particular expected return model are likely to be also characterised by the trade-off between theoretical value and practical usefulness discussed above. As such, they sharply demonstrate how deeply ingrained the joint-hypothesis problem actually is, and hence how crucial it is for EMH researchers to report on enough sensitivity checks with respect to the procedures used for the estimation of abnormal returns.

As stated earlier, the impact that the joint-hypothesis problem has had on the empirical EMH literature cannot be overestimated. In fact, it is the ambiguities involved in estimating abnormal returns that proponents of the EMH have continually seized upon to do away with basically all rejections of the null hypothesis of market efficiency. In all honesty, one can hardly blame them: the (effect) construct validity of empirical EMH studies is indeed almost inherently low and since a chain is only as strong as its weakest link, so is those studies' overall validity. Does this mean that academics should just give up on subjecting the theory of efficient markets to empirical testing? Clearly not. I do believe that it is researchers' duty, however, to see to it that all results that are apparently inconsistent with market efficiency be complemented with as much relevant side evidence as possible to help discriminate between inefficiency-based explanations and risk-based explanations. That is why, indispensable as they may be, the sensitivity analyses with regard to the operationalization of the abnormal return concept will generally not suffice. After all, no expected return model is flawless. As a result, true adherents of the theory of efficient markets may ascribe even an

anomaly that appears to be insensitive to the method used for calculating abnormal returns to some source of risk that the various examined expected return models fail to capture. This is where the overlap with the selection threat to internal validity[27] surfaces: due to the troublesome operationalization of the effect construct, it is always debatable whether an apparent relationship between some piece of (accounting) information and subsequent "abnormal" returns evidences inefficient use of that information item or simply the fact that the latter serves as a risk proxy. With a view to protecting both internal and construct validity in the best conceivable manner, I therefore think that further action is highly recommendable. At least four potential measures come to mind that might help reduce the likelihood that the seemingly abnormal profitability of a certain investment strategy is in reality merely rewarding ex ante risk:

- first of all, one may want to be careful as far as the operationalization of the *cause* concept is concerned. Specifically, I think it is useful to steer clear of information items that are related to equity risk by construction. Examples of such items are the P/E ratio, the B/M ratio and company leverage. Ceteris paribus, the stock of highly levered firms, for example, is theoretically expected to earn higher returns than that of companies with low leverage. Naturally, those differences should disappear at the level of risk-adjusted returns. However, due to the difficulties involved in adequately adjusting returns for risk, this may very well not be the case in practice. Therefore, I feel that by examining market efficiency with respect to information items that are known to be risk-related to begin with, one surely increases the proneness of one's research design to the possibility of simply detecting inappropriate risk adjustment instead of real inefficiencies;
- secondly, Bernard, Thomas and Wahlen (1997), among others, have described how actual security misvaluations are more likely to be corrected around subsequent information releases, because the new information is supposed to elicit revisions in investors' prior (incorrect) beliefs. Assuming that this conjecture holds, a disproportionate part of the security price correction should occur at future earnings announcements, for example, relative to non-announcement periods. Note how examining the

[27] See section 5.1.

concentration of abnormal returns at specific subsequent dates is particularly compatible with indirect tests of the EMH. This approach may indeed be quite useful to separate mispricing from mismeasured risk. As Bernard, Thomas and Wahlen explicitly state, Ball and Kothari's (1991) contention that risk premiums need not arise smoothly through time and might be concentrated around information events certainly deserves to be acknowledged, but that does not make it any easier to formulate theories that predict the shifts in systematic risk required to explain the return behaviour for some regularities[28]. Nevertheless, this approach is not without its flaws. The main problem is that if abnormal returns are not concentrated around subsequent information releases, this does not necessarily constitute evidence in favour of market efficiency. In such a case, mispricings may very well have been corrected in response to anticipatory news events. It does not automatically imply that the observed unusual abnormal returns are attributable to faulty risk measurement, however. Therefore, by itself, this approach may fail to discriminate unambiguously between risk and market inefficiency;

- a third possible measure is fairly closely related in that it also works with the pre-specified informativeness of the item under consideration with respect to some future value-relevant event. This third approach consists of examining whether the estimated abnormal returns persist beyond the occurrence of that future value-relevant event. If so, risk mismeasurement obviously becomes the more likely explanation. It is on the basis of additional analyses in this direction that Stober (1992), for example, has questioned Ou and Penman's (1989a) findings;
- the fourth measure considers the potential role of risk in a more direct manner. Concretely, it departs from the premise that "risky" abnormal returns would be positive on average, but negative in certain sub-periods. This has brought Bernard, Thomas and Wahlen (1997), for example, to the suggestion to investigate the consistency with which zero-investment portfolios, i.e. portfolios representing long (short) investments in stocks expected to perform well (poorly), generate positive returns. If such portfolios

[28] Post-earnings announcement drift (see section 2.2.) is a good example of such a regularity.

consistently do well, risk-based explanations are strained of course. After all, holding on to the contention that the ex ante probability of incurring losses with a particular investment strategy is significant, although losses are (next to) never observed during the sample period, is definitely not self-evident. Conversely, observing quite many losses obviously increases the likelihood that the apparently abnormal profitability of the investment strategy concerned should in fact be attributed to inappropriate risk measurement.

On the whole, it should be clear that although tests of market efficiency are almost naturally endowed with low construct validity, there are certainly opportunities to substantially mitigate the consequences thereof. Stated differently, I feel that it would take a genuine market efficiency "fundamentalist" to continue making play with the joint-hypothesis issue in the face of empirical evidence that is inconsistent with the EMH, if that evidence withstands the various sensitivity analyses and additional tests I have described above.

7. EXTERNAL VALIDITY

With regard to the issue of external validity, Cook and Campbell (1979) distinguish between the generalizability of results *to* populations and *across* populations. They emphasize the latter form of external validity and therefore I shall also devote most of my attention to it. Nonetheless, I feel that the generalizability *to* target populations also deserves some comment. In most empirical EMH studies the target population of interest is the whole of companies listed on a particular stock market. Ideally, one would use a random selection procedure to obtain a sample representative of such a population. As explained in section 5.2., researchers have, however, typically made use of commercial databases to conduct empirical tests of market efficiency. It is important to recognize that such databases rarely to never contain similar amounts of data for all firms of the target population, which may obviously prohibit the construction of a representative sample. It is not inconceivable, for example, that a database contains relatively more (detailed) data of larger firms. If one is to adopt the approach consisting of the usage of earlier database versions that I have suggested in section 5.2. to mitigate the threat of

mortality, construction of a representative sample might be made even more difficult. Since the decision to use a particular database is generally subject to budgetary and possibly other constraints, it could be argued that this issue is largely beyond the researcher's control. Up to a certain point I am inclined to agree with this contention, but in my opinion it does not free the researcher from acknowledging the problem nor from giving the reader the opportunity to assess the representativeness of the sample used. Therefore, I find it quite important for authors to always devote sufficient attention not only to the description of their sample selection procedure, but also to the presentation of descriptive statistics pertaining to how their final sample relates to the target population. Especially as far as this second requirement is concerned, it is my opinion that the bulk of the literature on EMH anomalies has failed to come up to the mark.

7.1. INTERACTION OF SAMPLE SELECTION AND TREATMENT

The problematic construction of a sample that may be deemed representative of a particular stock market and the resulting questionable generalizability *to* a typical target population is very closely related to the first threat to generalizability *across* populations: the possible **interaction of sample selection and treatment**[29], that is. After all, the latter threat also takes root in the fact that the sampling procedure used might lead to the construction of a sample that is quite atypical of the target population as a whole. Any results obtained are then perhaps not generalizable across subpopulations of research units with dissimilar features. For example, as I have already mentioned before, it is far from inconceivable that most commercial databases contain relatively more data for large-sized companies. Empirical evidence with respect to an alleged market inefficiency for such a sample in which large companies are over-weighted need of course not apply to smaller firms as well. This does obviously not alter the fact that, in view of the practical impediments involved in most alternative data-gathering processes, the convenience of

---

[29] Cook and Campbell (1979) call this threat simply the interaction of *selection* and treatment. Personally, I do not feel entirely comfortable with this labelling. The reason is that the term *selection* that is used here is distinct from the so-called *selection* threat to internal validity, which I discussed in section 5.1. The threat to internal validity refers to the fact that initial differences among the research units may unjustly make outcomes appear to be treatment-caused, whereas here the term selection pertains to the possibility that the sampling technique used leads to the construction of a relatively atypical sample composed of research units that share some common characteristic(s), making it difficult to generalize the obtained results across subpopulations of research

commercial databases is generally of overriding importance. The pervasive use of such databases is, therefore, largely justified and in fact merely logical. As claimed in the preceding paragraph, though, researchers must on the other hand not be blind to the limitations of databases, of which the possible bias in coverage is an example. Consequently, I suggest once more that authors dedicate attention to the presentation of descriptive statistics so as to at least enable readers to get a sound understanding of the features of the samples on the basis of which empirical results are produced. In my opinion, none of the many studies I have cited in section two has satisfyingly done so.

7.2. INTERACTION OF SETTING AND TREATMENT

A second threat to the generalizability of results across populations is that of **interaction of setting and treatment**. With the exception of a few studies (e.g. Fama and French (1998)), market efficiency tests with respect to accounting information have generally focused on one stock market (or at least one country) at a time. Not surprisingly, the U.S. markets in particular have thereby attracted most attention. It is important to explicitly acknowledge the potentially setting-specific nature of the results obtained in such studies. Findings pertaining to a particular stock market must not be taken to be automatically relevant for other markets as well. Due to differences in market characteristics such as, for instance, size, liquidity, microstructure and listed firms, the limitations in terms of external validity ought to be at least considered in all EMH tests. However, in the special case of examinations of market efficiency with respect to financial statement information, another country-specific element is of course introduced into the analysis. Not only market particularities, but also features of accounting practices and underlying standards may then interact with the treatment. As such, the necessity to adopt an internationally oriented research approach if one is to achieve some level of external validity is all the more pertinent in accounting-based studies on market efficiency. Furthermore, an international research design has the major advantage that additional insights might be gained from it. Apart from the fact that EMH tests conducted for stock markets in various countries can serve as sensitivity

---

units that do not bear those particular feature(s). As the usage of the same term for two different concepts could lead to confusion, I prefer adding the word *sample* in this context.

checks, they may also provide indications as to whether and how the institutional and regulatory context influences the extent to which accounting information is efficiently processed by investors. Since the relative nature of the market efficiency concept as well as the contextual factors that might have an impact on the degree of efficiency are issues that have remained largely unexamined so far, international comparative EMH tests are in my opinion highly recommendable.

The threat of interaction of setting and treatment can also manifest itself in quite a different manner. After all, there may not only be a problem of generalizability across market settings; generalization from a research setting to a real-world setting might not be self-evident either. In fact, the simple abnormal returns that appear to be earned by simulated trading strategies will normally not be attainable in practice. This is where the very important distinction between statistical and economic significance surfaces. A statistically significant relationship between some information item and subsequent abnormal returns need not automatically constitute a rejection of market efficiency. This will only be the case if one holds on to Fama's (1970) extremely strict definition of the EMH, which requires that costless access to information and absence of transaction costs are assumed. In that case, statistical and economic significance may be considered synonymous. Those assumptions are, however, so unrealistic that Fama's version of the EMH is surely false. Therefore, there is not much point in subjecting it to empirical testing, although many studies have implicitly done so by totally ignoring the issue of market frictions. In this respect, Jensen's (1978) interpretation of the market efficiency concept is in my opinion much more useful. Yet since Jensen introduces the costs involved in investment strategies into the analysis, the distinction between statistical and economic significance is no longer trivial. Whereas statistical testing procedures are usually aimed at simply inferring whether abnormal returns are significantly different from zero, judgement as to the economic significance of abnormal returns basically requires that they be compared to the sum of information processing and transaction costs that have to be incurred to earn them. Conducting such a comparison is unfortunately far from straightforward. Especially indirect transaction costs like price pressure effects are thereby very hard to quantify. Before enumerating the major potential ways to go about incorporating transaction costs into empirical EMH studies, I would like to stress

something else. The necessity to consider the economic significance of abnormal returns makes that purely statistical testing procedures such as regression analysis can never tell the whole story[30]. Meaningful examination of the descriptiveness of the EMH requires that one always explicitly calculate the abnormal returns accruing to an investment strategy designed to exploit the alleged inefficiency under consideration. Consistent with the point that I made in section 4.1. on statistical conclusion validity, I hereby claim again that the regression and portfolio approaches to market efficiency testing ought to be regarded as being complementary.

Now how might the costs involved in investing concretely be accounted for? Fama (1991) is a proponent of simply reporting the measured abnormal returns and leaving it up to the reader to decide whether those returns are within costs. According to him, this makes it possible to (p. 1575) *"sidestep the messy problem of deciding what are reasonable information and transaction costs"*. This is true of course, but in my opinion a researcher must not avoid taking responsibility. A researcher's task does not consist of providing the reader with some output and leaving the actual analysis up to him! Consequently, I feel that the costs inherent in a particular investment strategy should always be considered more explicitly in empirical examinations of market efficiency.

One way to do so is through the use of proxy variables. Karpoff and Walkling (1988) and Bhushan (1994), for example, suggest share price, trading volume, firm size in terms of market value, and the number of shares outstanding as possible proxies under the assumption that these variables are negatively related to transaction costs. The major advantage of this approach is that an attempt can be made to assess the impact of transaction costs without having to deal with the problematic issue of actually estimating them. On the other hand, the use of proxy variables does not permit direct estimation of the effects of transaction costs. In addition, the proxies may capture effects that are not related to transaction costs.

---

[30] A very notable exception in this respect is the regression analysis used by Abarbanell and Bushee (1998), which was originally developed by Fama and MacBeth (1973). Abarbanell and Bushee transform the independent variables in such a manner that (p. 26) *"an individual coefficient represents the abnormal return to a zero-investment portfolio optimally formed to exploit the information in the associated independent variable that is orthogonal to the information in the other independent variables."*

That is probably why some authors have taken up the challenge to produce explicit transaction cost estimates anyway. Most of those authors (e.g. Stoll and Whaley (1983) and Bhardwaj and Brooks (1992)) have used the technique of "spread plus commission" to estimate transaction costs. The latter are thereby calculated as the sum of the proportional bid-ask spread and a "representative" commission from a brokerage firm. It has been shown (e.g. Lee and Ready (1991) and Petersen and Fialkowski (1994)), however, that this technique tends to yield estimates that systematically overstate actual transaction costs for the marginal trader, as many trades take place inside the quoted bid-ask spread. On top of that, the data required to calculate spread-plus-commission estimates may not be easy to come by for many stock markets. Their practical usefulness for (internationally oriented) EMH research purposes might therefore be quite limited.

That is why I have a personal preference for the approach that has recently been suggested by Lesmond, Ogden and Trzcinka (1999). Starting from the premise that the cost of transacting can be perceived as a threshold that must be exceeded before a security's price will be adjusted following the advent of information, Lesmond, Ogden and Trzcinka build on the incidence of zero daily returns to obtain estimates of round-trip transaction costs. Clearly, this procedure is not flawless either. What might be considered particularly troublesome in the context of market efficiency research is the fact that investor rationality is basically the main assumption underlying the procedure. Nevertheless, I feel that this is a classic example of the end justifying the means. For most investment strategies transaction costs are likely to be way too large to be regarded as trivial. As a result, ignoring them must not be considered to be an option. And since, as Lesmond, Ogden and Trzcinka demonstrate, their procedure (1) produces transaction cost estimates that are probably more accurate than the traditional spread-plus-commission estimates and (2) makes use of relatively easily available data, its underlying assumptions should not be prohibitive. And to totally come to terms with themselves, EMH researchers may always fall back on the proxy variable approach described above for sensitivity analyses.

7.3. INTERACTION OF HISTORY AND TREATMENT

The final threat to external validity is the possible **interaction of history and treatment**. It has to do with the generalizability of results across time periods. Concretely, empirical evidence of a market inefficiency might be specific for the sample period used and may not persist in the future. Apart from the fact that the inevitable variability in market conditions through time can always cause findings to be period-dependent, this threat could be especially pertinent in the context of EMH research as a result of the tendency toward self-destruction that is to some extent inherent in publicly available evidence of market inefficiencies. Unfortunately, as easy as the threat of interaction of history and treatment may be to understand, as difficult is it to protect one's research design from it.

Cook and Campbell (1979) suggest two possible "commonsense" solutions in this respect. The first one consists of examining whether one's findings are corroborated by existing literature in which different time periods are under scrutiny. That obviously requires that such literature is available, which is certainly not self-evident (e.g. comparative international stock market efficiency studies with respect to accounting information are still very rare). The second and probably most natural approach to deal with the threat of interaction of history and treatment comes down to replicating the research for various time periods. This solution may, however, also encounter some practical obstacles. For example, dividing the time period for which data are at hand into sub-periods while seeing to it that each sub-period contains a sufficient number of observations may not be feasible in all instances. Furthermore, even if the available dataset is large, division into sub-periods will only really enhance a study's external validity if there is enough variability across sub-periods in terms of prevailing market conditions. Otherwise, any necessary preconditions for the validity of the obtained results may very well remain unidentified. Therefore, I find arbitrary (i.e. chronological in most cases) division into sub-periods not particularly helpful in the context of external validity. It might only provide indications as to the spuriousness of findings or as to the risk involved in an allegedly profitable investment strategy. However, these are issues concerned with statistical conclusion validity and internal/construct validity, respectively. With a view to increasing the level of external validity of a research

design, sub-periods should be composed on the basis of criterions that are specified in advance and that are linked to characteristics of the general economic situation. That is the only way to reach an actual understanding of the market's behaviour under different circumstances and, hence, of the likelihood that an apparent inefficiency will continue to exist in the future. The Jensen, Johnson and Mercer (1997 and 1998) studies with respect to the size effect are excellent examples in this respect. Regrettably, researchers will of course not always have datasets at their disposal which allow that such contextual analyses be conducted. Even then, the issue should not be ignored. In my opinion, the potentially fallible external validity ought to be explicitly acknowledged and some attention would have to be devoted to a description of the prevailing macro-economic conditions during the period under investigation. In other words, the reader should always be given the opportunity to assess the market circumstances under which the reported results were obtained.

8. CONCLUDING REMARKS

In this paper, I have relied on the classical validity framework developed by Cook and Campbell (1979) to critically analyse the available evidence on market efficiency with respect to accounting information. In the process, quite a number of potential threats to the validity of accounting-based EMH tests have been identified. On the basis thereof, a substantial number of suggestions for future research in this area has been given. Repeating all of those recommendations here is unnecessary, but the most important ones are certainly worth mentioning.

As such, I have indicated my marked preference for indirect tests of the EMH, not in the least because of the inherently indirect nature of the relationship between financial statement data and subsequent long-term abnormal returns. To avoid detecting what is merely spurious covariation, I have also stressed the importance of using as treatments accounting(-based) variables for which the informativeness with respect to future value driver realisations may be specified beforehand. Another element that I have emphasized repeatedly is the fact that due to the joint-hypothesis issue empirical examinations of the EMH are naturally characterised by a low degree of (effect) construct validity. With a view to at least mitigating the

impact of the joint-hypothesis problem, I have suggested the use of normative expected return models in combination with extensive sensitivity analyses and a number of additional measures to help further reduce the potential influence of inappropriate risk adjustment. As far as statistical procedures are concerned, I have described the complementary nature of sorting-based tests and regression analyses, as well as the need to take the specific data characteristics into account when assessing statistical significance. Finally, I have underlined the importance of providing readers with sufficient side information to enable them to assess the extent to which the obtained results may be generalised both to the target population and across settings, subpopulations and time periods.

Quite remarkable is the fact that, to my knowledge, not one single study on market efficiency with respect to accounting information has satisfyingly safeguarded against all the potential threats to validity that I have identified throughout this paper. As a matter of fact, some anomalies have been discovered on the basis of research designs that are so prone to certain threats that it is extremely doubtful whether they actually pose empirical challenges to the paradigm of market efficiency. Typical examples are of course the Ou and Penman and the Holthausen and Larcker anomalies. The data mining exercises underlying these studies render their findings highly dubious. Concretely, the validity of the Ou and Penman and Holthausen and Larcker results is threatened at the very basis: their proneness to the threat of fishing is so important that it is most questionable whether there is actually any covariation whatsoever between their accounting-based treatment constructs on the one hand and their abnormal return measures on the other hand. As far as the value-versus-growth anomaly and the leverage anomaly are concerned, the documented statistical covariation is likely to be real. However, in addition to the fact that these anomalies relate financial statement information directly to subsequent abnormal returns, they entail the major disadvantage of being based on accounting measures that are inherently risk-related. Consequently, both their internal validity and construct validity are endangered. The acquisition probability anomaly on its part is interesting in that it links accounting information with subsequent abnormal returns through the occurrence of an intermediary event. Unfortunately, the prediction of corporate mergers using financial statement information is not a particularly well-researched topic, which adds a bit of an ad

hoc flavour to the specification of the accounting-based treatment. Therefore, my personal preference in terms of research design validity definitely goes to the studies that have relied on future earnings as the intermediary variable. Anomalies that have been unveiled on the basis of such studies are PEAD, the accrual-based anomaly and the Abarbanell and Bushee anomaly. Not surprisingly, these anomalies constitute in my opinion the most serious challenges to the EMH. Nevertheless, virtually all of the studies that have investigated the latter three anomalies are also capable of improvement. A very typical minus, for instance, has been the estimation of abnormal returns using positive expected return benchmarks that adjust for empiricism-based factors such as size and B/M. Also, not one study has adequately dealt with the concerns that I have expressed with regard to the external validity of research findings. The complete lack of international evidence on these three anomalies is a telling example.

Thus, although I feel that the research approaches adopted in the literature on PEAD, the accrual-based anomaly and the Abarbanell and Bushee anomaly indicate the general avenue to be followed in future investigations of market efficiency with respect to accounting information, it should be clear that there is still considerable room for methodological improvements. In my opinion, the implementation of those improvements should constitute one of the main aims of future work in this area. Only then will it be possible to obtain unambiguous results that considerably extend our understanding of the way in which financial statement information is incorporated into stock prices.

REFERENCES

Abarbanell, J. and Bushee, B. (1998), 'Abnormal Returns to a Fundamental Analysis Strategy', *The Accounting Review*, 73: 19-45.

Ali, A., Hwang, L.-S. and Trombley, M.A. (2000), 'Accruals and Future Stock Returns: Tests of the Naïve Investor Hypothesis', *Journal of Accounting, Auditing and Finance*, 161-81.

Ball, R. (1978), 'Anomalies in Relationships Between Securities' Yields and Yield-Surrogates', *Journal of Financial Economics*, 6: 103-26.

Ball, R. (1992), 'The Earnings-Price Anomaly', *Journal of Accounting and Economics*, 15: 319-45.

Ball, R. and Bartov, E. (1996), 'How Naive Is the Stock Market's Use of Earnings Information?', *Journal of Accounting and Economics*, 21: 319-37.

Ball, R. and Brown, P. (1968), 'An Empirical Evaluation of Accounting Income Numbers', *Journal of Accounting Research*, 6: 159-78.

Ball, R. and Kothari, S.P. (1991), 'Security Returns Around Earnings Announcements', *The Accounting Review*, 66: 718-39.

Barber, B.M. and Lyon, J.D. (1997), 'Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics', *Journal of Financial Economics*, 43: 341-72.

Bartov, E., Radhakrishnan, S. and Krinsky, I. (2000), 'Investor Sophistication and Patterns in Stock Returns After Earnings Announcements', *The Accounting Review*, 75: 43-63.

Basu, S. (1977), 'Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis', *Journal of Finance*, 32: 663-82.

Bernard, V.L. (1987), 'Cross-Sectional Dependence and Problems in Inference in Market-Based Accounting Research', *Journal of Accounting Research*, 25: 1-48.

Bernard, V.L and Thomas, J.K. (1989), 'Post-Earnings Announcement Drift: Delayed Price Response or Risk Premium?', *Journal of Accounting Research*, 27: 1-36.

Bernard, V.L and Thomas, J.K. (1990), 'Evidence that Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings', *Journal of Accounting and Economics*, 13: 305-40.

Bernard, V., Thomas, J. and Wahlen, J. (1997), 'Accounting-Based Stock Price Anomalies: Separating Market Inefficiencies from Risk', *Contemporary Accounting Research*, 14: 89-136.

Bhandari, L.C. (1988), 'Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence', *Journal of Finance*, 43: 507-28.

Bhardwaj, R.K. and Brooks, L.D. (1992), 'The January Anomaly: Effects of Low Share Price, Transaction Costs, and Bid-Ask Bias', *Journal of Finance*, 47: 553-74.

Bhushan, R. (1994), 'An Informational Efficiency Perspective on the Post-Earnings Announcement Drift', *Journal of Accounting and Economics*, 18: 45-65.

Brav, A., Constantinides, G. and Geczy, C.C. (1999), 'Asset Pricing With Heterogeneous Consumers and Limited Participation: Empirical Evidence', Unpublished Working Paper, Duke University.

Breeden, D.T. (1979), 'An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities', *Journal of Financial Economics*, 7: 265-96.

Brown, L.D. and Han, J.C.Y. (2000), 'Do Stock Prices Fully Reflect the Implications of Current Earnings for Future Earnings for AR1 Firms?', *Journal of Accounting Research*, 38: 149-64.

Brown, S.J. and Warner, J.B. (1980), 'Measuring Security Price Performance', *Journal of Financial Economics*, 8: 205-58.

Campbell, J.Y. and Cochrane, J.H. (1999), 'By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior', *Journal of Political Economy*, 107: 205-51.

Campbell, J.Y. and Cochrane, J.H. (2000), 'Explaining the Poor Performance of Consumption-Based Asset Pricing Models', *Journal of Finance*, 55: 2863-78.

Campbell, J.Y. and Mankiw, N.G. (1989), 'Consumption, Income, and Interest Rates: Reinterpreting the Time-Series Evidence', in: Blanchard, O.J. and Fischer, S. (eds.), *National Bureau of Economic Research Macroeconomics Annual 4*. Cambridge, Massachusetts: MIT Press.

Canina, L.M.R., Thaler, R. and Womack, K. (1998), 'Caveat Compounder: A Warning About Using the Daily CRSP Equal-Weighted Index to Compute Long-Run Excess Returns', *Journal of Finance*, 53: 403-22.

Chan, L.K.C., Hamao, Y. and Lakonishok, J. (1991), 'Fundamentals and Stock Returns in Japan', *Journal of Finance*, 46: 1739-64.

Chan, L.K.C., Jegadeesh, N. and Lakonishok, J. (1995), 'Evaluating the Performance of Value versus Glamour Stocks: The Impact of Selection Bias', *Journal of Financial Economics*, 38: 269-96.

Chandra, R. and Balachandran, B.V. (1992), 'More Powerful Portfolio Approaches to Regressing Abnormal Returns on Firm-Specific Variables for Cross-Sectional Studies', *Journal of Finance*, 47: 2055-70.

Cochrane, J.H. (2001), *Asset Pricing*. Princeton, New Jersey: Princeton University Press.

Collingwood, R.G. (1940), *An Essay on Metaphysics*. Oxford, England: Clarendon Press.

Collins, D.W. and Hribar, P. (2000), 'Earnings-Based and Accrual-Based Market Anomalies: One Effect or Two?', *Journal of Accounting and Economics*, 29: 101-23.

Cook, T.D. and Campbell, D.T. (1979), *Quasi-Experimentation – Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.

De Bondt, W.F.M. and Thaler, R.H. (1987), 'Further Evidence on Investor Overreaction and Stock Market Seasonality', *Journal of Finance*, 42: 557-81.

Desai, H. and Jain, P.C. (2000), 'Long-Run Common Stock Returns Following Financial Analysis by Abraham Briloff', Unpublished Working Paper, Southern Methodist University, Dallas.

Elgers, P.T. and Clark, J.J. (1980), 'Merger Types and Shareholder Returns: Additional Evidence', *Financial Management*, 9: 66-72.

Elton, E. and Gruber, M. (1973), 'Estimating the Dependence Structure of Stock Prices: Implications for Portfolio Selection', *Journal of Finance*, 28: 1203-32.

Elton, E., Gruber, M. and Padberg, M. (1976), 'Simple Criteria for Optimal Portfolio Selection', *Journal of Finance*, 31: 1341-57.

Elton, E., Gruber, M. and Padberg, M. (1977), 'Simple Criteria for Optimal Portfolio Selection: Multigroup Model', *Journal of Financial and Quantitative Analysis*, 12: 329-45.

Fama, E.F. (1970), 'Efficient Capital Markets: A Review of Theory and Empirical Work', *Journal of Finance*, 25: 383-417.

Fama, E.F. (1991), 'Efficient Capital Markets: II', *Journal of Finance*, 46: 1575-617.

Fama, E.F. (1998), 'Market Efficiency, Long-Term Returns, and Behavioral Finance', *Journal of Financial Economics*, 49: 283-306.

Fama, E.F. and French, K.R. (1992), 'The Cross-Section of Expected Stock Returns', *Journal of Finance*, 47: 427-65.

Fama, E.F. and French, K.R. (1993), 'Common Risk Factors in the Returns on Stocks and Bonds', *Journal of Financial Economics*, 33: 3-56.

Fama, E.F. and French, K.R. (1998), 'Value versus Growth: The International Evidence', *Journal of Finance*, 53: 1975-99.

Fama, E.F. and MacBeth, J.D. (1973), 'Risk, Return, and Equilibrium: Empirical Tests', *Journal of Political Economy*, 607-36.

Foster, G., Olsen, C. and Shevlin, T. (1984), 'Earnings Releases, Anomalies, and the Behavior of Security Returns', *The Accounting Review*, 59: 574-603.

Frankel, R. and Lee, C.M.C. (1998), 'Accounting Valuation, Market Expectation, and Cross-Sectional Stock Returns', *Journal of Accounting and Economics*, 25: 283-319.

Franks, J.R., Broyles, J.E. and Hecht, M.J. (1977), 'An Industry Study of the Profitability of Mergers in the United Kingdom', *Journal of Finance*, 32: 1513-25.

Freeman, R.N. and Tse, S. (1989), 'The Multiperiod Information Content of Accounting Earnings: Confirmations and Contradictions of Previous Earnings Reports', *Journal of Accounting Research*, 27: 49-79.

Graham, B. and Dodd, D.L. (1934), *Security Analysis: Principles and Technique*. New York: McGraw-Hill.

Grossman, S.J. and Laroque, G. (1990), 'Asset Pricing and Optimal Portfolio Choice in the Presence of Illiquid Durable Consumption Goods', *Econometrica*, 58: 25-52.

Gujarati, D.N. (1995), *Basic Econometrics*. New York: McGraw-Hill.

Holthausen, R.W. and Larcker, D.F. (1992), 'The Prediction of Stock Returns Using Financial Statement Information', *Journal of Accounting and Economics*, 15: 373-411.

Ikenberry, D., Lakonishok, J. and Vermaelen, T. (1994), 'Market Underreaction to Open Market Share Repurchases', *Journal of Financial Economics*, 39: 181-208.

Jacobs, B. and Levy, K. (1988), 'Disentangling Equity Return Regularities: New Insights and Investment Opportunities', *Financial Analysts Journal*, 18-43.

Jaffe, J., Keim, D.B. and Westerfield, R. (1989), 'Earnings Yields, Market Values and Stock Returns', *Journal of Finance*, 45: 135-48.

Jagannathan, R. and Wang, Z. (1996), 'The Conditional CAPM and the Cross-Section of Expected Returns', *Journal of Finance*, 51: 3-53.

Jensen, M.C. (1978), 'Some Anomalous Evidence Regarding Market Efficiency', *Journal of Financial Economics*, 6: 95-101.

Jensen, G.R., Johnson, R.R. and Mercer, J.M. (1997), 'New Evidence on Size and Price-to-Book Effects in Stock Returns', *Financial Analysts Journal*, 53: 34-42.

Jensen, G.R., Johnson, R.R. and Mercer, J.M. (1998), 'The Inconsistency of Small-Firm and Value Stock Premiums', *Journal of Portfolio Management*, 24: 27-36.

Johnson, N.J. (1978), 'Modified t Tests and Confidence Intervals for Asymmetrical Populations', *Journal of the American Statistical Association*, 73: 536-44.

Johnston, J. (1984), *Econometric Methods*. Singapore: McGraw-Hill.

Jones, C.P. and Litzenberger, R.H. (1970), 'Quarterly Earnings Reports and Intermediate Stock Price Trends', *Journal of Finance*, 25: 143-8.

Joy, O.M., Litzenberger, R.H. and McEnally, R.W. (1977), 'The Adjustment of Stock Prices to Announcements of Unanticipated Changes in Quarterly Earnings', *Journal of Accounting Research*, 15: 207-25.

Karpoff, M.K. and Walkling, R.A. (1988), 'Short Term Trading Around Ex-Dividend Days: Addition Evidence', *Journal of Financial Economics*, 21: 291-98.

Keim, D.B. (1988), 'Stock Market Regularities: A Synthesis of the Evidence and Explanations', in: Dimson, E. (ed.), *Stock Market Anomalies*. Cambridge: Cambridge University Press, 16-39.

Kothari, S.P., Shanken, J. and Sloan, R. (1995), 'Another Look at the Cross-Section of Expected Returns', *Journal of Finance*, 50: 185-224.

Kothari, S.P. and Warner, J.B. (1997), 'Measuring Long-Horizon Security Price Performance', *Journal of Financial Economics*, 43: 301-39.

Latané, H.A. and Jones, C.P. (1977), 'Standardized Unexpected Earnings – A Progress Report', *Journal of Finance*, 32: 1457-65.

Latané, H.A. and Jones, C.P. (1979), 'Standardized Unexpected Earnings – 1971-77', *Journal of Finance*, 717-24.

Lee, C. and Ready, M. (1991), 'Inferring Trade Direction from Intraday Data', *Journal of Finance*, 46: 733-46.

Lesmond, D.A., Ogden, J.P. and Trzcinka, C.A (1999), 'A New Estimate of Transaction Costs', *The Review of Financial Studies*, 12: 1113-41.

Lettau, M. and Ludvigson, S. (2001), 'Consumption, Aggregate Wealth, and Expected Stock Returns', *Journal of Finance*, 56: 815-49.

Libby, R. (1976), 'Discussion of Cognitive Changes Induced by Accounting Changes: Experimental Evidence on the Functional Fixation Hypothesis', *Journal of Accounting Research*, 14: 18-24.

Loughran, T. and Ritter, J.R. (2000), 'Uniformly Least Powerful Tests of Market Efficiency', *Journal of Financial Economics*, 55: 361-89.

Lucas, R.E. Jr (1978), 'Asset Prices in an Exchange Economy', *Econometrica*, 46: 1429-45.

Lyon, J.D., Barber, B.M. and Tsai, C.-L. (1999), 'Improved Methods for Tests of Long-Run Abnormal Stock Returns', *Journal of Finance*, 54: 165-201.

Lys, T. and Sabino, J.S. (1992), 'Research Design Issues in Grouping-Based Tests', *Journal of Financial Economics*, 32: 355-87.

Maddala, G.S. (1988), *Introduction to Econometrics*. New York: Macmillan.

Mandelker, G. (1974), 'Risk and Return: The Case of Merging Firms', *Journal of Financial Economics*, 1: 303-36.

Mankiw, N.G. and Zeldes, S.P. (1991), 'The Consumption of Stockholders and Non-Stockholders', *Journal of Financial Economics*, 29: 97-112.

Marais, M.L. (1986), 'An Analysis of a Multivariate Regression Model in the Context of a Regulatory Event Study by Computer Intensive Resampling', Unpublished Working Paper, University of Chicago.

Marshall, D.A. and Parekh, N.G. (1999), 'Can Costs of Consumption Adjustment Explain Asset Pricing Puzzles?', *Journal of Finance*, 54: 623-54.

Nicholson, S.F. (1960), 'Price-Earnings Ratios', *Financial Analysts Journal*, 43-50.

Ou, J. and Penman, S. (1989a), 'Financial Statement Analysis and the Prediction of Stock Returns', *Journal of Accounting and Economics*, 11: 295-330.

Ou, J.A. and Penman, S.H. (1989b), 'Accounting Measurement, Price-Earnings Ratio, and the Information Content of Security Prices', *Journal of Accounting Research*, 27: 111-42.

Petersen, M.A. and Fialkowski, D. (1994), 'Posted versus Effective Spreads: Good Prices or Bad Quotes?', *Journal of Financial Economics*, 35: 269-92.

Reinganum, M. (1981), 'A Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings Yields and Market Values', *Journal of Financial Economics*, 9: 19-46.

Rendleman, R.J., Jones, C.P. and Latané, H.A. (1982), 'Empirical Anomalies Based on Unexpected Earnings and the Importance of Risk Adjustments', *Journal of Financial Economics*, 10: 269-87.

Rendleman, R.J., Jones, C.P. and Latané, H.A. (1987), 'Further Insight into the Standardized Unexpected Earnings Anomaly: Size and Serial Correlation Effects', *Financial Review*, 131-44.

Ritter, J.R. (1991), 'The Long-Run Performance of Initial Public Offerings', *Journal of Finance*, 46: 3-28.

Roll, R. (1977), 'A Critique of the Asset Pricing Theory's Tests: Part I', *Journal of Financial Economics*, 4: 129-76.

Rosenberg, B., Reid, K. and Lanstein, R. (1985), 'Persuasive Evidence of Market Inefficiency', *Journal of Portfolio Management*, 11: 9-17.

Ross, S.A. (1976), 'The Arbitrage Theory of Capital Asset Pricing', *Journal of Economic Theory*, 13: 341-60.

Rubinstein, M. (1976), 'The Valuation of Uncertain Income Streams and the Pricing of Options', *Bell Journal of Economics*, 407-25.

Salamon, G.L. (1985), 'The Econometric Properties of Alternative Security Return Methods in the Presence of Industry and Time Period Clustering', Unpublished Working Paper, University of Florida.

Sefcik, S.E. and Thompson, R. (1986), 'An Approach to Statistical Inference in Cross-Sectional Models with Security Abnormal Returns as Dependent Variable', *Journal of Accounting Research*, 24: 316-34.

Senchack, A. and Martin, J. (1987), 'The Relative Performance of the PSR and the PER Investment Strategies', *Financial Analysts Journal*, 46-56.

Simkowitz, M. and Monroe, J. (1971), 'A Discriminant Analysis Function for Conglomerate Targets', *Southern Journal of Business*, 6: 1-15.

Sloan, R.G. (1996), 'Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?', *The Accounting Review*, 71: 289-316.

Soffer, L.C. and Lys, T. (1999), 'Post-Earnings Announcement Drift and the Dissemination of Predictable Information', *Contemporary Accounting Research*, 16: 305-31.

Spiess, D.K. and Affleck-Graves, J. (1995), 'Underperformance in Long-Run Stock Returns Following Seasoned Equity Offerings', *Journal of Financial Economics*, 38: 243-67.

Stevens, D.L. (1973), 'Financial Characteristics of Merged Firms', *Journal of Financial and Quantitative Analysis*, 8: 149-58.

Stober, T. (1992), 'Summary Financial Statement Measures and Analysts' Forecasts of Earnings, *Journal of Accounting and Economics*, 15: 347-72.

Stoll, H.R. and Whaley, R.E. (1983), 'Transaction Costs and the Small Firm Effect', *Journal of Financial Economics*, 12: 57-79.

Theil, H. (1971), *Principles of Econometrics*. New York: Wiley.

Thomas, W.B. (2000), 'A Test of the Market's Mispricing of Domestic and Foreign Earnings', *Journal of Accounting and Economics*, 28: 243-67.

Van Uytbergen, S. (2002), 'Empirical Research on the Efficient Stock Markets Hypothesis: The State of Affairs', Unpublished Working Paper, University of Antwerp.

Wansley, J.W., Roenfeldt, R.L. and Cooley, P.L. (1983), 'Abnormal Returns from Merger Profiles', *Journal of Financial and Quantitative Analysis*, 18: 149-62.

Watts, R.L. (1978), 'Systematic "Abnormal" Returns After Quarterly Earnings Announcements', *Journal of Financial Economics*, 6: 127-50.

White, H. (1980), 'A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity', *Econometrica*, 48: 817-38.

Xie, H. (2001), 'The Mispricing of Abnormal Accruals', *The Accounting Review*, 76: 357-73.