OEA Memorandum

TO: Investment Company Governance File S7-03-04

FROM: Chester S. Spatt, Chief Economist

DATE: 12/29/06

RE: Power Study as Related to Independent Mutual Fund Chairs[◊]

Executive Summary

Existing empirical studies of the effects of mutual fund governance have failed to consistently document a statistically significant relation between fund governance and performance, particularly with respect to board chair independence. We suggest that the lack of such evidence may be a result of the limits of standard statistical methods in identifying such a relation and is not necessarily indicative of the failure of such a relationship to exist.

In this memorandum, we describe the factors affecting the probability that a statistical test will identify a relation, if that relation actually exists—a concept known in statistics as the *power* of the test. In order to highlight the importance of power with respect to tests of differences in mutual fund returns, we assess how the power of statistical tests may be affected by the randomness in mutual fund returns. We provide evidence of the amount of data necessary to identify a given magnitude of effect and the magnitude necessary to identify a relation for a given data sample size. Our analysis suggests that most studies assessing the impact of chair independence on returns do not have sufficient power to reliably conclude that a relation does or does not exist.

In addition, empirical analysis in this area is constrained by limitations on the data and data collection. The data is limited because while there are over 8,400 mutual funds in the U.S.,[1] fund governance is typically determined at the fund complex level. While many fund families have a unitary board structure where a single board oversees all funds within the family, a non-trivial proportion of families have multiple boards. On average, mutual fund families have two boards, with considerable overlap in directors across boards. Further, because mutual fund boards do not change much from year to year, even a long time series of observations will likely have limited variation. These data issues are compounded by the difficulty in gathering fund governance data, which has historically been difficult to retrieve from paper filings. In fact, few studies actually collect more than a year or two of governance data.

[◊] This memorandum was drafted by the staff of the Office of Economic Analysis (OEA) and has been updated through December 2006. The Commission has expressed no view about its contents.

[1] See Investment Company Institute, 2006 Investment Company Fact Book, Washington D.C., 2006.

In sum, given the degree of randomness in mutual fund returns and the paucity of available fund governance data, standard statistical approaches have low power to identify relatively small return differences.

Introduction

Statistical power is defined as the probability that a statistical method will identify an effect or relation when the effect or relation in fact exists. This memorandum addresses issues of power in statistical analyses, particularly as applied to recent and ongoing studies of the effects of mutual fund board independence on performance, fees, and compliance. Economic analyses, including those published in peer-reviewed scholarly journals, place strong emphasis on the statistical significance of results. Discussions of economic significance and the power of statistical methods employed are often given a back seat to this pursuit of statistical significance.[2] This memorandum will explain how empirical analyses suffer as a result of this imbalance.

In particular, we will suggest that broad, cross-sectional analyses attempting to identify significant differences in financial performance attributable to differences in governance in mutual funds will suffer from low power (i.e., a low probability of identifying a relationship). As we will describe, several factors, including the degree of noise in mutual fund returns, the existence of many other factors that lead to predictable differences in returns across mutual funds and limitations in the data typically employed by empirical researchers will contribute to the low power of the test. Given the lower probability of identifying a relation between fund governance characteristics and returns (were it to exist), it is difficult to draw a strong inference from the lack of extant academic evidence on this topic.

This memorandum does not speak to the question of whether a relation between mutual fund chair independence and performance does exist, and if so what is the magnitude of that relation. Rather, it addresses a more focused question – if such a relation were to exist, what would be necessary – in terms of data and strength of relationship – for standard econometric techniques to identify it with confidence. Indeed, economic theory suggests that if there are no impediments to markets working efficiently, mutual funds and their shareholders would select governance characteristics in an optimal manner and there should be no expected relation between those characteristics and fund performance.[3]

We begin by addressing some of the underlying statistical issues, using a simple example. We then provide an overview of statistical hypothesis testing and related power issues. We apply these concepts to the analysis of mutual fund performance and present a series of power curves for relevant cases. We address a number of issues and complications, and apply our results to the nascent literature on the effects of mutual fund board chair independence. We conclude with a discussion of limitations in the statistical approach and suggest prescriptions to address these concerns. While we apply our analysis to studies of the effect of chair independence on performance, parallel analyses could be conceived with respect to fees and compliance issues.

---

[2] For a more thorough treatment, see L. Harris, Trading and Exchanges, Oxford University Press, New York, 2003, pp. 454-466.

[3] For a more complete discussion of these points, see OEA's companion memorandum reviewing the mutual fund governance literature.

Key Concepts in Statistical Analyses

This section is intended to give the reader a brief introduction to some of the concepts underlying common statistical analyses, in particular *random sampling*, *estimation* and *hypothesis testing*. While presented in the form of a simple example, these concepts and related issues are directly applicable to the mutual fund issues that are our current concern. We provide this background because we will apply these concepts in the context of the effects of mutual fund board chair independence in a subsequent section.

We will first highlight the difference between *economic significance* and *statistical significance*. Economic significance is a fairly imprecise concept and is typically case dependent. Generally, it is the characteristic of being substantive and meaningful in an economic sense, and is perhaps the opposite of the legal term *de minimis*. Statistical significance, alternatively, is a well-defined measure of the probability with which a finding may occur in the data by chance. Economic literature has in recent decades focused on the level of statistical significance of empirical findings and in many cases failed to differentiate between economic and statistical significance.[4] It is important to keep in mind that assessing economic significance and assessing statistical significance are two separate steps in a well thought out analysis.

A Simple Example

Many introductory statistics and probability textbooks employ a simple example to convey the basic precepts of random sampling, parameter estimation, and hypothesis testing, and so we will proceed in this manner. Suppose you have an urn filled with 1000 balls, some of which are blue and some of which are red. Suppose further that you do not know the number of blue and red balls and that you are unable to remove all of the balls and count them, but that you wonder whether there are more red balls or more blue balls.

Estimation

In order to estimate the proportion of red balls in an urn, you might conduct a statistical experiment by randomly drawing balls from the urn (call this a random sample), and noting the color of each ball. In this manner, the *proportion* of red balls in your sample is an estimate of the proportion of red balls in the population (i.e. the 1000 balls in the urn). Of primary concern is the question: How accurate is my sample estimate? That is, how close is the (observed) proportion of red balls in my sample to the (unobserved) proportion of red balls in the population?[5] The statistician might rephrase this question as follows: What is the probability of a random sample of N balls containing R red balls, *if* the true proportion is 50% red balls? For the sake of exposition, let's assume that you draw 10 balls from the urn randomly with replacement, and that 6 of them are red and 4 of them are blue.[6] Thus, your sample estimate of the proportion of red balls in the population is 6/10.

---

[4] See D. N. McCloskey and S. T. Ziliak, "The Standard Error of Regression," Journal of Economic Literature, March 1996, and D. N. McCloskey and S. T. Ziliak, "Size Matters: The Standard Error of Regressions in the *American Economic Review,*" Journal of Socio-Economics, November 2004.

[5] A key point is that the larger the sample size, the more accurate the sample estimate. The trivial case in our example, in which all of the balls can be observed, is equivalent to the unrealistic situation in which mutual fund performance is entirely *deterministic*—that is, there is no unexplained noise in returns.

[6] We will assume, without loss of generality, that each ball is replaced before drawing the next, a process called *sampling with replacement*.

Hypothesis Testing

Intuitively, if the realized sample (6/10) differs substantially from our guess of the true proportion (1/2), our guess is less likely to be correct.[7] A statistician may formalize this concept in a *statistical hypothesis test*. Put simply, a statistical hypothesis test is conducted by setting up a straw man, in the form of a *null hypothesis*. The researcher then uses available data to either *reject* or *fail to reject* the null hypothesis. In the urn example, the statistician may form these simple null and alternative hypotheses; [8]

- *Null Hypothesis: The proportion of balls in the urn is* 1/2.
- *Alternative Hypothesis: The proportion of balls in the urn is not* 1/2.

If we are able to reject the null hypothesis we can state with some degree of statistical confidence that the proportion of balls in the urn is not 1/2. However, the converse is not necessarily true. If we are unable to reject the null hypothesis, we have shown only that the data does not provide statistically significant evidence that the true proportion is not 1/2. In fact, there are a number of potential reasons why the null hypothesis may not be rejected, only one of which is that the true proportion is 1/2. Another possible reason is that there is insufficient data –i.e. that our sample size is too small. This may seem a trivial discussion in this setting, but should be well understood since it plays an important role in making inferences from studies of mutual fund board chair independence.

A hypothesis test is conducted by forming a *test statistic* (in our simple case, using the sample proportion 6/10, the hypothesized true proportion 1/2, and the sample size 10), which is compared to a table of *critical values* for the particular test. If the test statistic is greater in absolute value than the critical value, then the test is said to *reject* the null hypothesis, otherwise the test is said to *fail to reject* the null hypothesis.[9]

The reliability and efficacy of hypothesis tests are functions of the data on which they are based (in our simple example the sample size and estimated proportion). One of the key characteristics of a statistical test is the *size* of the test,[10] which determines the critical value used to determine whether or not the null hypothesis is rejected. A smaller *size* is equivalent to a higher critical value of the test statistic. *Size* denotes the probability of a *Type I error*—that is, the probability that we reject the null hypothesis when it is true. In other words, it is the probability that the test will determine that the true proportion is not 1/2, when in fact the true proportion is 1/2. Typically, economists set the *size* of a test at 5%; they will conclude that the

[7] More rigorously, given assumptions about the distribution of the sample estimate (for example, Normal with mean equal to the hypothesized true population value), the researcher can estimate the probability that a randomly drawn sample will have a value at least as far from the mean as that of the observed estimate. The higher this probability, the less likely is the true mean to be equal to the hypothesized value.

[8] This is referred to as a two-sided hypothesis test, because it describes the situation where a researcher wishes to test that the true proportion is either greater than or less than 6/10. The reader may note that the alternative hypothesis embodies the researcher's prior belief about the effect, with the null hypothesis its complement.

[9] For computational details, see P. Newbold, Statistics for Business and Economics, Prentice Hall, New Jersey, 1988.

[10] Statisticians and empirical researchers denote the size of a test as *alpha*. The reader should not confuse this with the common use of the symbol *alpha* to refer to a measure of abnormal performance in returns. To avoid this confusion we will simply use the italicized *size*.

true proportion is not statistically significantly different from 1/2 at the 5% level if it is determined that there is only a 5% chance that the true proportion (of which the observed sample estimate is a potentially noisy estimate) is actually not 1/2. Medical researchers will often use a *size* of 1% or even 0.1%, which is logical in light of the fact that such researchers are often studying the effects of, for example, drug dosage, and describing as safe a dosage at which the drug is not safe can have lethal implications.

A related characteristic of a statistical test is the probability of a *Type II error*—the probability that we accept the null hypothesis when it is false. In other words, it is the probability that the test will indicate that the true proportion is 1/2 when in fact it is not.[11] The complement of this probability, calculated as 1- (the probability of a *Type II error*), is called the *power* of the test, and is defined as the probability that the test will correctly indicate that the true proportion is (in this example) 1/2. The following table summarizes these points.

Table 1: Error Types in Statistical Hypothesis Tests

	Test is Significant	*Test is Not Significant*
Null Hypothesis is True	Type I Error (Size)	Correct Outcome
Null Hypothesis is False	Correct Outcome (Power)	Type II Error

* Correct Outcome indicates that the null hypothesis is rejected when it is in fact false (conversely, not rejected when true)
Size = Probability of a Type I Error
Power = Probability of a Correct True Decision = 1 - Probability of a Type II Error

A researcher must weigh the cost of both Type I errors (concluding a relationship exists where it does not) and Type II errors (failing to find a relationship where one does exist) in designing analyses. It can be shown that, all else held constant, there is a trade-off between Type I errors and Type II errors (i.e. between *size* and power). Intuitively, the more certain you are that a statistical test will not incorrectly find a relation that does not exist, the less certain you can be that you will identify the relation when in fact it does exist.[12]

Mutual Fund Returns, Hypothesis Testing and Power

It may be useful to briefly touch on the characteristics of mutual fund returns that drive power issues in statistical analyses. A mutual fund is a portfolio of assets. Economic and financial researchers think of there being an underlying process generating asset returns, and thus portfolio returns.[13] There exists a broad literature on the properties of portfolio returns, in particular, the econometric modeling of returns—that is, the identification of variables that drive returns. A key finding of this research is that the degree of noise in returns—that is, the amount of variability of returns unexplained by variability in other determinants of returns—is substantial.[14]

[11] Statisticians and empirical researchers denote the probability of a *Type II error* as the *beta* of a test. As with *alpha*, we will avoid this usage to ensure the reader does not confuse this idea with the more common financial meaning of *beta*.
[12] Thus, it can be seen that the practice of defining the alternative hypothesis so as to embody the researcher's prior belief about the effect serves to set the probability of erroneously concluding that his prior beliefs are true at the chosen *size*.
[13] Econometricians refer to this process as the *data generation process*, or DGP. Assumptions about the DGP drive many econometric results, and so are of much import to empirical researchers.
[14] Economists refer to this unexplained, or random, return as the *stochastic* component.

A key question in analyses of mutual funds, and in particular of mutual fund governance, involves determining the appropriate unit of measurement. The researcher must determine whether analyses should be conducted at the mutual fund family/complex level, the board level, the fund level, or the class level.[15] If one wishes to account for differences in fee structures, clearly class-level differences must be taken into account, as differences in fee schedules are the primary characteristics differentiating classes within a fund. Alternatively, if the effect of governance structure is of primary concern to the researcher, then the board is the appropriate unit of measurement.[16] A non-trivial ramification of choosing the fund- or class-level is that there are variables which do not vary across funds within a family (or board). We make this point so that the reader may avoid the trap of naïvely believing that choosing the fund-or class-level over the family- or board-level will eliminate the power issues we discuss below.

The Simple Probability Example Applied to Mutual Fund Returns

The example above of the urn is purposefully simple, but can be extended quite easily to empirical questions regarding mutual fund returns. This section applies our previous discussion to the effects of mutual fund board chair independence. We consider the case where a researcher wishes to analyze the difference in returns across two groups of mutual funds, specifically those chaired by independent directors and those chaired by management directors.

Hypothesis Testing and Mutual Fund Returns

Empirical analyses, whether quantitative or qualitative in nature, rely on the statistical properties of the estimators used to make statements about the relation between variables.[17] A characteristic of the *statistical hypothesis testing* approach taken by most empirical researchers (and applied as part of *regression analysis*) is that failure to identify a relationship represents lack of evidence that the relationship exists, but should not be interpreted as positive evidence that the relationship does *not* exist. This discussion will highlight why this is the case.[18]

As discussed above, a statistical hypothesis test involves setting up a 'straw man' in the form of a *null hypothesis*. The researcher then uses available data to either *reject* or *fail to reject* the null hypothesis. For example, a researcher may suspect that mutual funds chaired by independent directors perform better than those chaired by management directors. He may formulate the following simple null and alternative hypotheses:[19]

- *Null Hypothesis: Mutual funds with an independent chair produce returns no higher than those with a management chair.*

[15] A mutual fund family is made up of a collection of funds, each of which may offer one or more classes. In addition, many larger families have more than one board of trustees, each governing a subset of that family's funds.
[16] In fact, some existing studies of mutual fund governance are carried out at the fund family level, in effect marginalizing differences across multiple boards within a family.
[17] Qualitative analyses are those in which the variable of interest is represented by two or more categories, such as in a Logit or Probit analysis.
[18] For a more in-depth introduction of statistical modeling and hypothesis testing, see P. Newbold, Statistics for Business and Economics, Prentice Hall, New Jersey, 1988.
[19] This is an example of a one-sided hypothesis test because it describes the situation where a researcher wishes to test that the true difference in returns is positive.

- *Alternative Hypothesis: Mutual funds with an independent chair produce higher returns than those with a management chair.*

If we can state with some degree of statistical confidence that it is not the case that mutual funds with an independent chair produce returns no better than those with a management chair (reject the null hypothesis), then we can say with similar confidence that independent chaired funds produce higher returns than do management-chaired funds. Again, the converse is not necessarily true. If we are unable to reject the null hypothesis, we have shown only that the data does not provide statistically significant evidence of higher performance among independent chaired funds. There are a number of potential reasons for this, only one of which is that the hypothesized relationship does not exist. Other possible reasons are that there is insufficient data or that the degree of noise in the data swamps the magnitude of the 'true' difference in returns across the two groups, rendering the difference difficult to identify using standard statistical methods.

Hypothesis testing has many applications in empirical analyses. The example described above is a *two-sample test of means*. This type of test is designed to examine differences in average performance across two groups. However, this approach fails to account for other differences across the two groups which may drive performance. A more rigorous approach, typically employed by researchers, is to use *regression analysis* to correct for other known factors that might affect performance.

More specifically, a natural approach to analyzing the impact of independent chairs would be to start with a regression model in which performance is related to a collection of variables thought to explain fund returns, and then add an additional variable that indicates whether the fund has an independent chair (in other words, a variable that takes on a value of 1 for funds with an independent chair and 0 for funds with a management chair). Regression output includes a *coefficient* for each of the independent variables, along with a *standard error*, a statistical measure of the precision of the estimate. The researcher can then use the coefficient estimate, along with the corresponding precision measure, to test the null hypothesis that the variable in question (in this case, board chair independence) has no significant impact on fund performance. Again, rejecting the null hypothesis provides evidence in support of the alternative, that funds with independently chaired boards outperform those with management chairs, while failure to reject the null hypothesis should be interpreted as inconclusive. In the special case we have been describing, where we analyze differences across two groups, the two-sample hypothesis test of means and the regression-based test produce similar inferences.

Power Curves for Relevant Cases

As discussed above, economic and financial researchers typically fix *size* at 5%. In the context of mutual fund performance, the power of a 5% statistical test will be a function of the *signal-to-noise ratio* in the data and the number of observations. The signal-to-noise ratio describes the relationship between the magnitude of the true effect (signal) and the amount of unrelated variability in the data (noise). The higher the magnitude of the true effect relative to the amount of unrelated variability in the data (i.e. the higher the signal-to-noise ratio), the more likely a statistical test will identify an effect in a given data set. Accordingly, the higher the signal-to-noise ratio and the more observations a researcher uses in his study, the higher the

power of the resulting test, and the more likely the researcher can identify a relationship when in fact it does exist. In our mutual fund example, the signal-to-noise ratio will be determined by the 'true' performance difference between independent chaired and management chaired funds and by the standard deviation of fund performance. It is left for the researcher to select the data to be used for analysis, in particular, to specify the sample size.[20]

Researchers typically use power calculations to answer one or both of the following questions:

- *How powerful is my statistical test given the data that I have?*
- *How much data do I need to construct a test which is 'powerful enough'?*

At the heart of these questions is the idea of how much power is 'enough'. As in the case of determining the *size* of a test, the question of how much power is enough is largely a function of context. We take as given from our understanding of academic studies that economic and financial researchers prefer tests with power no lower than 0.8, representing an 80% probability of identifying a relationship if one exists.

The power that any researcher ultimately accepts for her test may ultimately and critically depend upon the ability to acquire data and at what cost. Experimentalists operate largely within the confines of expensive and difficult to obtain survey data, and therefore may estimate power so as to minimize the expense of an experiment. Empiricists in the areas of economics and finance are typically driven largely by the paucity of data in most applications outside of microstructure studies. In fact, power issues are largely ignored in most economic studies, presumably on the basis that data constraints are binding. The McCloskey study noted above finds that during the 1980s and 1990s, only 4.4% and 8%, respectively, of papers published in a leading economic journal considered the power of tests employed.[21]

Perhaps the most effective way to present power calculations is with a series of power curves, which map out calculated power for a range of sample sizes given assumptions about the signal-to-noise ratio.

How to Read a Power Curve

Continuing our example of examining the performance of independent versus management chaired mutual funds, we have calculated asset-weighted average annual returns for every fund family in the CRSP mutual fund database.[22] The cross-sectional standard deviation of these returns for 2002 was approximately 12%. Figure 1 presents power curves for 'true' differences ranging from 1% to 5% (100 to 500 basis points), with sample size ranging up to

[20] We note that a sophisticated researcher may attempt to minimize the signal-to-noise ratio either by decreasing the noise or by increasing the signal. This can be accomplished by carefully accounting for the determinants of performance or by, for example, defining discrete performance measures.

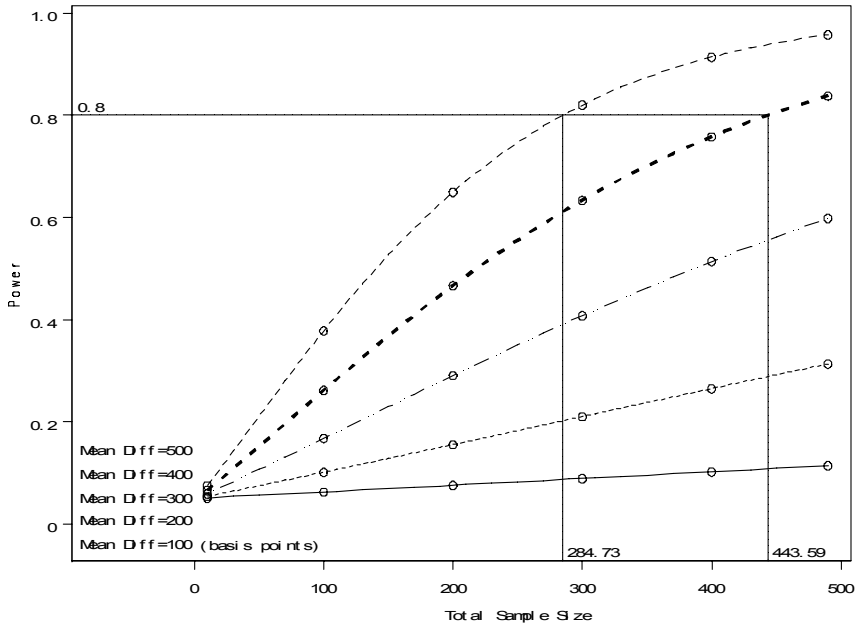
[21] See D. N. McCloskey and S. T. Ziliak, "Size Matters: The Standard Error of Regressions in the *American Economic Review*", Journal of Socio-Economics, November 2004.

[22] This is the CRSP Survivor-Free US Mutual Fund Database, produced by the Center for Research in Securities Prices at the University of Chicago Graduate School Of Business.

500, consistent with the sample sizes of existing family and board-level studies of the effects of independent mutual fund chair.[23] A few tips on reading the graph:

- Identify the curve related to the 'true' difference consistent with your beliefs.
- Identify the sample size consistent with your dataset on the horizontal axis.
- Read the power estimate for that data point on the relevant power curve from the vertical axis.

Figure1: Power Curves for a Test with 5% *Size* and Various 'True' Performance Differences, Given Standard Deviation of 12%



It bears noting that all of the 'true' performance differences presented here are economically significant. Assuming the 'true' difference in performance between independent chaired and management chaired mutual funds is 3% (300 basis points), and that we have data on 448 mutual fund families for a single year,[24] we can see that our test will have power of (approximately) 55%. This means that a researcher testing for a statistically significant difference in returns will fail to identify a difference 45% of the time.

An immediate conclusion from examining these power curves is that given the amount of variability in mutual fund performance, even at the family level, it takes a large amount of data to identify even sizable differences in performance across groups using standard statistical

[23] Note that we have also assumed that 1 in 5 fund families are chaired by an independent director or trustee.
[24] The figure of 448 families represents the same breadth of coverage as the Ferris-Yan study. The literature review to which this memorandum is a companion piece provides further details. See Ferris, S. and X. Yan, "Do Independent Directors and Chairmen Really Matter? The Role of Boards of Directors in Mutual Fund Governance," working paper, University of Missouri – Columbia, 2005.

approaches.[25] If realistic differences in performance are substantially smaller, the power problem is amplified.

Regression Analysis and Power

There are a number of approaches to minimizing the noise (and thereby maximizing the signal-to-noise ratio) in a given data set. Regression analysis allows the researcher to explain some percentage of the variation in fund performance as attributable to variation in other variables. For example, a researcher might believe that performance is strongly related to a variety of fund- and family-level characteristics other than chair independence. Regression analysis allows the user to examine the effect of chair independence conditional on, or net of, the effects of other variables. A measure of the extent to which this variation is 'explained away' is given by R^2, which reports the percentage of the variance of the dependent variable (performance) explained by the independent variables.

Continuing with the fund board chair independence as our example, it may be useful to examine what happens to the power curves subsequent to this increase in the signal-to-noise ratio. Specifically, we propose two hypothetical cases; the first assumes that the researcher's model explains 50% of the variance in performance (i.e. R^2=0.50), and the second assumes the researcher's model explains 75% of the variance (i.e. R^2=0.75). It should be noted that both of these are high R^2 values not typically achieved in modeling mutual fund returns, and are presented as hypothetical 'what-if' scenarios.[26]

Figures 2 and 3 present power curves for each of these cases, respectively. These reductions in noise are represented by decreases in the standard deviation of performance from 12% to 8.5% and 6%, respectively.[27]

[25]Even the smallest 'true' difference we consider, 1%, would be thought of as highly economically significant.
[26] As will be discussed below, the Meschke study reports R^2 values ranging from 0.01 to 0.50 in performance regressions. See J. Meschke, "An Empirical Examination of Mutual Fund Boards," working paper, Carlson School of Business, University of Minnesota, Minneapolis, MN, 2006.
[27] As variance is the square of standard deviation, a 50% reduction in variance (from 144 to 72) implies a reduction in standard deviation from 12 to 8.5.

Figure 2: Power Curves for a Test with 5% *Size* and Various 'True' Performance Differences, Given Standard Deviation of 12% and R^2 of 0.50

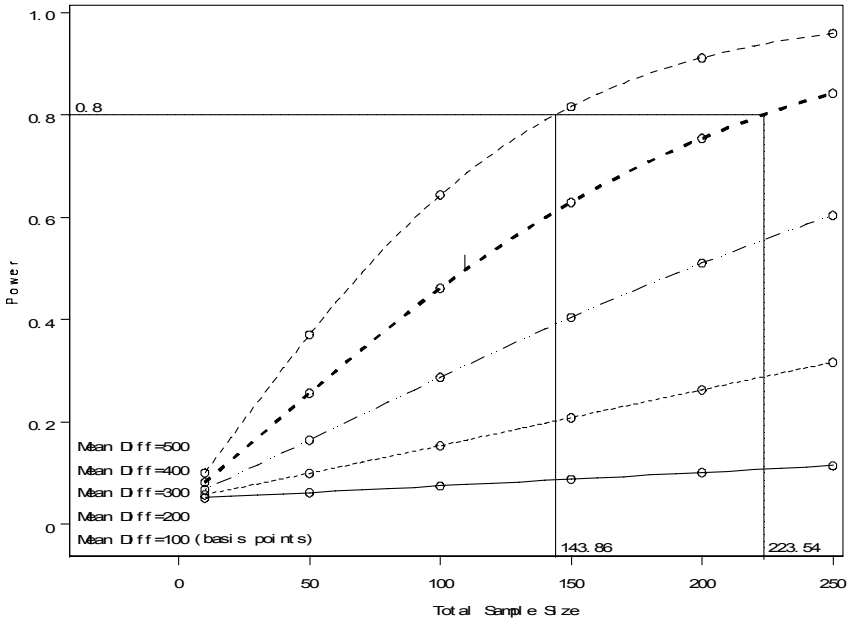
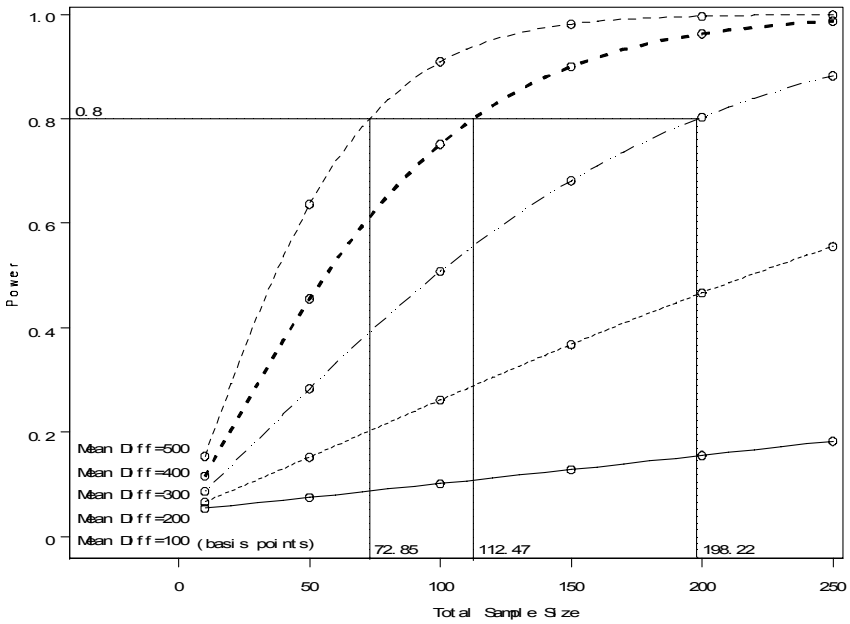


Figure 3: Power Curves for a Test with 5% *Size* and Various 'True' Performance Differences, Given Standard Deviation of 12% and R^2 of 0.75



The results suggest that successfully modeling a large fraction of the variance in performance results in increased power, and in fact for large 'true' differences in performance existing studies achieve an acceptably high degree of power. A researcher could achieve the

desired level of 80% power if the 'true' performance difference were 3%. However, for smaller differences (2% or lower) power is lower. We will discuss in a subsequent section the likely magnitude of 'true' performance differences.

An Alternative Approach: Calculating Required Sample Size

Alternatively, a researcher may wish to identify the sample size needed to achieve acceptable power under given conditions. Below we present a chart of sample sizes required to achieve power of 50%, 80%, and 95%, for the same standard deviation and R^2 used in figure 3 above, and 'true' performance differences of 100, 200, and 300 basis points.

Note that for a 5% test to achieve 80% power with a 'true' performance difference of (an economically significant) 100 basis points, a researcher would require a sample size of at least 1,770. These results mirror those of the power curves presented above. In fact, there are far fewer than 1,770 mutual fund families, so that given these assumptions a cross-sectional test based on one year of family-level data cannot achieve 80% power.

Table 2: Required Sample Sizes for Various 'True' Performance Differences and Power Requirements, Given Standard Deviation of 12% and R^2 of 0.75

'True' Performance Difference (BP)	Desired Power	Required Sample Size (N)		
		***Size* = 1%**	***Size* = 5%**	***Size* = 10%**
100	50%	1500	870	610
100	80%	2640	1770	1400
100	95%	4020	2930	2440
200	50%	380	220	160
200	80%	670	450	350
200	95%	1010	740	620
300	50%	170	100	70
300	80%	300	200	160
300	95%	450	330	280

Issues and Complications

There are a number of power-related issues worth mentioning. These are: 1) the effects of cross-correlations in investment strategies on mutual fund returns; 2) the choice of aggregation level—that is, should analyses be conducted at the fund, board; or family-level, 3) differences in performance measures used by researchers; 4) the general lack of a structural model for mutual fund returns; and 5) questions of cross-sectional versus time series or panel data analyses. We will briefly touch on each of these issues.

Cross-correlations. There are more than 8,400 mutual funds in the US, and while there are many dimensions to the investment strategies implemented by these funds, there are groups of funds that share similar characteristics. This is evidenced both by the proliferation of categorization schemes used throughout the industry as well as the practice of benchmarking to an appropriate index. In many cases, mutual funds have stated indexes against which their

performance is measured. The financial press often benchmarks fund performance against these indexes or, alternatively, against a peer group defined by synthetic categorizations such as the popular Morningstar style box. This focus on benchmarking can lead to what is termed *closet indexing*, where nominally actively managed funds are in fact following investment strategies nearly identical to their stated or implied index. Funds that practice similar investment strategies will produce highly *correlated* returns.

There are other dimensions across which groups of funds share strong similarities. Funds governed by the same board will clearly share the same governance characteristics, while funds within a family will share the same family-level characteristics, such as family size and distribution channel(s). The measures of statistical significance often employed in empirical analysis to quantify the probability that two samples (e.g., management-chaired vs. independent chaired-funds) have different average performance or expenses rely on the assumption that all observations within the sample are independent of each other. Because analyses conducted at the fund level violate this assumption, this method makes it more difficult to identify the true statistical significance of differences observed across the samples.[28]

Aggregation Level. A similar issue arises from the choice of aggregation level. There are four primary approaches, ranging from the individual mutual fund class level to the fund family level. In between are the portfolio (i.e., series) level and the board level. Broadly, fund families are composed of portfolios (or series) with different investment strategies, each of which offers a number of share classes differentiated by distribution channel and fee structure. Each family has one or more boards, each of which oversees one or more funds.[29] The appropriate unit of measurement for a particular analysis is not always clear. Raw returns vary across portfolios, while fees and expenses vary across classes. Family and board-level characteristics are likely to be key determinants of both returns and fees. Many existing studies address unit-of-measurement questions by performing analyses at multiple levels.[30]

It is difficult to make concrete assertions about the effects of unit of measurement choices and related *cross-sectional correlations* in strategy and fund management structure, in the absence of deeper knowledge about the structure of the cross-correlations. However, it is clear that fund class or portfolio-level analyses may not be the quick fix to the power issues discussed above that they may appear to be, relative to family or board-level analyses. The cross-sectional correlations inherent in class- and portfolio-level analyses may act to reduce the *effective sample size*, which may have a deleterious effect on statistical power.

---

[28] See e.g., D. Latzko, "Economies of Scale in Mutual Fund Administration," *Journal of Financial Research* 22 (1999), pp. 331-339.

[29] The staff of the Office of Economic Analysis believes there are cases in which different share classes are overseen by different boards, but that these are few.

[30] For example, Ferris and Yan (2005) adopt this approach. The authors comment on the potential for fund-level analyses in the presence of cross-correlations to result in overstating the statistical significance of hypothesis tests. We wish to remind the reader that their comment speaks to the probability of a Type I error (*size*), while our concern is with the power of the test. Our assertion is that cross-correlations result in a less-than-expected increase in power when comparing fund- and family-level analyses, and is not inconsistent with Ferris and Yan's comment. See Ferris, S. and X. Yan, "Do Independent Directors and Chairmen Really Matter? The Role of Boards of Directors in Mutual Fund Governance," working Paper, University of Missouri – Columbia, 2005.

Performance Measures. There are numerous approaches used by economic and financial researchers to measure performance. These range from using raw returns or returns minus an appropriate benchmark, to more complex multi-factor models in which the fund's level of risk relative to a set of benchmarks is accounted for. Other approaches involve using rankings or some combination of approaches (such as the Morningstar star ratings). While several of these approaches may explain some of the variability of returns, in particular risk-adjusting approaches, the resulting performance measures remain noisy relative to the likely magnitude of 'true' performance differences between independent and management chaired funds, and low power remains an issue.

Lack of a Structural Model. An even more difficult issue involves the general lack of a conceptual or theoretical framework for the relationships being examined. The reader may wish to refer to the related Office of Economic Analysis literature review for a more detailed discussion of this issue, but it should be noted here that in the absence of a theoretical framework on which to base empirical analyses, a researcher increases the risk of data-mining, in which case his personal biases may unduly influence the results. In this case it can be difficult to evaluate the researcher's findings and distinguish among competing interpretations.

Cross-sectional vs. Time Series Issues. A final issue relates to differences between *cross-sectional* and *time series analysis*.[31] Existing studies of mutual fund board independence have examined fee and performance differences in relation to differences in fund governance across groups of funds or families at a single point in time, and so are considered to be cross-sectional in nature. Time series analysis involves the study of changes in the characteristics of a single unit of observation over time, and is subject to statistical assumptions and issues not faced by purely cross-sectional analyses. Studies that combine these two aspects—for example, a study of all mutual fund families over several years—are called *panel data analyses*, and present a number of interesting econometric issues.[32] Expanding the time series dimension of a dataset dramatically increases the number of observations and is highly valuable in low power studies of limited cross-sectional dimension, such as the fund chair independence example described here.[33]

However, expanding a dataset from a simple cross-section to a panel requires that the researcher gather all information for each period (e.g. year) in the sample. It bears noting since it appears that some studies have collected a 'snapshot' of governance data and applied those characteristics to multiple years of performance and expense data.[34] We presume that this approach has been taken due to the difficulty and expense in collecting multiple years of governance data. This approach raises several questions. First, if any funds changed governance structure during the sample period the data may be misapplied and thus any inferences from this analysis may be inappropriate. Second, if governance characteristics evolved in predictable ways

[31] See e.g., J. D. Hamilton, Time Series Analysis, Princeton University Press, Princeton, 1994.
[32] See e.g., B. H. Baltagi, Econometric Analysis of Panel Data, John Wiley & Sons, Ltd, New York, 2001.
[33] It bears noting here that the naïve approach of maximizing the time series dimension of a panel of data by using a variable measured at a smaller interval, say daily rather than annual returns, will not help resolve the power issue. This approach will not yield the increase in power one might expect from such a dramatic increase in the number of observations. This holds because of: 1) the difference in signal-to-noise ratios between daily and annual returns and, 2) the absence of variation in governance characteristics at the daily level.
[34] See e.g., the discussions of the Bobroff-Mack and Meschke studies in OEA's companion literature review.

(e.g., if boards generally had higher proportions of independent directors in 2002 than in 1998), this misapplication will systematically bias the results of the analysis. Third, to the extent that this approach understates the true variation in the data across the sample period, any benefits that may have accrued to the power of the test by expanding the sample are substantially lower than predicted.

Power Analyses of Existing Mutual Fund Studies

Economically Significant Differences in Mutual Fund Returns

The foregoing discussion highlights the difficulty of identifying differences in mutual fund performance across groups, in the presence of some degree of noise and some amount of data. We have been purposefully agnostic with respect to whether a 1% or a 3% difference in fund performance is either economically significant or likely in practice.

The hypothetical cases above have used differences measured in percentage points (100s of basis points) primarily because such differences are necessary to generate informative power curves. 'True' differences may be more likely to be measured in basis points, rather than percentage points.

To address the issue of economic significance, let us use the archetypical example of investor A, who invests $100,000 in The Management-Chaired Fund, which subsequently generates returns net of expenses of 8% per year, and investor B, who invests $100,000 in The Independent-Chaired Fund, which subsequently earns 8.5% per year, net of expenses.[35] At the end of 10 years, investor A will have $215,892, while investor B will have $226,098. That 50 basis point difference in annual net returns resulted in a difference after 10 years of $10,206, or 4.7%. The lesson here is that the relatively long run horizon of many mutual fund investors combined with the effects of compounding suggest that a 50 basis point difference in net returns is an economically significant difference.

An alternate view of the economic significance of apparent small differences in net returns can be arrived at by looking at the size of the industry. At the end of 2005, the mutual fund industry managed approximately 9 trillion dollars.[36] One half of one percent (50 basis points) of 9 trillion dollars is 45 billion dollars.

Comments on Power in Existing Studies

This brings us to specific examples of empirical studies of the relationship between fund board chair independence and fund performance, fees, and compliance issues. The reader may wish to refer to the Office of Economic Analysis literature review to which this is a companion piece for details about these existing studies. However, there are two key points to be made about the data sets on which these studies are based.

---

[35] In this example we remain agnostic as to whether the hypothesized difference in net returns accrues from portfolio performance or fee differences. This example is not meant to imply that the true difference in net returns between management-chaired and independent-chaired funds is on the order of 50 basis points. Rather, 50 basis points represents a difference in returns smaller than any captured in the power curve analyses considered to this point. For the power curves shown and the sample sizes considered, a difference of 50 basis points would be detected with a probability of less than 20%.

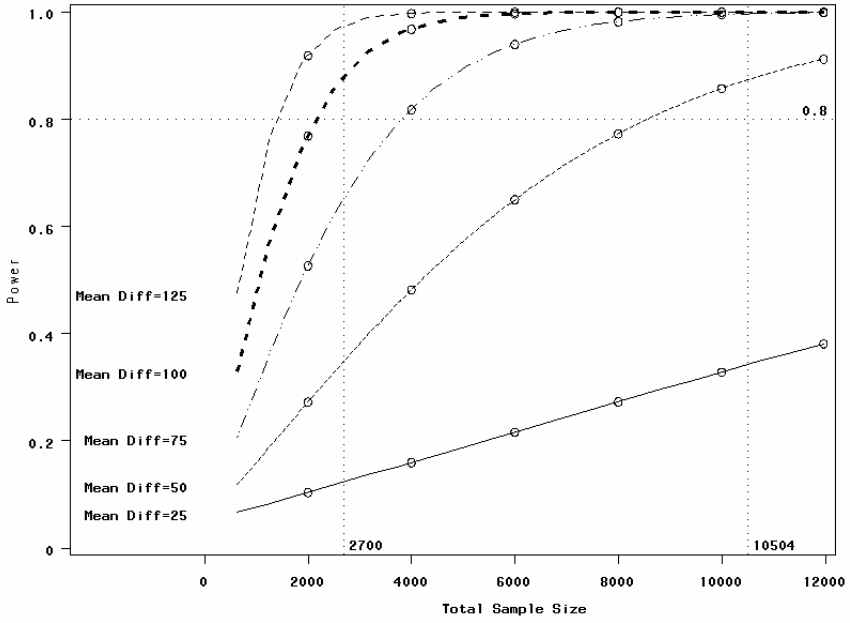
[36] 2006 Investment Company Fact Book, Investment Company Institute, Washington D.C., 2006, p.2.

- The data samples are limited in scale and scope.
- The signal-to-noise ratios are low.

As an example, consider a case where the cross-sectional standard deviation of abnormal returns across 10,504 funds is 6.6%. Using this data, regression results of abnormal performance on a collection of variables both including and excluding board characteristics result in an R^2 value of 0.13. Figure 4 below presents power curves for this scenario.[37]

The curves suggest that with a sample size of 10,504, this approach will have acceptable power (e.g. greater than 0.80) only if the 'true' difference in abnormal performance is roughly 50 basis points or greater. For example, if the 'true' difference between independent and management-chaired funds is 25 basis points per year, this relationship will appear statistically significant less than 40% of the time.[38]

Figure 4: Power Curves for a Test with 5% *Size* and Various 'True' Performance Differences, Given Standard Deviation of 6.6% and R^2 of 0.13



Empirical Limitations

There are a number of issues that suggest limitations in the application of common statistical approaches to identify differences between independent- and management-chaired mutual funds. These limitations derive primarily from the relatively small number of

[37] This example is taken from J. Meschke, "An Empirical Examination of Mutual Fund Boards," working paper, Carlson School of Business, University of Minnesota, Minneapolis, MN, 2006. We note that an R^2 of 0.50 is unusually high in modeling returns.

[38] In fact, Meschke's data collection process may result in lower power than this prediction, because he appears to have applied a snapshot of governance data to historical returns.

independent-chaired mutual funds and the fact that many of these funds moved from management chairs to independent chairs only during the last few years. Not only does this limit the universe of available data to the point where, for small 'true' differences, high power tests may be difficult to design, but it calls into question any study that employs historical data without accounting for the date of transition from management to independent chair.

Conclusion

The purpose of this memorandum is to discuss the factors that affect the power of common statistical methods to identify small yet economically meaningful empirical relations in noisy data, with an application to empirical studies of mutual fund chair independence and performance, fees, and compliance. Specifically, we focus on potential performance differences between independent and management chaired funds, and how the issue of low power of the statistical tests used in existing empirical analyses affects the interpretation of results. We address this question as part of our analysis of the empirical evidence relating independent chairs to performance.

Power refers to the accuracy with which a statistical test can identify whether or not a hypothesized relation exists in the data. For instance, power tells us something about the likelihood that a given test—with a given amount of data—will be able to reliably identify a relation between independent fund chairs and performance, given the assumption that such a relation does exist.

Our discussion of statistical hypothesis testing highlights the fact that the power of a statistical test is largely a function of the sample size, the signal-to-noise ratio in the data employed and the magnitude of the relation being tested. Tests conducted with small data sets or data that is subject to a great deal of noise are less likely to identify a relation. Further, the smaller the magnitude of the true relation, the lower the power of the test. This implies the smaller the magnitude of the real difference in returns that might accrue to independent-chaired and management-chaired funds (should it exist), the less likely a statistical test will be to identify it reliably.

The issue of low power is particularly germane when considering the small set of studies relating mutual fund chair independence to performance. The few studies of which the staff of the Office of Economic Analysis is aware are characterized both by small sample sizes and low signal-to-noise ratios. As part of our analysis, we show, by example, how likely it is that the tests conducted in this literature are to reliably identify a relation between independent chairs and performance. We conclude from our analysis that the magnitude of the relationship would have to be quite large for these tests to have any reasonable likelihood of identifying such a relationship. If the true relation were only 1% (still an economically significant difference), our analysis suggests that a study would have identified the difference as statistically significant only one time in five.

Low statistical power implies that researchers are more likely to fail to reject the hypothesis that chair independence, and governance more generally, is not related to performance. Coming to such a conclusion, though, is not at all equivalent to concluding that no relation exists. We would caution readers of this literature against coming to any strong

conclusions about the nature of the relation between independent fund chairs and performance on the basis of the evidence currently available.

There are several approaches that may be available to researchers to potentially increase the power of statistical tests. These typically involve either increasing the sample size through expanding the cross-sectional or time-series dimensions of the data used, or increasing the signal-to-noise ratio. Sample size may be increased by including all mutual funds over multiple years. Researchers may decrease the noise in returns by, for example, carefully accounting for all relevant determinants of fund performance. An alternative method to increase the signal may be careful construction of a test to identify instances where differences in returns will be greatest, such as in Khorana, Tufano, and Wedge (2006).[39] Depending on the true magnitude of the difference, however, these approaches may not result in sufficiently powerful tests.

[39] See Khorana, A., P. Tufano and L. Wedge, "Board Structure, Mergers and Shareholder Wealth: A Study of the Mutual Fund Industry," Working paper, Harvard Business School, 2006.