A First Look At The Accuracy Of The CRSP Mutual Fund Database And A Comparison Of The CRSP And Morningstar Mutual Fund Databases

by

Edwin J. Elton*
Martin J. Gruber*
Christopher R. Blake**

First Draft: June 2000

*Nomora Professors of Finance, Stern School of Business, New York University
**Associate Professor of Finance, Graduate School of Business, Fordham University

In recent years there has been an enormous increase in the number of mutual fund studies. There is hardly a professional meeting without at least one session devoted to the topic.[1] One of the major driving forces behind this increase in research is the availability of large computer-readable databases on fund characteristics and fund returns. The most widely used mutual fund databases in recent studies are those provided by the Center for Research in Security Prices (CRSP) and Morningstar[2].

While Morningstar has provided mutual fund data for some time, the CRSP mutual fund database is more recent. CRSP has constructed a mutual fund database that is likely to challenge Morningstar's as the source of fund data for academic research. The advantage of Morningstar's database is that it contains additional data on mutual funds not found in the CRSP database. The major advantage of the CRSP database is that it contains information on funds that have ceased to exist and thus allows correction for survivorship bias.

Despite the excellent job that was done in constructing the CRSP mutual fund database, like any new database it is not free of errors. The purpose of this article is to examine the potential errors in the CRSP database and to compare the CRSP return data with Morningstar return data. It is particularly appropriate for us to do so, since we constructed the first survivorship-bias-free databases of mutual fund monthly returns, and thus we have a data set which we can compare to the CRSP database.[3]

---

[1] For example, the AFA meetings have had at least one such session in each of the last five years.

[2] For example, Chevalier and Ellison (1999), and Hsiu-Lang Chen (2000) use Morningstar as a database while Wermers (2000), Zheng (1999) and Klaas Baks (2000) use CRSP.

[3] See Elton, Gruber, Das and Hlava (1993), Blake, Elton and Gruber (1993) and Elton, Gruber and Blake (1996a). The latter articles construct a database that is particularly appropriate to compare with the CRSP database.

Before doing so, however, a few comments are in order. First, although fund total returns can be large, properly adjusted performance measures typically are small. Annual risk-adjusted performance (alphas), properly measured, are on the order of minus 70 basis points per year (see Elton, Gruber and Blake (1996a) or Gruber (1996)). Furthermore, cross-category differences (e.g., aggressive growth versus growth) are much smaller still (see Elton, Gruber, Das and Hlavka (1993)). Finally, the evidence on mutual fund predictability is based on small differences (see Elton, Gruber and Blake (1996b) or Gruber (1996)). Thus, inaccurate data that produce small differences in alpha can lead to incorrect inferences.

All data sets have errors. The types of errors that are most harmful are systematic errors that cause biases. The presence of these biases in the CRSP database will be the subject of the first two sections of this paper. In the first section we will show that the returns in the CRSP database are upward biased in any month where there are multiple distributions on the same day. Often when a fund makes a capital gain distribution, there is also a dividend distribution on the same day, so overestimated returns are a persistent feature of the CRSP database. At the end of the first section we show how this can be corrected. In the second section, we restate a well-known feature of the Morningstar database: it has survivorship bias. However, we also show that the CRSP database, which does not have survivorship bias, does have an *omission* bias that causes the same type of problems as does survivorship bias. We discuss how to control for this bias. One of the nice features of the CRSP database is that it contains detailed tables on the dates of mergers and liquidations and, for merged funds, the name of the fund merged into (the surviving fund). We show that the name of the surviving fund is accurate, but that the merger and liquidation dates are often very inaccurate. Furthermore, the CRSP monthly returns table often does not contain monthly returns for several months before the merger

month. We discuss whether this is a problem. Finally, in the last section we compare the monthly fund returns listed by Morningstar with those listed by CRSP. We find a large number of differences in monthly returns and large differences in measures of risk-adjusted returns (alphas). Thus, the results of any study might differ depending on the database used. We discuss rules to determine which returns to cross check in order to eliminate differences in alpha across databases.

I. Upward-Biased Monthly Returns in the CRSP Mutual Funds Database

While CRSP has done an excellent job of reporting dividend payments and capital gains distributions in its mutual fund database, there is a systematic bias in CRSP's calculation of fund returns. Because of the formula CRSP uses to adjust returns for distributions, there is a consistent overstatement of returns for any period in which more than one distribution occurs on the same day. In this section, we review CRSP's formula for computing returns and show how this error occurs. We illustrate the error with a calculation using data from the CRSP mutual fund database. Once understood, the error can be easily corrected. In order to understand the impact of the bias in already published studies, and to understand the need to correct it in any future studies, we correct the data for a sample of funds from the CRSP returns file and measure the impact of the correction on return and mutual fund performance. In this section we employ monthly returns, for most performance studies are done using monthly returns.

The formula CRSP uses for calculating monthly returns (actually returns over any period) in its mutual fund returns database is presented in the *Survivor Bias Free U.S. Mutual Fund Data Base File Guide*, published by The Center for Research in Security

Prices. The formula is given on page 18 of the 1998 guide (version 1998.1.0). If there are no splits in a given period, the formula is:

$$R_{t-1,t} = \left(\frac{NAV_t}{NAV_{t-1}} \right) \left(\prod_{j=1}^{J} \left(1 + \frac{X\_AMT_j^D}{RE\_NAV_j^D} \right) \right) - 1$$

where

$R_{t-1,t}$ is the return on a fund between time $t-1$ and time t;

NAV_{t-1} is the fund's net asset value (NAV) at the end of the previous period;

NAV_t is the fund's NAV at the end of the current period;

J is the number of dividend or capital gains distributions during the period;

$X\_AMT_j^D$ is the jth dividend or capital gains distribution during the period, in dollars;

$RE\_NAV_j^D$ is the NAV at which the jth dividend or capital gains distribution was reinvested.

While the above formula works perfectly for distributions that occur on different days, it overstates returns for any period with more than one distribution occurring on the same day. It is not uncommon for a fund to pay a dividend and capital gain on the same day. When this occurs, the formula assumes that a capital gain can be used to purchase shares and that the dividend is received on the old shares plus the new shares purchased by the capital gain. But this cannot occur when the two payments are received simultaneously. This can best be illustrated with an example from the CRSP mutual fund database. For the month of December 1994, CRSP lists two distributions for the Vanguard Windsor Fund both occurring on December 14, 1994: a dividend of 24 cents per share and a capital gain of 86 cents per share. The reinvestment NAV reported by CRSP for each of those distributions is $12.53. CRSP lists the Vanguard Windsor Fund's

NAV at the end of November 1994 as $13.71 and the fund's NAV at the end of December 1994 as $12.59[4]. The fund's December 1994 return using the CRSP formula is

$$R_{\text{Dec. 1994}} = 12.59/13.71 \times (1 + (0.24/12.53)(1 + 0.86/12.53) - 1 = 0.00013$$

or 0.013%. This is also the (rounded) amount shown in CRSP for the Vanguard Windsor Fund's December 1994 return.

However, realizing that the two cash payments are paid at the same time, so that one does not receive a dividend on the shares purchased with the capital gain, the correct reinvestment assumption is a single payment of $1.10 reinvested in shares at an NAV of $12.53. The correct computation of return is therefore:

$$R_{\text{Dec. 1994}} = 12.59/13.71 \times (1 + 1.10/12.53) - 1 = -0.001075$$

or – 0.1075%. This amount is in fact the return reported by Morningstar for the Vanguard Windsor Fund in December 1994, and it is also the actual December 1994 return received by holders of the that fund[5].

In summary, unless the CRSP formula for return is corrected for payments on the same day, it will overstate the return for those months where two returns occur on the same day by $\frac{NAV_t}{NAV_{t-1}} \times \left\lceil \frac{A}{C} \times \frac{B}{C} \right\rceil$, where A and B represent the amounts of the two payments and C represents the reinvestment NAV. The example presented above, while

[4] The dividend, capital gain and reinvestment NAV are from the CRSP distribution table. The month end NAV's are from the monthly net asset values/share table.

[5] Morningstar rounds their returns.

actual, was selected for dramatic effect: the correction turned the return from positive to negative. The question we examine now is the importance of this bias.[6]

The first issue we examine is how often funds distribute capital gains and dividends on the same day. CRSP lists distribution data for 14,099 funds in its distributions table.[7] Of these funds a significant number, 6,731, or 47.7%, had at least one month in which multiple distributions excluding splits were reported on the same day.

We selected the last five years (1994-1998) for which we had CRSP data to study more intensely the bias in the CRSP method of handling multiple distributions. We selected from the CRSP database the 25 largest funds in terms of total net assets as of December 1993 in each of five ICDI fund objectives. The objectives selected were Aggressive Growth (AG), Income (IN), Growth and Income (GI), Long-Term Growth (LG), and Total Return (TR). The largest funds were selected because those were the funds whose records were least likely to contain random data errors.

Since our sample contains 125 funds for 60 months, there are 7,500 observations. Out of those observations, 551, or 7.3%, have two or more payments on the same date. The pattern of multiple payments differs across types of funds and months. In Panel A of Table 1 we show, for each type of fund, the number of monthly observations (out of 1,500 observations per type) with same-day multiple payments, as well as the percentage of months with same-day multiple payments. Note that the aggressive growth group has the lowest occurrence of simultaneous payments, while the growth and income group has

[6] The CRSP stock return databases do not contain this problem. It is not an issue for the CRSP daily stock returns file, because CRSP computes the daily returns as we suggest. The CRSP monthly stock returns file assumes all distributions occur at the end of the month, and ignores intra-monthly reinvestment.

[7] There are a total of 15,554 funds listed in the "Funds List" table of the 1998 CRSP mutual funds database, version 1998.1.0; however, some of those funds do not appear in the "Distributions" table of that database. (All CRSP mutual fund data used in this study are from version 1998.1.0 of the CRSP mutual funds database.)

the highest. Why does this occur? Recall that multiple payments on the same date almost always involve the payment of both a dividend and a capital gain distribution. The aggressive growth funds have fewer dividend payments even though they pay capital gains more often than other groups, while the growth and income funds, because of their dual objectives, often have both dividend payments and capital gain payments on the same day.

Panel B of Table 1 tells an even stronger story. Over 75% of the errors due to same-day multiple payments occur in the month of December. December is the standard month for most funds to pay capital gains because of the uncertainty of the capital gains earned in any year until the end of the year, and because of the legal requirement to pay out almost all of the net realized capital gains in order to be taxed as an investment company. Notice that other months which have multiple payments tend to be the end of the calendar quarters. These results strongly indicate that the use of the CRSP mutual fund data will introduce a seasonal influence into mutual fund data even if none exists.

We now turn to the major part of this section: how important is this bias in the CRSP data. Examining the difference in return caused by this bias, we find that for the months in which the problem arises it results in an overstatement of return by 4.2 basis points per month, or about 50 basis points yearly. If we look across all months for all funds in our sample, it amounts to 0.31 basis points per month or about 3.7 basis points per year. While this is a small difference in average return, the more important question to answer is how much does it affect the risk-adjusted performance as computed by most industry and academic studies?

This presents us with the problem of selecting a model to compute risk-adjusted performance. We use a model that we have employed in our previous research:

$$R_{it} = \alpha_i + \beta_{i1} I_{1t} + \beta_{i2} I_{2t} + \beta_{i3} I_{3t} + \beta_{i4} I_{4t} + \varepsilon_{it} \qquad (1)$$

where

R_{it} is the excess total return (total return net of the 30-day T-bill return) for fund *i* in month *t*;

α_i is the alpha for fund *i*, the risk-adjusted performance measure;

β_{ik} is the sensitivity of fund *i*'s excess return to index *k*;

I_{1t} is the excess total return of the S&P 500 in month *t*;

I_{2t} is the excess total return of the Lehman Aggregate Bond Index in month *t*;

I_{3t} is the return on a small-cap portfolio minus the return on a large-cap stock portfolio based on Prudential Bache indexes in month *t*;[8]

I_{4t} is the return on a growth stock portfolio minus the return on a value stock portfolio based on Prudential Bache indexes in month *t*;

ε_{it} is the random error for fund *i* in month *t*.

When equation (1) is used to estimate the average alpha for the 125 funds in our sample using uncorrected CRSP return data, the average monthly alpha is –0.0406 or –4.06 basis points per month. When the average alpha is computed from re-running the four-index model on the corrected data, the average monthly alpha is –0.0438 or –4.38 basis points. The difference of 0.32 basis points per month or 3.84 basis point per year means that the average alpha is overstated by 7.8%.[9] However, the error is more important than the average would imply. Out of the 125-fund sample, the five funds with the largest error in alpha had an error of 21.61 basis point per year, while for the top ten funds the average error was 16.40 basis points and for the top 25 funds it was 11.07 basis

[8] See Elton, Gruber and Blake (1996b) for a detailed description of the Prudential Bache indexes used in the model.

[9] If the results are obtained using a one-index model employing the excess returns of the S&P 500 index, the difference in alpha is 0.30 basis points per month rather than the difference of 0.32 basis points that occurs with the four-index model.

points. Errors of this size are important in evaluating the performance of individual funds and have the potential of producing predictability when none exists since the same funds tend to have multiple distributions (and hence overstated returns) over time.

Should we care about correcting this error? After all, the error in return only occurs 7.3% of the time and when it occurs it only averages 0.31 basis points per month. We believe the answers is "yes," for the error introduces a clear bias. Any fund which pays out capital gains and dividends at the same time has upward-biased returns. The amount of upward bias differs from one type of fund to another. Where an appropriate risk model is introduced so that risk-adjusted returns can be measured, we see that the bias can make a large difference. The average difference in alpha is 7.8%. Furthermore, if we look at the 10 funds for which the error is largest out of the 125-fund sample, we see that the error is 16.4 basis points per year or a percentage error of 21.8%. This difference is large enough that it can cause us to draw incorrect conclusions about individual funds, can change the ranking of funds, and might impact studies of performance evaluation and predictability. However, it is small enough that we probably do not want to disregard all prior studies which do not correct for it.

II. Omission and Survivorship Bias

Survivorship bias comes about when databases don't include funds that merge or liquidate and when those funds have characteristics different from funds that survive. For example, if one created a database consisting solely of funds that existed at the end of 1990 with ten years of prior history, then funds that existed in 1980 and merged or liquidated before the end of 1990 would be excluded from the database. If these funds had different performance characteristics, then the database would have survivorship bias.

The Morningstar database has survivorship bias.[10] As shown in Elton, Gruber and Blake (1996a), this causes overall performance measures to be inflated by between 40 basis points and 1%, depending on the length of the sample period used in the study. This bias is sufficiently large, given the underperformance usually observed, that a sample of funds with survivorship bias can appear to have a positive average alpha when the true average performance is negative. In addition, since survivorship bias affects funds with different investment objectives by different amounts, one could inaccurately conclude that funds with different objectives had different levels of performance when in fact they performed the same.[11] Finally, survivorship bias can lead to an appearance of predictability when none is present (see Brown, Goetzman, Ibbotson & Ross (1992)).

While the CRSP database is free of *survivorship* bias, it has an *omission* bias that has a similar effect on results for anyone using its monthly return data. The return data on the CRSP files is monthly for some funds and annual for others. The merger and liquidation rates are very different for funds that have monthly CRSP return data than they are for funds with annual CRSP return data. Thus researchers using monthly CRSP fund return data will have a sample that understates the proportion of mergers and liquidations and thus overstates performance, inaccurately measures differences across funds with different objectives, and may demonstrate predictability where none exists. In short, a set of data that exhibits all the problems of a sample with survivorship bias.

To examine the magnitude of this problem we will use the survivorship-bias-free sample that we constructed in Elton, Gruber and Blake (1996a) (EGB). This sample contains all U.S. equity funds that listed "common stock" as their objective in the 1977

[10] A survivorship bias free database can be constructed by getting all past Morningstar quarterly distributed data and tracing through all funds.

[11] See Elton, Gruber, Das and Hlavka (1993) for an analysis of the difference in performance across different classifications.

annual edition of Wiesenberger's *Investment Companies* (which lists year-end 1976 data) and that listed total net assets of $15 million or greater.[12] Table 2 shows the number of common stock mutual funds listed in the 1998 CRSP mutual funds database that existed at the end of 1976, classified in two groups by year-end 1976 total net asset value (under $15 million and $15 million or greater).[13] The table also shows whether complete monthly return data, partial monthly return data, or no monthly return data exist in the CRSP database.[14] All but two funds of the funds listed in Table 2 with $15 million or more in total net assets have complete monthly return data in the CRSP database.[15] CRSP's assessment of which funds merged or liquidated for funds with total net assets

[12] One of the missing pieces of data in CRSP is information about which funds are restricted and who can purchase them. For example, one is restricted to Lutheran ministers and another to GE employees. There are many such funds, see Elton, Gruber and Blake (1996a). Any researcher studying the profitability of trading rules such as those contained in predictability studies will need to control for this.

[13] CRSP lists 207 common stock funds with year-end 1976 total net assets of $15 million or greater, which is the same count we documented in Elton, Gruber and Blake (1996a). However, in our sample, we dropped 19 of those funds which restricted the type of individual who could own the fund, were closed to new investors, or were variable annuity funds. This left us with a remaining sample of 188 large common stock funds. In table 2 and in this analysis, we exclude those 19 funds from the CRSP sample as well, so that we may compare results from our data set with those from the CRSP data set.

[14] CRSP has a count of 155 small common stock funds rather than the count of 154 funds that appears in Table 2. CRSP misclassified one fund. Wiesenberger was the source both we and CRSP used to get year-end 1976 investment objectives. This fund is classified by Wiesenberger as a "specialized" fund (Weisenberger code "Spec") in the *Investment Companies* editions surrounding 1977; in the 1977 edition there is a typographical error, and the fund's code is listed simply as "S," a code that has no Wiesenberger definition.

[15] Those two funds have missing CRSP return data for a few months in the middle of their history. For a fund missing return data in the middle of its history, CRSP records in the first month after the missing month(s), the total return over the missing month(s), and that month. For example, if there is a single missing month, the return given in the next month is a two-month return. Those two funds are included with the funds for which complete data was supplied since only a few months of data were missing and the data was readily available from other sources.

over $15 million closely agrees with our assessment. CRSP shows 43 funds merged and one liquidated; we show 42 merged and none liquidated.[16]

The reason we did not include common stock funds with total net assets less than $15 million in our survivorship-bias-free returns sample is that most of those funds were not listed by NASDAQ at the beginning of our sample period, and NASDAQ was the source of monthly fund returns provided to data vendors in the early years of our sample. Thus, for many defunct small funds, monthly returns can not be found in any source. Table 2 classifies the funds into three groups: funds for which CRSP reports complete monthly returns, those for which it reports monthly returns only part of the time, and those for which it reports only annual returns (no monthly returns) or no returns at all. Table 2 also shows the funds that merged or liquidated in each of these categories. Funds with no monthly returns are clearly excluded from any performance study employing monthly data. The funds that are less than 15 million in size with partial monthly data are likely to be excluded, for six of those funds have five or less years of data, and unless the starting date of the monthly returns for those funds happened to coincide with the beginning date of the study, the fund would have less usable data. However, to be conservative in estimating omission bias, we will first exclude and then include the small funds with partial monthly returns from our sample of funds with monthly data.

We now estimate the bias due to the differential impact of mergers and liquidations on mean alpha from omitting funds for which CRSP does not include monthly data. In EGB (1996a) for the large-fund group examined here, we calculated an average yearly alpha of −0.1269% for surviving funds and −2.8779% for non-surviving funds. Using only those funds with complete monthly data from the CRSP data source,

[16] CRSP classifies one fund as liquidated that we had tracked (from the fund's investment company) as a name change. See section III for further discussion.

one would have a sample of 281 funds and could observe which of those funds did not survive.[17] The alpha one would obtain is:

$(215 / 281) \times (-0.1269) + (66 / 281) \times (-77.3) = -0.773$, or −77.3 basis points.

The total number of funds, including funds with no monthly data, is 342 funds of which 125 merge or liquidate. Thus the population alpha is:

$(217/342) \times -0.1269 + (125 / 342) \times -1.132 = -1.132$, or −113.2 basis points.

This gives a bias of 36 basis points. If those funds with partial monthly data are included in the sample with monthly data, then the bias is reduced to 30 basis points.

A second way to estimate the magnitude of the omission bias is to look at the differential performance for those funds for which CRSP reports annual returns (but not monthly returns) compared to the funds for which CRSP reports monthly returns. Since funds with only annual returns in the database tend to exist for a small number of years, using a time series to adjust for risk is not possible. All we can do is compare unadjusted annual returns. We have nine years in which we have at least five funds with only annual data. In each of the nine years, the average annual return for the funds with monthly data in the small fund group was higher than the average annual return for those with annual data. The difference from year to year in average annual return ranged from 2.1% to 9.8%, with an overall average difference of 6.1%.[18] If we use these figures to estimate omission bias, recognize that 60 funds had no data, annual data, or a very small number of monthly observations, and assume that the difference in return would be reflected in difference in alpha (same average risk for annual and monthly return samples of small

[17] We include, in the CRSP data set of complete monthly returns data, two funds with more than $15 million in assets and some missing monthly returns, since the data could be and was found from other sources. Those funds with missing monthly returns and total net assets under $15 million have monthly returns data that is much sparser, and we could find no way of filling in the missing data.

[18] The 6.1% difference seems larger than we would expect. One possible explanation is that data for funds that were more successful and have monthly data were back filled.

funds), then omission bias is 60 / 339 × 6.1 = 1.08% or 108 basis points. This is considerably higher than our prior estimate, and probably overstates omission bias because it doesn't correct for risk and doesn't consider what strategy an investor might follow after a fund ceases to exist. While the second method shows omission bias is important, we believe the first method for estimating it is more accurate. Elton, Gruber and Blake (1996a) estimated survivorship bias, using methodology similar to the first of these techniques, to be 90.6 basis points. If we use our best estimate of omission bias, 36 basis points, then omission bias is about 40% as large as survivorship bias. Studies using CRSP monthly data for all funds in the CRSP database still have a bias sufficient enough to have a serious effect on mean alpha, and one may well find predictability where none exists.

An easy way to avoid this problem is to restrict the sample of funds studied to contain only those funds that have over $15 million in total net assets at the beginning of any observation period. CRSP has monthly data in all months for most funds with over $15 million in total net assets, since these funds report data to NASDAQ. This does not introduce a bias, since size (total net assets over $15 million) is known before fund performance is studied. On the other hand, the results from such a study only apply to investors who consider funds with over $15 million in total net assets. If this is not appropriate for the purpose of a particular research project, then either more data must be found or techniques such as those employed above must be used to correct for omission bias.

III. Merge Data

CRSP supplies for funds that merged or liquidated the name of the fund that the original fund merged into (the surviving fund), the date of the merger or liquidation, and

monthly returns for the merged or liquidated fund, stopping before or at the time of the merger or liquidation. How accurate is this data, and what are the problems associated with it? To analyze this we examined from EGB (1996a), the funds that existed in 1977 and were over $15 million in size. In that study, we show 42 funds that merged. Since in that study we were interested in “following the money”, for each fund that merged we had to obtain from the surviving fund's investment company the exact date of the merger and the merger terms.

CRSP has fairly accurate data on which funds merge and on the names of the merge partner funds. Of the 188 funds in our 1977-1993 sample, CRSP shows that 42 funds merged and one fund liquidated; we found 42 merged and none liquidated. The difference occurs because the fund CRSP classifies as liquidating instead simply experienced a name change on that date. In addition, of the 42 merged funds we have in common with CRSP, we agree on 40 of the names of the surviving funds. In one case where we disagree, CRSP classifies the fund as merging in a month where we show a name change. That fund subsequently merged at a later date; hence, CRSP misses the subsequent merger. In the second case, CRSP does not list a name for the surviving fund. In short, of the 43 funds that CRSP classifies as merging or liquidating, 40 were accurately classified by CRSP, in terms of both the merge event and the name of the surviving fund. CRSP is less accurate about the dates of the mergers and recording return data for all months until the merger dates.

Table 3 shows a breakdown of how CRSP merger dates compares with the EGB merger dates. Why are we confident that the EGB merger dates are accurate? First, because we used a "follow the money" approach in calculating returns, we had to contact the surviving fund's management company to obtain merger terms and dates of the merger for each of the 42 mergers, and we were able to obtain returns consistent with

these dates. Second, all of CRSP merged fund returns end before our merger dates. Furthermore, for four of the eight funds where CRSP says they merge later, the CRSP return data is consistent with our merge date and in the one case where they show an earlier merge date we were able to obtain returns for the fund after they say it merged. How accurate is the CRSP merge date? For five of seven funds that merged on the last day of the month, CRSP shows the correct date, while for the remaining two funds CRSP reports the date as the first day of the next month or the last day of the prior month. For the 22 mid-month mergers, CRSP shows ten occurring on the first day of the merger month, eleven occurring on the last day of the prior month, and one on a different day during the correct month. Finally, for nine cases CRSP reports a date more than one month from the actual merger date, and in four cases CRSP reports a merger without listing a merger date at all. In five of the nine cases where CRSP merger dates differ by more than one month from EGB merger dates, the availability of monthly return data in the CRSP database is consistent with the merger date reported in EGB and inconsistent with the merger date reported by CRSP. In the other four cases, return data in the CRSP database stops before EGB states the fund merged, even though in three of these cases CRSP shows a merger date many months later than EGB. The rule that would give a researcher using the CRSP database the highest degree of accuracy in identifying the merger month is to assume that mergers that are shown to occur on the last day of the month occur in the next month, and those shown at the beginning of a month and middle of the month occur in that month. This would have resulted in correctly identifying the merger or liquidation month in 24 of 43 cases including the mis-classified fund.

As shown in Table 4, the month the return data ends for a merged fund in the CRSP database also has a high error count if the intent is to have the data complete up to

the merger month.[19] If one accepts the merger dates used in EGB, then in 32 of 42 cases CRSP data would be consistent with the rule of "show return data in the month of merger" if the fund merges at the end of the month and "show data to the month prior to the merger" if it merges before the end of the month. This is a sensible and consistent way to collect returns. In the ten remaining cases, CRSP ends returns some months before the time of the merger. The number of months that CRSP stops early are 1, 1, 4, 4, 5, 7, 12, 12, 23 and 36.

If one accepts the CRSP date of merger as correct, then there are 23 of 38 cases where CRSP return data are consistent with the rule mentioned earlier.[20] In 14 of the 15 cases where CRSP does not follow the rule, the monthly returns end earlier, ranging from one month to two years[21]. In the remaining case, CRSP has data after the CRSP merger date. In four cases, CRSP does not show a merger date. Thus, using the CRSP merger date or CRSP monthly return data to infer merger months is accurate only slightly more than half the time.

There are two issues that are worth further examination. What is the impact on alpha of excluding return data for months prior to the merger? What is the impact of excluding the merger month?[22]

If we accept the EGB merger dates, we have return data for ten funds where CRSP stopped reporting the data one or more months before the beginning of the merger

[19] A lot of other possibilities would be reasonable. For example, a researcher could assume investors leave the fund when a merger is announced. However, for some funds, the CRSP data include returns in the merger month, so that this alternative rule was not followed. The rule stated in the text seems to be the most consistent with the CRSP data.

[20] This number excludes one firm that was mis-classified as liquidated. This number is the sum of A1, B1, B2, and B4 from Table 4.

[21] The sum of A3, B3, C1, C2, and C3 from Table 4.

month (Table 4, Group C). For nine of these funds CRSP simply stopped reporting returns early; for one fund they stopped because they mis-classified a name change as a merger. For the nine funds where CRSP simply stopped reporting return data early, we used the CRSP return data and estimated the four-index model discussed earlier (equation (1)) using three years of data (corrected for multiple distribution on the same day) ending in the last month CRSP shows data. We then used the alpha and betas from the regression along with the appropriate index values and the EGB return data for each of these funds to calculate the average residual of the fund, starting with the month CRSP stopped reporting the fund's returns through to either the month just prior to where the merger occurred if it occurred mid-month or to the end of the month if the merger occurred at the end of the month. There was no systematic pattern. Roughly half of the average residuals across the nine funds were positive and half were negative. Cumulating across all nine funds and all months, and treating each observation equally, produced an average residual close to zero. However, the small number of observations and the sufficiently large standard error make it hard to draw definitive conclusions.

To examine the average risk-adjusted performance in the merger month, we ran the same four-index model using CRSP return data (corrected for multiple distributions) over all funds for which CRSP had complete monthly return data up to the month just before the merger (the 25 funds in Group B of Table 4). We then used the alphas and betas from the regression and the EGB spliced merge-month returns for those funds to calculate each fund's residual risk-adjusted return for the month of merger. Of the 25 funds, we dropped two funds with a few missing mid-series CRSP monthly returns and one fund that only had five monthly returns prior to merging. Of the remaining 22 funds,

[22] Since mergers need to receive shareholder approval, the merger month is known. Thus there is no bias in assuming the investor leaves the fund at the beginning of the merger month. Nevertheless, merger-month performance is interesting by itself.

13 had negative residuals in the merger month, and the average residual was a negative 87 basis points. The reader should view this result with caution because the betas may well be different for the surviving fund and the alphas are not statistically different from zero at the 5% level.

In summary, anyone studying mutual fund mergers should use CRSP merger data only as a starting point and to obtain the names of the merge partner funds. The CRSP data on merger dates are inaccurate enough to require that all merger dates be independently validated. For purposes of merged fund performance measurement, there is no evidence that the CRSP fund return data, which in many cases stops months before the merger, introduces a systematic bias. However the sample we used to draw that conclusion is sufficiently small that care should be exercised.

IV. Consistency of CRSP and Morningstar Data

As stated previously, we believe CRSP and Morningstar will be the most widely used databases for mutual fund research in the future. CRSP has a major advantage in that its database includes data on funds that merge and liquidate (albeit with many timing errors). The advantage of the Morningstar database is that it includes lots of data on composition and performance, and is widely used and quoted.

The question we examine in this section is, if funds are selected for which Morningstar and CRSP both have return data, are the data the same, and, if there are differences, can these differences lead to serious differences in performance measurement results? To answer this question we selected common stock funds from the Growth and Income group in the CRSP database that had over $15 million in total net assets in 1998 and that also had complete sets of monthly returns in both the CRSP and Morningstar databases for a 20-year period from January 1979 through December 1998. By doing so

we eliminated any differences due to Morningstar survivorship bias and CRSP omission bias. We corrected the CRSP sample for multiple distributions on the same day by using the procedure outlined earlier in this paper. Because we were interested in the effects of size, we studied the 25 largest non-index funds and the 25 smallest non-index funds with over $15 million in assets in the growth and income group. While we were interested in the entire 20-year period, we emphasize four five-year subperiods for two reasons. First, most studies of performance measurement use five years of monthly data in estimating performance. Second, by looking at four five-year subperiods, we can see if the data in the two databases are getting closer together over time.

Information about the difference in alphas from applying the four-index model (equation (1)) to CRSP and Morningstar data for each of the four five-year samples is presented in Table 5. First, note that across the four five-year samples the alphas are different for the two data sources in 99 out of 100 cases. Two facts are immediately apparent from Table 5. The differences in alphas using the two data sources are most serious in the first five-year period and are much more serious for small funds than they are for large funds. In the first five-year period for large funds the average difference in alpha amounts to about 16 basis points per year. For the sample of small funds it amounts to 61 basis points per year. Clearly these differences are important. For example, average underperformance of actively managed mutual funds using the model employed in this paper is about 70 basis points per year (see EGB (1996b) or Gruber (1996)). If one is studying mutual fund performance before the mid-1980's, differences in alpha are sufficiently large that conclusions might well be affected depending on whether one uses the Morningstar or CRSP databases. The average differences in later periods are much smaller. For the last five year period for large funds the average yearly difference drops to about 1.7 basis points, and for small funds it is about 3.1 basis points.

When studying the performance of individual funds, whether to identify good managers or to look at performance persistence, the differences in alphas across different data sources are more important. There are a number of large differences in alpha. For example, when examining the small funds, 34 out of the 100 observations have differences in alpha greater than 12 basis points per year, 14 have differences greater than 60 basis points per year, and nine have differences greater than 120 basis points per year. The five largest differences are 6.8%, 4.9%, 4.9%, 3.1% and 2.5% per year. There are fewer large differences for the large fund sample (18 greater than 12 basis points per year, 6 greater than 60 basis points, and 3 greater than 120 basis points). Furthermore, the largest differences are smaller; the five largest are 4.9%, 3.1%, 1.9%, 1.6% and 1.4%. For large funds, the big differences are all in the first five years. For small funds, big differences continue through the 20 years.

While most studies use a five-year period to measure alphas, some studies might use a longer period. To see if errors still remain important, we examined results from estimating equation (1) over the full 20-year period, a period which is longer than we would expect any researcher to employ. Over the 20-year period, over 10% of the alphas have errors greater than 60 basis points.[23]

Differences in alphas are important. They can change the conclusions about individual funds or a group of funds. Furthermore, any researcher using data that is more than 15 years old must be extremely careful about overall conclusions.

Obviously the differences we observe in alphas are caused by differences in the returns reported by CRSP and Morningstar. Table 6 presents summary data on the

[23] These results are not shown in Table 5.

differences in returns between the two databases[24]. The table makes it clear that the differences between the databases and hence their accuracy has gotten a lot better in recent years. This is true whether we base the judgment on the number of differences, the size of the average difference, the size of the average absolute difference, or the number of large differences. The other fact that shows up dramatically in Table 6 is that there is more of a problem with data in small funds than there is with large funds. Small funds tend to have more differences and larger differences.

While we have not investigated the cases where differences exist to see which data source is accurate, we have shown that there are a large enough number of cases of sufficient magnitude to be of concern to a researcher.

We now examine how large the differences in return have to be to significantly impact alphas. More specifically, we examine whether differences in alpha of greater than 0.01% per month could be identified from differences in return. This is roughly 12 basis points or more on an annual basis, a figure that clearly should be of concern to a researcher. When we examine all funds in any five-year period that have a difference in return for at least one month of more than 0.5%, we find that we correctly identify 51 out of the 52 five-year alphas that are greater than 0.01% per month. While we are very successful in identifying which funds will have large differences in alpha from examining differences in return data, there is a cost. This rule also identifies 22 funds out of 200 that should not be checked since they have a difference in alpha of less than 0.01%. However, as a screening device to identify those funds where a data problem might exist in estimating alpha, it seems very effective.

[24] Return differences are caused by differences in month-end NAV and differences in distributions. Differences in month-end NAV frequently cause return differences to be of opposite sign and roughly equal in magnitude in adjacent months.

V. Conclusion

The CRSP database is a fairly new publicly available database on mutual funds. It is comprehensive and is corrected for survivorship bias. It and the Morningstar database are likely to be the standard databases used by researchers in the future. Despite the care that has been exercised in compiling the CRSP database, it needs to be corrected for certain types of problems. The most obvious bias in the CRSP database is that it calculates fund returns for months with multiple distributions on the same day in a way that causes returns in those months to be overstated. This overstatement has an impact on overall returns and alphas which is of economic significance. The Morningstar database is free of this problem.

We have shown that while CRSP does not suffer from survivorship bias, it does suffer from omission bias. Because only some small funds under $15 million in total net assets have monthly data on the CRSP database, and because the omitted funds have much greater merge and liquidation rates, we show that the returns reported for that group of funds which have monthly data overstate the population returns and alphas. We then examine the data CRSP provides on mergers. While these data are quite good in identifying mergers, we show that there are major problems in merger dates and reporting return data up to the time of the merger.

Finally, after correcting for all of these influences, we compare the data in the CRSP database with the data in the Morningstar database. We examine differences in alphas and return over four five-year periods. There are many differences. The differences are most severe for the smallest funds. For all funds, the differences are larger as we go back in time. We develop a rule for differences in return that allows us to determine when differences in alpha are likely to arise.

References

Blake, C. R., E. J. Elton and M. J. Gruber, 1993 "The Performance of Bond Mutual Funds," *The Journal of Business*, 66, 371-403.

Brown, S. J., W. N. Goetzmann, R. G. Ibbotson, and S. A. Ross, 1992, "Survivorship Bias in Performance Studies," *The Review of Financial Studies*, 5, 553-580.

Carhart, M., 1997, "On The Persistence of Mutual Fund Performance," *The Journal of Finance*, 52, 57-82.

Chevalier, J. and G. Ellison, 1999, "Are Some Mutual Fund Managers Better than Others" Cross-Sectional Patterns in Behavior and Performace," *Journal of Finance*, 875-899.

CRSP, 1998. *CRSP Survivor Bias Free US Mutual Fund Data Base*, Version 1998.1.0, Center for Research in Security Prices, Graduate School of Business, The University of Chicago.

Elton, E. J., M. J. Gruber, C. R. Blake, 1999, "Common Factors in Active and Passive Portfolios," *The European Finance Review*, 3, 53-78.

Elton, E. J., M. J. Gruber, C. R. Blake, 1996a, "Survivorship Bias and Mutual Fund Performance," *The Review of Financial Studies*, 9, 1097-1120.

Elton, E. J., M. J. Gruber, C. R. Blake, 1996b, "The Persistence of Risk-Adjusted Mutual Fund Performance," *The Journal of Business*, 69, 133-157.

Elton, E. J., M. J. Gruber, S. Das, and M. Hlavka, 1993, "Efficiency with Costly Information: A Reinterpretation of Evidence from Managed Portfolios," *The Review of Financial Studies*, 6, 1-22.

Gruber, M. J., 1996, "Another Puzzle: The Growth in Actively Managed Mutual Funds," *The Journal of Finance*, 51, 783-810.

Morningstar, 1999, *Principia Pro Plus for Mutual Funds*, January 1999 CD, Morningstar, Inc., Chicago.

Wermers, R., 2000, "Mutual Fund Performace: An Empirical Decomposition Charts Stock Picking Talent, Style, Transactions Costs and Expenses," *Journal of Finance (*forthcoming).

Wiesenberger Financial Services, 1976-1978, *Investment Companies*, Warren, Gorham & Lamont, Inc., Boston.

Zheng, Lu, 1999, "Is Money Smart? A Study of Mutual Fund Investors' Fund Selection Ability," *Journal of Finance*, 54, 3, 905-935.

TABLE 1

Breakdown Of Multiple Distributions
This Table Analyzes Multiple Distributions Which Occur On The Same Date Over The Period January 1994 Through December 1998
The Distributions Are For The Largest 25 Funds Ranked By Total Net Assets As Of Year-End 1998 In Each Of Five Investment Objectives
(125 Total Funds; 1,500 Observations Per Investment Objective Group)

Panel A: Breakdown By Investment Objective

	Investment Objective				
	Aggressive Growth	Total Return	Income	Long-Term Growth	Growth and Income
Number Of Observations With Same-Date Multiple Distributions	61	105	106	119	160
Percentage Of Observations With Same-Date Multiple Distributions By Group	4.1%	7.0%	7.1%	7.9%	10.7%

Panel B: Breakdown By Month

	Jan.	Feb.	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Percentage Of All Observations Where Same-Date Multiple Distributions Occur In Each Calendar Month	0.0%	0.9%	6.4%	0.4%	2.9%	2.7%	2.5%	1.1%	3.6%	0.0%	3.6%	75.9%

Notes:
Within each group, top 25 funds do not include index funds or multiple classes of the same fund.

All CRSP fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

TABLE 2

Merger Experience and CRSP Return Data Availability
(January 1977 - December 1993)

	Year-End 1976 Total Net Assets			
	$15 Million Or More		Less Than $15 Million	
	All Funds [a]	Merged/Liquidated [b]	All Funds [c]	Merged/Liquidated [d]
Total Funds	188	42	154	83
Funds With Complete Monthly Returns [e]	188 [f]	42 [f]	93	24
Funds With Some Monthly Returns	0	0	10	9
Funds With No Monthly Returns	0	0	51 [g]	50 [g]

Notes:

[a] Excludes 19 funds that were restricted, closed to new investors, or variable annuity funds.

[b] Numbers shown are from the 188 "All Funds" group and exclude one fund that CRSP shows as a liquidation but that we tracked as a name change.

[c] Excludes one fund CRSP incorrectly categorizes as a common stock fund.

[d] Numbers shown are from the 154 "All Funds" group.

[e] Includes funds with complete monthly returns up to exit month in CRSP monthly returns file.

[f] Includes two funds that were missing a few mid-series CRSP monthly returns that were obtainable from other sources.

[g] Includes five funds where CRSP is unsure of what happened to the fund.

All fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

TABLE 3

CRSP Merger Dates Relative To Actual Merger Dates

Status	Occurrences
A. CRSP Date Differs By One Month Or Less From Actual Date	**29**
1. Actual Merger Date On Last Day Of Month	7
a. Identical To Actual Merger Date	5
b. First Day Of Month After Actual Merger Date	1
c. Last Day Of Month Prior To Actual Merger Date	1
2. Actual Merger Date In Mid Month	22
a. First Day Of Actual Merger Month	10
b. Last Day Of Month Prior To Actual Merger Month	11
c. Different Mid-Month Date	1
B. CRSP Date Differs By More Than One Month From Actual Date	**9**
1. CRSP Date After Actual Date	8
2. CRSP Date Before Actual Date	1
C. CRSP Date Not Given	**4**

Note:
All CRSP fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

TABLE 4

CRSP Monthly Return Ending Month Relative To Actual Merger Mont

Status	Occurrences
A. CRSP Reports Returns For The Actual Merger Month And Fund Merges End Of Mont	**7**
1. CRSP Merger Date Identical To Actual Merger Date	5
2. CRSP Merger Date One Month Before Actual Merger Month	1
3. CRSP Merger Date More Than One Month After Actual Merger Month	1
B. CRSP Returns End One Month Before Actual Merger Month And Fund Merges Mid Mont	**25**
1. CRSP Merger Date At End Of Month Prior To Actual Merger Month	11
2. CRSP Merger Date At beginning Of Actual Merger Month	6
3. CRSP Merger Date More Than One Month After Actual Merger Month	4
4. CRSP Merger Date Is Mid Month In Actual Merger Month	1
5. CRSP Merger Date Not Listed	3
C. CRSP Returns End More Than One Month Before Actual Merger Dat	**10**
1. CRSP Merge Month Identical To Actual Merge Month	4
2. CRSP Merge Month Later Than Actual Merge Month	4
3. CRSP Misclassifies Name Change As Merger	1
4. CRSP Merger Date Not Listed	1
D. CRSP Returns End Due To Misclassification Of Name Change As Liquidation And Fund Surviv	**1**

Note:
All CRSP fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

TABLE 5

Differences In Monthly Alphas Estimated From Four-Index Model (Equation (1)) Using CRSP and Morningstar Monthly Return Data (All Values Expressed In Basis Points)

Large Funds

Sample Period	Number Of Funds With Non-Zero Difference	Avg. Difference[a]	Avg. Absolute Difference	Number Of Differences Greater Than Or Equal To: 10 Basis Points	5 Basis Points	1 Basis Point
1979 - 1983	25	–1.34	4.84	3	6	11
1984 - 1988	25	0.09	0.60	0	0	3
1989 - 1993	25	–0.12	0.21	0	0	1
1994 - 1998	24	–0.14	0.26	0	0	3
Overall	99	–0.38	1.48	3	6	18

Small Funds

Sample Period	Number Of Funds With Non-Zero Difference	Avg. Difference[a]	Avg. Absolute Difference	Number Of Differences Greater Than Or Equal To: 10 Basis Points	5 Basis Points	1 Basis Point
1979 - 1983	25	–5.08	7.71	5	7	14
1984 - 1988	25	–0.11	2.32	2	3	9
1989 - 1993	25	–0.81	1.20	0	2	7
1994 - 1998	24	–0.26	1.29	2	2	4
Overall	99	–1.56	3.13	9	14	34

Notes:

[a] Differences measured as alpha using CRSP data minus alpha using Morningstar data

All CRSP fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

All Morningstar fund data obtained from Principia Pro January 1999 CD.

TABLE 6

Differences In Monthly Total Returns Using CRSP and Morningstar Monthly Return Data
(All Values Expressed In Percent)

Large Funds

Sample Period	Number Of Months With Non-Zero Difference	Avg. Difference[a]	Avg. Absolute Difference	Number Of Differences Greater Than Or Equal To: 5.0%	1.0%	0.5%
1979 - 1983	532	0.001	0.151	14	44	59
1984 - 1988	421	–0.002	0.030	0	8	19
1989 - 1993	297	0.000	0.015	0	2	6
1994 - 1998	185	0.000	0.009	0	4	7
Overall	1435	0.000	0.052	14	58	91

Small Funds

Sample Period	Number Of Months With Non-Zero Difference	Avg. Difference[a]	Avg. Absolute Difference	Number Of Differences Greater Than Or Equal To: 5.0%	1.0%	0.5%
1979 - 1983	639	–0.030	0.280	20	91	145
1984 - 1988	473	–0.004	0.122	8	31	57
1989 - 1993	231	–0.007	0.029	0	13	28
1994 - 1998	190	0.002	0.013	1	3	7
Overall	1533	–0.010	0.111	29	138	237

Notes:

[a] Differences measured as return using CRSP data minus return using Morningstar data.

All CRSP fund data obtained from CRSP Mutual Funds Database, Version 1998.1.0.

All Morningstar fund data obtained from Principia Pro January 1999 CD.