A First Look at the Accuracy of the CRSP Mutual Fund Database and a Comparison of the CRSP and Morningstar Mutual Fund Databases

EDWIN J. ELTON, MARTIN J. GRUBER, and CHRISTOPHER R. BLAKE*

ABSTRACT

This paper examines problems in the *CRSP Survivor Bias Free U.S. Mutual Fund Database* (CRSP, 1998) and compares returns contained in it to those in Morningstar. The CRSP database has an omission bias that has the same effects as survivorship bias. Although all mutual funds are listed in CRSP, return data is missing for many and the characteristics of these funds differ from the populations. The CRSP return data is biased upward and merger months are inaccurately recorded about half the time. Differences in returns in Morningstar and CRSP are a problem for older data and small funds.

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. 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.[1]

Whereas 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. 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 that we can compare to the CRSP database.[2]

* Elton and Gruber are from New York University. Blake is from Fordham University.

[1] For example, Chevalier and Ellison (1999), Blake and Morey (2000), and Chen and Pennacchi (2000) use Morningstar as a database whereas Zheng (1999) and Wermers (2000) use CRSP.

[2] See Elton et al. (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, is on the order of −70 basis points per year (see Brown and Goetzmann (1995), Elton et al. (1996a), Ferson and Schadt (1996), Gruber (1996), Carhart (1997), Daniel et al. (1997), or Wermers (2000)). Furthermore, cross-category differences (e.g., aggressive growth versus growth) are much smaller still (see Grinblatt and Titman (1989) or Elton et al. (1993)). Finally, the evidence on mutual fund predictability is based on small differences (see Grinblatt and Titman (1992), Brown and Goetzmann (1995), 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 is the subject of the first two sections of this paper. In Section I, 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 traditional survivorship bias, does have a form of survivorship bias called *omission* bias that causes the same type of problems as does traditional survivorship bias. In Section II, we show that the returns in the CRSP database are upward biased in any month where there are multiple distributions on the same day. In Section III, we discuss the accuracy of an innovative feature of the CRSP database. The database 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 a random number of months before the merger month. We discuss whether this is a problem. Finally, in Section IV, after correcting both Morningstar and CRSP for the biases discussed in this paper, we compare the monthly fund returns contained in those data sources. 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. Omission Bias and Survivorship Bias

In this section, we show that the CRSP database, while free of traditional survivorship bias, has a form of survivorship bias, which we identify as omission bias, that is significant and has many of the same characteristics as traditional survivorship bias.

The Morningstar database has survivorship bias. As shown in Elton et al. (1996a), this causes overall performance measures to be inflated between 40 basis points and one percent, 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 significant positive average alpha when the true average performance is negative.[3] 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.[4] Finally, survivorship bias can lead to an appearance of predictability when none is present (see Brown et al. (1992)).

CRSP claims that their mutual fund database is free of survivorship bias. In fact, their database title is *CRSP Survivor Bias Free U.S. Mutual Fund Database*. All researchers using it seem to accept this and use it for that reason. Unfortunately, for performance measurement studies, this is not accurate. The CRSP database has a form of survivorship bias that we refer to as *omission* bias. Omission bias arises because the return data on the CRSP files is monthly for some funds, annual for others, and for some funds, no returns are recorded. The merger and liquidation rates are much lower for funds that have monthly CRSP return data than they are for funds with annual or no CRSP return data. Thus, researchers using monthly CRSP fund return data have a sample that understates the proportion of mergers and liquidations and thus overstates performance for the population of funds. It also inaccurately measures differences in performance across funds with different objectives, and may demonstrate predictability where none exists. In short, this set of data exhibits all the problems of a sample with traditional survivorship bias.

To examine the magnitude of this problem, we use the survivorship-bias-free sample that we constructed in Elton et al. (1996a; EGB). This sample contains all U.S. equity funds that listed “common stock” as their objective in the 1977 annual edition of Wiesenberger’s *Investment Companies* (1976 to 1978 which lists year-end 1976 data) and that listed total net assets of $15 million or greater.[5] Table I shows the number of common stock mutual funds listed in the 1998 CRSP mutual funds database that were in existence 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).[6] The table also shows whether

[3] See also Brown and Goetzmann (1995) and Carhart et al. (2000).

[4] See Elton et al. (1993) for an analysis of the difference in performance across different classifications.

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

[6] 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 et al. (1996a). However, in our sample, we dropped 19 of those funds that restricted the type of individual who could own the fund, were closed to new investors, or were variable annuity funds. In Table I and in this analysis, we

Table I

Merger Experience and CRSP Return Data Availability (January 1977 to 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]
1. Total funds	188	42	154	83
2. Number of funds with complete monthly returns[e]	188[f]	42[f]	93	24
3. Number of funds with some monthly returns	0	0	10	9
4. Number of funds with no monthly returns	0	0	51[g]	50[g]

All CRSP fund data obtained from CRSP Mutual Fund Database, Version 1.0 1998.
[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 categorized 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.

CRSP included complete monthly return data, only some monthly return data, or no monthly return data for funds that were listed in the CRSP database. All but two funds of the funds listed in Table I with $15 million or more in total net assets have complete monthly return data in the CRSP database.[7] CRSP's assessment of which funds merged or liquidated for funds with total net assets over $15 million closely agrees with our assessment. CRSP shows that 43 funds merged and one liquidated; we show that 42 merged and none liquidated.[8]

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 these funds were not listed by Nasdaq at the beginning of our sam-

exclude those 19 funds. CRSP has a count of 155 small common stock funds rather than the count of 154 funds that appear in Table I. CRSP misclassified one fund. Wiesenberger Financial Services (1976 to 1978) 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" (spec) except in 1977; in the 1977 edition the fund's code is listed simply as "S," a code that has no Wiesenberger definition.

[7] Those two funds have missing CRSP return data for a few months. They are included as having complete data since the data were readily available from other sources.

[8] 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).

ple period, and Nasdaq was the source of monthly fund returns provided to all data vendors in the early years of our sample.[9] Thus, for many defunct small funds, monthly returns cannot be found in any source. Note that all but one of the funds for which CRSP reported existence but did not record any monthly data merged or liquidated. Thus, excluding these funds from a study of mutual fund performance means that the characteristics of the sample are substantially different from the population. Funds that are less than $15 million in size with partial monthly data are also likely to be excluded since six of those funds have five or less years of data. Unless the starting date of the monthly returns for those funds happen 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 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 caused by omitting funds for which CRSP does not include monthly data. In Elton et al. (1996a) for the large-fund group examined here, we calculated an average yearly alpha using a three-index model for return generation of -0.1269 percent for surviving funds and -2.8779 percent for nonsurviving funds.[10] Using only those funds with complete monthly data in CRSP, one would have a sample of 281 funds (from row 2 of Table I) and could observe which of those funds did not survive. The alpha one would obtain is

$$(215/281) \times (-0.1269) + (66/281) \times (-2.8779)$$

$$= -0.773, \text{ or } -77.3 \text{ basis points.}$$

The total number of funds, including funds with no monthly data, is 342 funds (from row 1 of Table I) of which 125 merge or liquidate. Thus the population alpha is

$$(217/342) \times -0.1269 + (125/342) \times -2.8779$$

$$= -1.132, \text{ or } -113.2 \text{ basis points.}$$

This gives a bias of 35.9 basis points.[11]

[9] The rule for Nasdaq listing was over $15 million in assets or 1,000 investors.

[10] The three indexes used in Elton et al. (1996a) and in this study are the S&P 500 Stock Index minus the 30-day T-bill rate, an index of the return on a portfolio of small stocks minus the 30-day T-bill rate, and an index of the return on a portfolio of corporate and government bonds minus the 30-day T-bill rate.

[11] If those funds with partial monthly data are included in the sample with monthly data, then the bias is reduced to 30 basis points.

To test the robustness of our estimate of bias, we repeated the analysis above using the Fama–French (1996) three-factor model rather than our three-factor model to estimate alphas. When we used the three-factor Fama–French model, the bias is 44 basis points as compared to the 35.9 basis point estimate arrived at with the EGB three-factor model. Elton et al. (1996a) estimated survivorship bias to be 90.6 basis points, using methodology similar to the first of these techniques. If we use our best estimate of omission bias in the CRSP database, 35.9 basis points, then omission bias in the CRSP database is about 40 percent as large as traditional 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.

As a further check on the existence of omission bias, we calculated 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 an average of 16 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 in that group with annual data. From year to year, the difference in average annual return ranged from 2.1 percent to 9.8 percent, with an overall average difference of 6.1 percent.[12]

An easy way to avoid the problem of omission bias 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.

II. Upward-biased Monthly Returns in the CRSP Mutual Funds Database

There is a systematic bias in CRSP's calculation of fund returns because of the formula CRSP uses to adjust returns for distributions. The formula results in a consistent overstatement of returns for any period in which more

[12] The 6.1 percent 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. Subsequent to the acceptance of this paper, CRSP corrected the method of calculation that caused this bias.

than one distribution occurs on the same day. In this section, we first show the source of the bias and the correct formula. The bias is easy to correct, and all future researchers will want to do so. However, the presence of the bias means we need to examine its importance for already published papers.

The formula CRSP uses for calculating monthly fund returns in a month with no splits is[13]

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

where $R_{t-1,t}$ is the return on a fund between time $t-1$ and time t; NAV_t is the fund's net asset value at the end of the current period; NAV_{t-1} is the fund's net asset value at the end of the previous 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; and $RE\_NAV_j^D$ is the NAV at which the jth dividend or capital gains distribution was reinvested.

Although 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 CRSP 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. The impact on returns 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 is \$12.53. CRSP lists the Vanguard Windsor Fund's NAV at the end of November 1994 as \$13.71 and as \$12.59 at the end of December 1994. 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 percent. This is also the (rounded) amount shown in CRSP.

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. The correct computation of return is therefore

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

[13] See *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 1.0 1998).

Table II

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 ICDI 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.	Mar.	Apr.	May	June	July	Aug.	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%

Within each group, the top 25 funds do not include index funds or multiple classes of the same fund. All CRSP fund data obtained from CRSP Mutual Fund Database, Version 1.0 1998.

or −0.1075 percent. This amount is, in fact, the return reported by Morningstar (1999) for the Vanguard Windsor Fund, and it is also the actual return received by holders of that fund.

This bias is easy to correct and should be corrected in any future study. However, it is worth briefly examining its importance in evaluating prior studies.

We selected five years (1994 to 1998) of CRSP data to study the magnitude of 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 the five CRSP ICDI fund objectives, shown in Table II. The largest funds were selected because their 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 percent, have two or more payments on the same date. Seventy-five percent of the same-day multiple payments occur in December, since December is the standard month to pay capital gains. For the months with multiple payments, the bias in CRSP returns is 4.2 basis points in that month.

To examine the impact on alpha, we employed the Elton et al. (1996b) four-index model.[14] The effect on average alpha for the 125 funds in our sample of using the correct return data is 3.84 basis points per year. Thus, for most studies, the impact of using the uncorrected CRSP returns did not create a problem. However, it might be a problem in studies where alphas on individual funds are important. Out of the 125-fund sample, the five funds with the largest error in alpha had an error of 21.61 basis points per year, whereas for the top 10 funds, the average error was 16.40 basis points, and for the top 25 funds it was 11.07 basis 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.

In summary, there is an upward bias in CRSP returns in any month where there are multiple distributions on the same day. It is easily corrected. The impact on average alphas is small, but results for studies that use alphas of specific funds should be accepted with caution.

III. Merge Data

For funds that merged or liquidated, CRSP supplies 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. In this section, we analyze the accuracy of the CRSP merge data.

Why should we care about accurate dates for mergers? There are two reasons. First, studies of mutual fund organization and performance frequently use data on fund size, cash flows, and the impact of fund mergers. For such studies, information on merger dates and merger partners is extremely important. The second reason involves the impact of mis-estimated merger data on general performance studies.

To examine accuracy, we use the sample of funds from Elton et al. (1996a). EGB uses a "follow the money" approach to track the performance of all funds classified as "common stock" by Wiesenberger at the end of 1976 that had over $15 million in assets under management at the end of 1976. For each of the 42 funds from this sample that merged after 1976, they contacted the management company to get the exact date and terms of merger.[15]

Table III presents differences between the data reported by CRSP and the actual merger date. Note that in 21 percent of the cases, the CRSP merger date is more than one month removed from the actual merger month. For mergers that were midmonth and CRSP was within a month of the actual

[14] The four indexes are the S&P 500 index minus the risk-free rate (30-day T-bill), a small-cap minus large-cap stock index, a value minus growth stock index, and an aggregate bond index minus the risk-free rate; see Elton et al. (1996b) for details.

[15] CRSP shows 42 funds merged and 1 liquidated. The difference comes about because CRSP missed a name change and called it a liquidation.

Table III

CRSP Merger Dates Relative to Actual Merger Dates

Status	Occurrences
A. CRSP date differs by one month or less from actual date	19
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 midmonth	22
a. First day of actual merger month	10
b. Last day of month prior to actual merger month	11
c. Different midmonth 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

All CRSP fund data obtained from *CRSP Mutual Funds Database*, Version 1.0 1998.

date, about half the time the merger is listed as the last day of the previous month, and about half the time the last day of the month in which the merger actually occurred. The following rule would give a researcher using the CRSP database the highest degree of accuracy in identifying the merger month. Assume that mergers that are shown in CRSP to occur on the last day of the month occur in the next month, and those shown at the beginning 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 misclassified fund.

As shown in Table IV, 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. If one accepts the merger dates used in Elton et al. (1996a), 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 (Table IV, Panel A) and "show data to the month prior to the merger" if it merges before the end of the month (Table IV, Panel B). This is a sensible and consistent way to collect returns. In the 10 remaining cases (Table IV, Panel C), CRSP ends returns some months before the time of the merger. The number of months that CRSP stops early varies from 1 to 36, with 4 cases off by a year or more.[16]

The final issue to examine is whether the lack of return data in months prior to the merger introduces bias. We use the CRSP return data and estimate the four-index model discussed earlier using three years of data (cor-

[16] 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. In 14 of the remaining 15 cases, monthly returns end from one month to two years earlier. In the final case, CRSP has return data after the CRSP merger date.

Table IV

CRSP Monthly Return Ending Month Relative to Actual Merger Month

Status	Occurrences
A. CRSP reports returns for the actual merger month and fund merges end of month	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 midmonth	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 midmonth in actual merger month	1
5. CRSP merger date not listed	3
C. CRSP returns end more than one month before actual merger date	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 survived	1

All CRSP fund data obtained from *CRSP Mutual Fund Database*, Version 1.0 1998.

rected for multiple distribution on the same day) ending in the last month CRSP shows data. We then use 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. We calculated residuals 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 midmonth or to the end of the month if the merger occurred at the end of the month. There is no systematic pattern. Roughly half of the average residuals across the nine funds are positive and half are negative. Cumulating across all nine funds and all months, and treating each observation equally, produced an average residual close to zero. We find no evidence of bias due to missing observations prior to merger.

In summary, anyone studying mutual fund mergers should use CRSP merger data only as a starting point to obtain the names of the merger 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 use 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 some data on funds that merge and liquidate. The advantage of the Morningstar database is that it includes much more data on composition and performance, and it 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, and if the data is selected and adjusted to correct the known biases in each data source, 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 eliminate any differences due to Morningstar survivorship bias and CRSP omission bias. We corrected the CRSP returns for the bias caused by multiple distributions on the same day. Because we were interested in the effects of size, we studied the 25 largest nonindex funds and the 25 smallest nonindex funds with over $15 million in assets in the growth and income group. We use 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 discussed earlier to CRSP and Morningstar data for each of the four five-year samples is presented in Table V. Two facts are immediately apparent from Table V. 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 Brown and Goetzmann (1995), Elton et al. (1996b), Gruber (1996), or Carhart (1997)). If one is studying mutual fund performance before the mid-1980s, differences in alpha are sufficiently large that conclusions might well be affected depending on whether one uses the Morningstar or CRSP databases.

When trying to identify good managers or to look at performance persistence, the differences in alphas for individual funds across different data sources are important. There are a number of large differences in individual fund alphas. For example, when examining the small funds, 34 out of the 100

Table V

Differences in Monthly Alphas Estimated from Four-index Model Using CRSP and Morningstar Monthly Return Data (All Values Expressed in Basis Points)[a]

				Number of Differences Greater Than or Equal To		
Sample Period	Number of Funds with Difference	Avg. Difference[b]	Avg. Absolute Difference	10 Basis Points	5 Basis Points	1 Basis Point
Large Funds						
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	−0.378	1.48	3	6	18
Small Funds						
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

All CRSP fund data obtained from *CRSP Mutual Fund Database*, Version 1.0 1998.
All Morningstar fund data obtained from *Principia Pro* January 1999, CD.

[a] The alphas are intercepts of the regression of the excess return for each fund against the excess return on the S&P 500 Stock Index, the return on the small stocks minus the return on the large stocks, the return on growth stocks minus the return on value stocks, and the return on long term bonds minus the T-bill rate.

[b] Differences measured as alpha using CRSP data minus alpha using Morningstar data.

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 percent, 4.9 percent, 4.9 percent, 3.1 percent, and 2.5 percent 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 per year, and 3 greater than 120 basis points per year). Furthermore, the largest differences are smaller; the five largest are 4.9 percent, 3.1 percent, 1.9 percent, 1.6 percent, and 1.4 percent. For large funds, the greatest differences are all in the first 5 years. For small funds, large differences continue through the 20 years.[17]

[17] Although 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 the same equation over the full 20-year period, a period that is longer than we would expect any researcher to employ. Over the 20-year period, over 10 percent of the alphas have errors greater than 60 basis points.

Table VI

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

Sample Period	Number of Months with Difference	Avg. Difference[a]	Avg. Absolute Difference	Number of Differences Greater Than or Equal To 5.0%	1.0%	0.5%
			Large Funds			
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	1,435	0.000	0.052	14	58	91
			Small Funds			
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	1,533	−0.010	0.111	29	138	237

[a] Differences measured as return using CRSP data minus return using Morningstar data.
All CRSP fund data obtained from *CRSP Mutual Funds Database*, Version 1.0 1998.
All Morningstar fund data obtained from *Principia Pro* January 1999 CD.

Differences in alphas are important. They can change the conclusions about individual funds or a group of funds. Furthermore, any researcher using data 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 VI presents summary data on the differences in returns between the two databases. The table makes it clear that the differences between the databases, and hence their accuracy, has gotten a lot better in recent years and is more of a problem for small funds. When we apply a rule that any monthly return difference of more than 0.5 percent between the two databases leads to an alpha difference of 12 or more basis points per year, we can correctly identify 51 of 52 cases where alpha differences of 12 basis points or larger occur. This would be a useful rule to identify problem cases.

Although 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.

V. Conclusion

The CRSP database is a fairly new publicly available database on mutual funds. 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.

There are two bias problems. First, although CRSP does not suffer from traditional survivorship bias, it does suffer from a form of survivorship bias called omission bias. Because only *some* small funds under $15 million in total net assets have monthly data in the CRSP database, and because the omitted funds have much greater merger and liquidation rates, we show that the returns reported for the group of small funds that have monthly data overstate the population returns and alphas. Second, returns in the CRSP database for months with multiple distributions on the same day are overstated. The Morningstar database is free of this problem.

We also examine the data CRSP provides on mergers. Although these data are quite good in identifying mergers, we show that there are 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 serious 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, Christopher R., Edwin J. Elton, and Martin J. Gruber, 1993, The performance of bond mutual funds, *The Journal of Business* 66, 371–403.

Blake, Christopher R., Matthew R. Morey, 2000, Morningstar ratings and mutual fund performance, *The Journal of Financial and Quantitative Analysis* 35, 451–483.

Brown, Stephen J., and William N. Goetzmann, 1995, Performance persistence, *Journal of Finance* 50, 679–698.

Brown, Stephen J., William N. Goetzmann, Roger G. Ibbotson, and Steve A. Ross, 1992, Survivorship bias in performance studies, *The Review of Financial Studies* 5, 553–580.

Carhart, Mark, 1997, On the persistence of mutual fund performance, *The Journal of Finance* 52, 57–82.

Carhart, Mark, Jennifer Carpenter, Anthony W. Lynch, and David K. Musto, 2000, Mutual fund survivorship, Unpublished manuscript, New York University.

Chen, Hsiu-Lang, and George Pennacchi, 2000, Does prior performance affect a mutual fund's choice of risk? Unpublished manuscript, University of Illinois.

Chevalier, Judith, and Glenn Ellison, 1999, Are some mutual fund managers better than others? Cross-sectional patterns in behavior and performance, *Journal of Finance* 54, 875–899.

CRSP, 1998, *CRSP Survivor Bias Free U.S. Mutual Fund Database,* Version 1.0, 1998, Center for Research in Security Prices, Graduate School of Business, The University of Chicago.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic based benchmarks, *Journal of Finance* 85, 1088–1105.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 1996a, Survivorship bias and mutual fund performance, *The Review of Financial Studies* 9, 1097–1120.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 1996b, The persistence of risk-adjusted mutual fund performance, *The Journal of Business* 69, 133–157.

Elton, Edwin J., Martin J. Gruber, Sanjiv Das, and Matt Hlavka, 1993, Efficiency with costly information: A reinterpretation of evidence from managed portfolios, *The Review of Financial Studies* 6, 1–22.

Fama, Eugene, and Ken French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 35–84.

Ferson, Wayne E., and Rudi Schadt, 1996, Measuring fund strategy and performance in changing economic conditions, *Journal of Finance* 51, 425–461.

Grinblatt, Mark, and Sheridan Titman, 1989, Portfolio performance evaluation: Old issues and new insights, *Review of Financial Studies* 2, 393–421.

Grinblatt, Mark, and Sheridan Titman, 1992, Performance persistence in mutual funds, *Journal of Finance* 47, 1977–1984.

Gruber, Martin J., 1996, Another puzzle: The growth in actively managed mutual funds, *Journal of Finance* 51, 783–810.

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

Wermers, Russ, 2000, Mutual fund performance: An empirical decomposition into stock picking talent, style, transaction costs and expenses, *Journal of Finance* 55, 1655–1695.

Wiesenberger Financial Services, 1976 to 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, 905–935.