Data Appendix
Measuring Managerial Skill in the Mutual Fund Industry
By Jonathan B. Berk and Jules H. van Binsbergen

(Last updated 06/09/2011)

Preliminary Notes:

This documentation summarizes all the steps we have taken so far in merging, organizing and correcting errors in CRSP and Morningstar.

List of Contents

Raw CRSP Database Clean-Up:

We start the algorithm by sorting out the Morningstar and CRSP databases so they are ready for merging. This involves several error corrections on both databases. The raw CRSP database contains 3973218 observations, 43662 distinct *crsp_fundnos*, and spans from April 1961 to December 2009. We make the following changes to this raw CRSP database, in the order presented below:

1) We observe that there are cases where two "identical" observations with the same *crsp_fundno* during the same *year* and *month* coexist in our database. We consider two observations to be "identical" if they have the same *crsp_fundno*, *year*, *month*, *fund_name*, *mret* (monthly net returns), *mnav* (net asset value), and *mtna* (total net assets). Under this criteria, a total of 10 pairs of observations in CRSP are found to be repeated. All the observations differed from each other wrt fiscal year end --- in each case one of them had a fiscal year end different from the year of the observation, while one of them had a fiscal year end that matched the year of the observation. We judge that this repetition is a mistake made in the original CRSP database in trying to correct the error in fiscal year ends. Consequently, for every pair of repetitive observations, the observation in which the fiscal year end does not match the year of observation is removed from the database. By doing so, we removed 10 observations. After this removal the variables *crsp_fundno*, *year*, and *month* together can uniquely identify an observation in the CRSP database. It should be noted that in three cases the two repeated observations did differ in a material variable, the expense ratio.

2) We back-filled and forward-filled the ticker in CRSP using *crsp_fundno* as the benchmark. We first recognized that some *crsp_fundno*'s use more than one ticker. There are 109 such *crsp_fundnos*. To ensure that despite the ticker change, each *crsp_fundno* only identifies a single fund, we looked through these manually and followed through their *fund_names* before and after the *ticker* change. Judging from the *fund_names*, we found that despite the *ticker* changes, for all of the 109 *crsp_fundnos*, the actual fund remained the same. This meant that no *crsp_fundno* was assigned to two different funds. Because databases are merged based on *ticker*, we need to ensure that *ticker* also uniquely identifies a fund, so to ensure a one-to-one correspondence between *crsp_fundno* and *ticker* we first identify the last non-empty *ticker* used by a fund, where a fund at this point is still defined by a unique *crsp_fundno*. Then for every observation of that fund:

   a) If the observation has an empty *ticker*, we replace it with the last non-empty *ticker* used by that fund.
   b) If the observation already has a *ticker*, but this *ticker* is not the same as the last non-empty *ticker* used by the fund, we replace this *ticker* using the *last ticker* used by the fund.

   As a result of forward and backward filling, a total of 648467 *tickers* that were initially empty are replaced with non-empty *tickers*, and a total of 16861 *tickers* that were initially non-empty are replaced with different non-empty *tickers*. Note that after back-filling and forward-filling, our database has the following characteristics:

a) A *crsp_fundno* either corresponds to an empty *ticker*, or to a unique non-empty *ticker*, but never to both an empty *ticker* and non-empty *ticker*, or a multiple of non-empty *tickers*.
b) A *crsp_fundno* is assigned to an empty *ticker* in any observations only if this *crsp_fundno* is never assigned to a non-empty *ticker* in all of its observations in the original database.

3) There are also cases where two *crsp_fundnos* are assigned simultaneously to the same *ticker*. This means that in the same *year* and *month*, two *crsp_fundnos* use the same non-empty *ticker*. There are 2486 such *tickers* that, during the same *year* and *month*, have been used by more than one *crsp_fundnos*. A look into these cases suggests that the reasons multiple *crsp_fundnos* are assigned simultaneously to the same non-empty *ticker* are:

    i. CRSP further divides one *ticker* into multiple subclasses, and separately identifies these subclasses using different crsp_fundnos.
    ii. CRSP made a mistake in assigning ticker, and one of the crsp_fundno is using an incorrect ticker.

    To prevent incorrect matches in merging our CRSP database with the Morningstar database caused by cases i and ii as stated above, for any observations having one of those 2486 *tickers*, we replace its *ticker* to an empty *ticker*. We leave it to our merging algorithm later on to find the correct Morningstar match for these observations. In this step, we erased the *tickers* of 206010 observations.

4) Finally, at this stage we have 529299 observations in the database having empty *tickers*. For each of these observations, we replaced the empty *ticker* with the *crsp_fundno*.

After correcting the CRSP database using steps 1) to 4) as stated above, our modified CRSP has the following properties:

1. There is no empty *ticker* in this modified database, since there is no empty *crsp_fundno* in the raw database.
2. Each *crsp_fundno* is assigned to one and only one *ticker*.
3. Each *ticker* only corresponds to one *crsp_fundno* for any given *year* and *month*. However, it is not yet true that each *ticker* only corresponds to one *crsp_fundno* in general, because *tickers* can get reused by another fund after the fund initially using it dies. We correct for this problem later in the algorithm.
4. The three variables *ticker*, *year*, and *month* uniquely identify an observation in this modified database.

This modified CRSP database is saved for further use in merging.

Raw Morningstar Database Clean-Up:

The raw Morningstar database covers 2718257 observations, and a total of 39200 non-empty *tickers*. It spans December 1995 to December 2009. The following changes are made to the raw Morningstar database, in the order they are presented in this document below:

1. We back-filled and forward-filled the ticker in Morningstar using *fundname* in Morningstar as the benchmark. This means that if a non-empty *fundname* has an empty *ticker* at any time A, but the identical non-empty *fundname* has a non-empty *ticker* at another time B, then we replace the empty *ticker* at time A with the non-empty *ticker* at time B. Note that this algorithm of filling in *tickers* is not ambiguous, because there is never a case where the same *fundname* corresponds to two or more non-empty *tickers* in our original database. Thus after back-filling and forward filling we have:

   a) In the entire modified Morningstar database, each non-empty *fundname* either corresponds to a unique non-empty *ticker*, or to an empty *ticker*, but never to both.
   b) There can be multiple *fundname* assigned to the same non-empty *ticker*.

   In this step, 140830 observations had their empty *tickers* replaced with a non-empty *ticker*. There are still 115495 observations with empty *tickers*

2. There are cases where two observations with the same non-empty *ticker* during the same *year* and *month* coexist in our database. We treat such cases as follows.

   a) In 7 pairs of the repeating observations, not only do the two observations in each pair have the same *ticker*, *year* and *month*, they also have the same *totre1mo* (monthly net return), *monthendnav* (net asset value), and *netassetsmm* (total net assets). For these 7 pairs, we conclude that the repeating observations must refer to the same fund because all fields are identical for each pair. Consequently, one of the two observations in each pair is dropped.

   b) Otherwise, if both observations in the repeating pair have the same *ticker*, *year* and *month* but have different *totret1mo*, *monthendnav*, or *netassetsmm*, we conclude that Morningstar made a mistake in labeling the ticker for at least one of the observations in each of the repeating pairs. Consequently, to prevent carrying this mistake into the merged database later on, we replace this *ticker* with an empty *ticker* for all observations in the database using this *ticker*. 484 *tickers* out of a total of 37687 observations are erased in this step as a result.

Note that consequently, our modified Morningstar database holds the following properties:

a. Each non-empty *fundname* either corresponds to a unique non-empty *ticker*, or to an empty *ticker*, but never to both.
b. The variables *ticker*, *year* and *month* can identify a unique observation if the ticker is non-empty.

This modified Morningstar database is saved for further use in merging.

Merging Morningstar and CRSP databases

At this stage, we are ready to merge together the CRSP database with the Morningstar database. The identifiers that will be used to perform the merge are: *year*, *month*, and *ticker*. This means that we will match two observations if they are in the same *year* and *month*, and have the same *ticker*. In the first step, we merge the database we prepared in the two earlier sections. The statistics regarding an initial merge are outlined below:

	# total	# merged	# not merged
observations in Morningstar	2718250	2063237	655013
observations in CRSP	3973208	2063237	1909971

The statistics given above include those funds in CRSP prior to 1996. Since Morningstar only dates back to 1996, it is natural that we cannot match a large fraction of observations in CRSP onto Morninstar. If we instead only consider observations after 1995, the matching statistics look as follows:

	# total	# merged	# not merged
observations in Morningstar	2718250	2063237	655013
observations in CRSP	3212700	2063237	1149463

We notice that in a significant fraction of the observations in both databases, the ticker is missing. This means that there are four problems we need to resolve before successfully merging the two databases:

1. For some funds, the *ticker* was originally missing in CRSP. For these observations we filled in those *tickers* using the *crsp_fundno* (as described in the earlier section). This prevents a successful match.
2. For some funds, the *ticker* in CRSP is not missing, but the *ticker* in Morningstar is missing.
3. For some funds, the *ticker* in Morningstar is missing, and the ticker in CRSP was also originally missing and in the earlier section, we replaced the ticker in CRSP with *crsp_fundno*.
4. For some funds, the *ticker* used by Morningstar is not a real *ticker*, but seems to be Morningstar's own identifier (i.e. "F000N"). Consequently, it will differ from the "real *ticker*" that CRSP uses.

Before we describe the process in detail, here is a quick summary of what we will do. We first match all observations that have the same *year*, *month* and *ticker*. The remaining observations fall into one of the above 4 categories. To match within that category, we begin by finding, for each Morningstar observation, the set of CRSP observations in the same *year* and *month* having a return within 2 basis points of the return on Morningstar, and having a NAV within 0.02 of the NAV on Morningstar. We then search though this set and identify the fund with the closest fund name (to be formally defined later in this section) to the Morningstar fund. Next we use the time series information --- for each unique fund name on Morningstar we define a correct match as a case when over 60% of the above matches correspond to the same CRSP fund number. (Because Morningstar and CRSP returns do not agree in a significant fraction of cases when we have a definitive match on ticker, it does not make sense to use a higher percentage.) In that case we associate all observations under that Morningstar fund name to their corresponding observations

in CRSP with that *crsp_fundno*. We fill in the *ticker* symbol if it exists on either data base and if it does not exist, we use the *crsp_fundno* as the ticker symbol. In the next paragraphs we outline the process in greater detail.

Step One

First, we take every unmatched observation in Morningstar, count the number of words in it. We then check all unmatched CRSP observations that are from the same *year* and *month*, have *mret* not different from the Morningstar return by more than 2 basis points, and have NAV not different from the Morningstar NAV by more than 0.02 in value. From all unmatched CRSP observations that satisfy the above criteria, we select the candidate whose *crsp_fundname* matches with the Morningstar fund name the most. That is, the CRSP observation such that the highest percentage of words in the Morningstar fund name can be found in its c*rsp_fundname*. Further, the corresponding *crsp_fundname* found also needs to satisfy the following three criteria:

1. If the Morningstar fund name contains any single-letter words, then all of these words need to be found in the *crsp_fundname*.
2. If the Morningstar fund name contains words "ii" or "iii", then these words need to be found in the *crsp_fundname*.
3. The first word in the Morningstar fund name must be found in the *crsp_fundname.*

These three criteria are used to reduce possible error, and especially to ensure that we do not incorrectly match different subclasses. If two crsp_fundnames both satisfy these criteria and the same percentage of words in the Morningstar fund name can be found in them, we pick one randomly. If no fund satisfies the above criteria, then we judge that this Morningstar observation does not have a corresponding CRSP observation.

Step Two

In the earlier step, in each period we matched a Morningstar observation to a CRSP observation. The next step uses the time series properties --- that is, if the match is accurate, then the same fund should be matched in every period. We sort all Morningstar observations that were not matched based on the ticker by their *ms_fundname*. Then for observations having the same *ms_fundname:*

1. If by the mechanism of Step One, more than 60% of these observations are matched onto CRSP observations having the same *crsp_fundno*, we match all observations having this *ms_fundname* to CRSP observations having this corresponding *crsp_fundno*. For example, in the case given below, over 60% of the Morningstar observations under the fund name "Armada: Money Market: B" is matched to the *crsp_fundno* "4083", so we match the *ms_fundname* "Armada: Money Market: B" to the *crsp_fundno* "4083".
2.

Before:			
year	month	ms_fundname	crsp_fundno matched to
1996	1	Armada: Money Market: B	4083
1996	2	Armada: Money Market: B	.
1996	3	Armada: Money Market: B	4083
1996	4	Armada: Money Market: B	4083

After:			
year	month	ms_fundname	crsp_fundno matched to
1996	1	Armada: Money Market: B	4083
1996	2	Armada: Money Market: B	4083
1996	3	Armada: Money Market: B	4083
1996	4	Armada: Money Market: B	4083

3. If by the mechanism of Step One, we do not find that over 60% of unmatched observations having this *ms_fundname* is matched to CRSP observations with a unique crsp_fundno, then we conclude that those unmatched Morningstar observations should not be matched to any *crsp_fundno*. Consider the following example: the *ms_fundname* "AXA Enterprise Hlthcare A" is matched to *crsp_fundnos* "1023", "7370" and "242", but none of the *crsp_fundnos* is matched to over 60% of the Morningstar observations having that *ms_fundname*, so we do not match the fund name "AXA Enterpirse Hlthcare A" to any crsp_fundno.

Before:			
year	month	ms_fundname	crsp_fundno matched to
2007	6	AXA Enterprise Hlthcare A	1023
2007	7	AXA Enterprise Hlthcare A	7370
2007	8	AXA Enterprise Hlthcare A	1023
2007	9	AXA Enterprise Hlthcare A	.
2007	10	AXA Enterprise Hlthcare A	242
After:			
year	month	ms_fundname	crsp_fundno matched to
2007	6	AXA Enterprise Hlthcare A	.
2007	7	AXA Enterprise Hlthcare A	.
2007	8	AXA Enterprise Hlthcare A	.
2007	9	AXA Enterprise Hlthcare A	.
2007	10	AXA Enterprise Hlthcare A	.

At this stage we have matched the *crsp_fundnos* with *ms_fundnames*. Now we update these changes in the CRSP and Morningstar databases. We take the initially unmatched CRSP observation and Morningstar observation pair, we follow the procedure outlined below:

1. If CRSP's ticker is not its *crsp_fundno*, we replace the Morningstar ticker with the CRSP ticker.
2. If the Morningstar observation has a non-empty ticker and the CRSP's ticker is its *crsp_fundno*, we replace the CRSP ticker with the Morningstar ticker.
3. If the Morningstar observation has an empty ticker and CRSP's ticker is its *crsp_fundno*, we replace Morningstar's ticker with the *crsp_fundno*

We then merge the modified CRSP and Morningstar databases, the new statistics are:

	# total	# merged	# not merged
observations in Morningstar	2718250	2333586	384664
observations in CRSP	3973208	2333586	1639622

If we only look at observations after 1995, the statistics looks as follows:

	# total	# merged	# not merged
observations in Morningstar	2718250	2333586	384664
observations in CRSP	3973208	2333586	879114

After we merged together the files for the second time, we implement another series of checks to remove possible errors. If any of the following problems are found in the merged database, mistakes were likely introduced in Step Two:

1) Two *ms_fundnames* that have overlapping dates are assigned the same *crsp_fundno*, which indicates that at least one of these *ms_fundnames* is matched incorrectly. There are 80 such *crsp_fundnos* that are suspicious in this way, and these 80 *crsp_fundnos* have 5561 observations in our database.

2) Two Morningstar fund names sharing the same ticker are matched to different *crsp_fundnos* having overlapping dates, which also indicates that at least one of these matches is mistaken. There are 2 such *tickers* and 5 *crsp_fundos* that are suspicious in this way, affecting 562 observations.

We make a note of each of these suspicious cases and manually go through our database to check each of them. We find that for each of these suspicious cases, we made a mistake in Section Two when matching *ms_fundnames* to *crsp_fundnos*. The mistakes are caused by our inability to distinguish subclasses denoted by two or more letters. After correcting for these mistakes in the Morningstar and CRSP databases separately, we merge the two databases again, and the merge statistics we obtain are:

	# total	# merged	# not merged
observations in Morningstar	2718250	2335957	382293
observations in CRSP	3973208	2335957	1637251

If we only look at observations after 1995, the statistics look as follows:

	# total	# merged	# not merged
observations in Morningstar	2718250	2335957	382293
observations in CRSP	3212700	2335957	876743

Step Four

We repeat Step One to Step Three on this modified combined file, in order to check if we have missed any possible merges. Note that it is possible that we have missing possible merges, because after we generated the first merge batch using the earlier steps, the heuristic algorithm faces less noise when searching for matching observations in the subsequent runs. After repeating Step One to Step Three twice more we obtain the following final statistics:

	# total	# merged	# not merged
observations in Morningstar	2718250	2389371	328879
observations in CRSP	3973208	2389371	1543837

If we only look at observations after 1995, the statistics look as follows:

	# total	# merged	# not merged
observations in Morningstar	2718250	2389371	328879
observations in CRSP	3212700	2389371	783329

Correcting for Reused Tickers

Next, we must correct the problem that tickers are reused, that is, the same *ticker* gets used repeatedly by different funds. This correction is necessary because we want to make *ticker* the variable that uniquely identifies funds in our database. We use the following rules to correct repeated *tickers*:

We sort *ticker* by *year* and *month*. If a *ticker* is included in the CRSP database, then we say that a *ticker* is reassigned if the *ticker's* corresponding *crsp_fundno* in the previous observation is not the same as the corresponding *crsp_fundno* in this observation, and the two observations have a gap of more than a year. An example of a case where we decide that the ticker got reassgined is given below.

year	month	crsp_fundno	ticker
1999	2	7390	ACEXX
1999	3	7390	ACEXX
1999	4	7390	ACEXX
2000	7	13685	ACEXX
2000	8	13685	ACEXX
2000	9	13685	ACEXX

We find cases where the crsp_fundno changes, but the ticker is in a continuous time series. That is, the previous observation was always the data one month ago, and there is never a gap in the ticker. We manually checked these cases and judged that CRSP made a mistake in identifying them as a separate fund (because we can tell that before and after the *crsp_fundno* change, manager and firm information did not change). Consequently, our algorithm requires that not only do we need a *crsp_fundno* change, we also need an at least 1 month gap between the tickers to be able to decide that the ticker got re-used.

If the *ticker* is used only in Morningstar, then we say that the ticker got reassigned if the last time it was used was more than 2 years ago, and the previous observation and this observation has different managers as well as different management companies. An example of this is:

year	month	ticker	manager	firm
2005	5	ACTAX	Smith	American Century
2005	6	ACTAX	Smith	American Century
2008	12	ACTAX	Lee	Mason Street
2009	1	ACTAX	Lee	Mason Street
2009	2	ACTAX	Lee	Mason Street

Once we decide on which tickers got re-used, we add the following suffix to the re-used tickers to differentiate them from their earlier use:

1) "/2" is added as a suffix to the ticker for second time use
2) "/3" is added as a suffix to the ticker for third time use

We find that in this database, the ticker is never reused three times or more.

Index Fund Correction and Search

The objective of this section is to accurately distinguish which one of the following three groups the funds in our database belong to:

1. Ordinary index funds - a passively managed mutual fund that tries to mirror the unleveraged performance of a specific index, such as the S&P 500.
2. Leveraged index funds - a mutual fund that tries to mirror the leveraged or inversed performance of a specific index.
3. Actively managed funds - mutual funds that are neither ordinary index funds nor leveraged index funds.

First, we noticed that Morningstar has an internal label for index funds included in its database. That is, Morningstar gives a label of "I" in the variable called *specialcriteria* if an observation belongs to an index fund. This label has some problems associated with it:

1. For a lot of funds, Morningstar labeled it as an index funds only for a few of its observations, and it did not label the remaining observations. We noticed that in most cases this was not because those funds changed their index status during the sample, but because Morningstar failed to label them for every monthly period.
2. Morningstar is not very accurate in labeling index funds, we find both cases where an index fund is not labeled and cases where a non-index fund is labeled as an index fund. We will describe how we found these incorrectly labeled funds and how we corrected for them later in this section.
3. Morningstar does not distinguish between ordinary index funds and leveraged index funds.

To make corrections to the Morningstar index label, we first identify all funds that are ever labeled as an index fund by Morningstar, which includes all funds that are identified by Morningstar as index funds in at least one month. There are a total of 2020 such funds. Because we suspect that Morningstar's label of index fund is neither comprehensive nor accurate, we developed a mechanism to check for the validity of Morningstar's label. To do so, we first generate a list of phrases likely to appear in index fund names and a list that appears in actively managed fund names. We then isolate the funds that Morningstar labeled as an index but contained words that likely indicate it to be an actively managed fund, and isolate the funds that Morningstar did not label as an index but contained words that indicate it actually appears to be an index fund. We then manually go through SEC filings to verify funds on these suspicious lists. Then we apply a similar mechanism to isolate and verify suspicious funds that appear to be leveraged index funds. A detailed description of this procedure, and the statistics on how many funds are selected and verified in each step, is given in the paragraphs below.

First, we take the 2020 funds that are labeled by Morningstar and list them. From now on, in this section of our appendix, "fund name" can refer to either *crsp_fundname* (CRSP's fund name) or *ms_fundname* (Morningstar's fund name). For example, if we write "fund name containing a word X," that means either *ms_fundname* contains the word X or *crsp_fundname* contains the word X. For each observation, we decompose all the words in its fund name. We then make a list of all the words that ever appeared in any of the 2020 fund's fund names (either *ms_fundname* or *crsp_fundname*), and denote the total number of times that word appeared in the names of the

2020 funds. For every word that appeared more than 10 times in these 2020 fund's names, we manually went through them and picked certain words to create the following list, which we will refer to as Table 1.

TABLE 1: key words used to identify index funds					
index	indx	idx	s&p	standard	poor
dow	jones	etf	ishare	profund	russell
proshare	powershare	viper	spider	spdr	wilshire

We judged that any funds having these words in their names are likely to be an index fund, and any fund without any of these words are very likely to be actively managed.

Then we make a list of funds that contain any of these words, but are not labeled as an index fund by Morningstar. Again we decompose all the words in them, rank the words by their number of appearances, and manually went through all words that appeared for more than 10 times and picked certain words to create the following table. We will refer to it as Table 2 from now on.

TABLE 2: key words used to refute index funds					
select	adv	hedge	manage	enhance	Plus
1970	1975	1980	1985	1990	1995
2000	2005	2010	2015	2020	2025
2030	2035	2040	2045	2050	

We judge that any funds with these words are likely to be actively managed funds, even if they also have some of the words in Table 1 in their fund names.

Using these two tables we make the following two lists of "suspicious funds". We consider a fund to be suspicious if it falls into any of the following two categories:

1. We created a list of funds that are labeled by Morningstar as an index fund, but either does not contain any of the words or phrases listed in Table 1, or contain at least one of the words or phrases listed in Table 2 (we do not differentiate between upper & lower case letters). The intuition is that here we are trying to identify funds that Morningstar labeled as index funds but do not actually look like an index fund. We found a total of 596 such funds.

2. We created a list of funds that appear in the Morningstar database, are not labeled as index funds by Morningstar, but contain at least one of the words listed in Table 1 and do not contain any of the words listed in Table 2. The intuition is that we also want to identify funds that are not labeled by Morningstar as an index funds, but looks like an index fund judging from its fund name. We found a total of 61 such funds.

Now we have two list of "suspicious funds", the intuition is that one list contains funds labeled by Morningstar as indexes but look like non-index funds, and the other list contains funds not labeled by Morningstar as indexes but look like index funds. So in this next step, we checked whether each of these funds are an ordinary index fund, an leveraged index fund, or an actively

managed fund by going to their company fillings using the EDGAR database on the SEC website. We checked multiple prospectuses. One at or as close to the inception date as possible, another during the time Morningstar labeled the fund, and if necessary a third time after that period. Of course, this was not possible for all of the funds since there were cases in which prospectuses were not available for the exact time. In these cases we checked the nearest possible prospectuses. If a fund changed from a non-index fund to an index fund or vice-versa we used the first prospectus in which the change occurred. If prospectuses were not available, we assumed the status remained the same. Within each prospectus, we searched for the fund name and looked in the sections titled "Investment Objective" or "Investment Policies." To track what we have done, we stored all the prospectuses we looked through to decide on whether a fund is an index fund. We saved the prospectuses we checked for future reference. Finally, for each fund that we manually checked, we make the following corrections:

1. If a fund has always been an actively managed fund, or has been an actively managed fund in a certain period, but Morningstar incorrectly labeled it as an index fund, then we re-label the fund as an actively managed fund for the period we identified it to be actively managed.

2. If a fund has always been an index fund, or has been an index fund in a certain period, but Morningstar failed to label it as an index fund, then we re-label the fund as an index fund for the period we identified it to be indexing.

3. If a fund has always been an leveraged index fund, or has been an leveraged index fund in a certain period, we make a note of this and label the fund as an leveraged index fund for all of its observations, or only during that period, respectively.

The following table describes the statistics on the "index status" distribution for "suspicious" funds that we manually checked and when necessary relabeled.

suspicious fund type	looks like index but not labeled	labeled but does not look like index
# total	61	596
# index	3	153
# leveraged index	5	23
# actively managed	52	407
# changed status during life of fund	1	13

To verify that our approach is successful at identifying suspicious index funds that Morningstar labeled, we also manually checked a random sample of 100 funds that Morningstar identified as an index fund, and has not been yet identified by our previous algorithms as suspicious. Intuitively, these are funds that have successfully passed our check for suspicious funds, and if our check is effective, we would need most of funds in this list, if not all, to be ordinary index funds. We find that in the 100 funds random sample we checked, 90 were manually identified to be real ordinary index funds. Another 7 were identified to be leveraged index funds. The remaining 3 were actively managed. The statistics show that we are not doing very well in identifying leveraged index funds. Consequently, we find that it is necessary to add an additional step to more effectively identify leveraged index funds from the Morningstar database.

Next, we thus search for leveraged index funds and separate them from the ordinary index funds and actively managed funds. To do so, we first make a list of words that indicate a fund to be a leveraged index fund. This list is given in the table below, which we refer to as Table 3.

TABLE 3: key words to identify leveraged index funds from index funds					
Inverse	short	ultra	2x	3x	4x
5x	6x	7x	8x	9x	0x

Note that the words in Table 3 are selected based on our experience in reading the prospectuses of funds in the previous steps. In the earlier steps, we found a total of 28 leveraged index funds and they all contained one of the words listed in Table 3. We also use our experience from reading prospectuses in the earlier steps to create Table 4 below, which contains a list of words that indicate a fund is not a leveraged index fund, even if that fund has some of the words listed in Table 3.

TABLE 4: key phrases to refute a fund is an leveraged index fund					
short term	short tm	short bond	short bnd	long short	lg short

We judge that if a fund in Morningstar contains at least one of the words in Table 3, and does not contain any of the words in Table 4, we would label it as a leveraged index funds.

We have already taken care of identifying index funds for funds in Morningstar. But some of the funds in our combined database are only from CRSP and are not merged to Morningstar. This issue only applies to funds that are never merged to Morningstar. If a fund in CRSP started prior to 1996 but is later merged to Morningstar, we labeled it using the Morningstar information throughout its life. Consequently, we need to build a mechanism to identify index funds in CRSP as accurately as we can. Since CRSP does not have an internal label for index funds, we have to build our own label from scratch for these funds using the following criteria:

1. For a fund, if any of its *crsp_fundnames* satisfies a) and b) as defined below, we label it as a leveraged index fund.
   a) Contains a word in Table 3
   b) Does not contain any of the words in Table 2 and Table 4
2. For a fund, if it is not labeled as a leveraged index fund by (1), and if any of its *crsp_fundnames* satisfies a) and b) as defined below, we label it as an ordinary index fund.
   a) Contains a word in Table 1
   b) Does not contain any of the words in Table 2, Table 3 and Table 4.
3. For a fund, if it is not labeled as an index fund by either (1) or (2), we label it as an actively managed fund.

Note that for funds that appear in Morningstar, since we manually checked each "suspicious" fund that appeared, we are able to identify funds that changed from being an actively managed fund to an index fund, or vice-versa. However, in categorizing funds that only appear in the CRSP database, we have been assuming that if a fund appears to be an or leveraged index fund (or ordinary index fund) for any one month, it must be a leveraged index fund (or ordinary index

fund) for all periods. This assumption is not necessarily true, because we realized when doing the manual check on "suspicious" Morningstar funds that a small subset of the funds do change their index fund status. So the objective of this last step is to pick out such funds that changed their "index status". To achieve this, we first create a list of funds that we suspect to change its "index status" in our database. The mechanism for identifying such "suspicious" funds is as follows:

- If a fund that only existed in CRSP, judging from its *crsp_fundname*, appears to be having a fixed "index status" in every month before a certain period, but appears to be having a different "index status" in every month after that period, and if both periods before and after that month last for more than 12 months, we label that fund as "suspicious".

We found a list of 149 such funds that are suspicious of changing their "index status". To verify, we went back to EDGAR to check if these funds do actually change their index status. If they do not, then our database's labeling is correct. If they do, we modify our database accordingly to reflect this fact. We found that in those 149 funds, 16 changed their status. The following table describes this in more detail.

type of change in "index status"	# of funds
actively managed to ordinary index	9
ordinary index to actively managed	5
leveraged index to actively managed	2
actively managed to leveraged index	0
ordinary index to leveraged index	0
leveraged index to ordinary index	0

As a last check, we take all funds we so far classified as ordinary index funds. For these funds that we have not yet manually checked on EDGAR, we go to Bloomberg to verify that the fund is indeed an index fund. Bloomberg provides a short description of the investment objective for each fund in its database. A total of 1520 funds are checked this way. Of those we can identify on Bloomberg, we find that 42 funds we classified as index funds are reported as actively managed by Bloomberg. For these 42 funds, we went back to EDGAR to clarify their real status. Only 29 of the 42 funds are found on EDGAR, but all 29 cases EDGAR is consistent with Bloomberg. That is, all of these 29 funds are also reported in EDGAR as actively managed funds. Using this information, we judged that Bloomberg was accurate and relabeled all those 42 funds as actively managed funds.

The final distribution of funds and observations by "index status" is given in the table below:

index status	# of funds	# of observations
ordinary index funds	3060	172007
leveraged index funds	1557	138432
actively managed funds	47599	4048862
funds that changed status	30	n/a

Correction of Monthly Returns

There is a significant number of observations for which the monthly return reported by Morningstar and the monthly return reported by CRSP differ. The combined database contains a total of 4525081 observations, of which 2357848 observations have both *mret* and *totret1mo* reported. Of these, 60831 observations (2% of total observations) have *mret* (the CRSP reported monthly return) and the *totret1mo* (Morningstar reported monthly return) differ significantly (more than 10 basis points). Details on the differences between *totret1mo* and *mret* can be found in the table below:

Difference between mret and totret1mo	# of observations	% of observations
Do not differ	2152604	91%
1 basis point	4057	0.2%
2-10 basis points	140356	6.1%
11-100 basis points	40755	1.7%
> 100 basis points	20076	1.0%

In this section, we use the terms "differing significantly" or "inconsistent" when the absolute difference in the monthly return reported by Morningstar and by CRSP is bigger than 10 basis points (for example, one number is 2.03% and the other number is 2.14%). To ensure accuracy in our database, we decided to make corrections on these 60831 observations. Our correction mechanism in this section can be divided into four steps.

Step One

We apply several automated correction mechanisms to these inconsistent monthly returns. First, we recognize that both CRSP and Morningstar report funds' net asset values (NAV) and sometimes also report dividend values. From these NAVs, we can compute two additional sets of monthly returns, one from the NAV reported by Morningstar and one from the NAV reported by CRSP, which we will now call ms_ret and crsp_ret, respectively. More specifically, they are calculated as:

$$crsp\_ret_{i,t} = \frac{crsp\_nav_{i,t} + crsp\_dividend_{i,t} - crsp\_nav_{i,t-1}}{crsp\_nav_{i,t-1}}$$

$$ms\_ret_{i,t} = \frac{ms\_nav_{i,t} + ms\_dividend_{i,t} - ms\_nav_{i,t-1}}{ms\_nav_{i,t-1}}$$

The dividend value is missing. We apply the following set of rules to fill in the dividend values as best as we can:

1) If dividend is missing in one database (either CRSP or Morningstar), but not the other, then we fill in the dividend value for that database using the dividend value of the other database.
2) If (1) cannot resolve the missing dividend problem for an observation, we assume the dividend paid for that observation is 0.
3) If under the assumption in (2), we find that the difference between *mret* and *crsp_ret* is equivalent to the difference between totret1mo and *ms_ret*, then we can infer that the

difference is caused by dividends and since the two differences are consistent, the inferred dividends of the two databases are consistent, and we fill in the difference as the dividend ratio. In the following example, note although dividends are missing, the difference between *crsp_ret* and *mret* and the difference between *ms_ret* and *totret1mo* are both 0.07, indicating that the dividend ratio is 0.07.

Before:					
Mret	totret1mo	crsp_ret	ms_ret	crsps_dividend	ms_dividend
0.17	0.18	0.10	0.11	.	.
After:					
mret	totret1mo	crsp_ret	ms_ret	crsps_dividend	ms_dividend
0.17	0.18	0.10	0.11	0.07	0.07

Next, for a given observation with a monthly return inconsistency, we apply the following set of rules:

1. If *mret* is consistent with both *crsp_ret* and *ms_ret*, then we accept *mret* as the correct monthly return
2. If *totret1mo* is consistent with both *crsp_ret* and *ms_ret*, then we accept *totret1mo* as the correct monthly return
3. If *mret* is consistent with *crsp_ret* but not with *ms_ret*, and *totret1mo* is not consistent with *ms_ret*, we accept *mret* as the correct monthly return
4. If *totret1mo* is consistent with *ms_ret* but not with *crsp_ret*, and *mret* is not consistent with *crsp_ret*, we accept the *totret1mo* as the correct monthly return.
5.

This set of rules allows us to correct for 11319 return inconsistencies in the database.

Step Two

One major reason why there are still significant inconsistencies remaining is because there are many cases where the computed *crsp_ret* is consistent with *mret*, and the computed *ms_ret* is consistent with *totret1mo*, but the returns are inconsistent across the two databases. An example of such a case is presented below:

Year	month	Ticker	mret	totret1mo	crsp_ret	ms_ret
1997	7	ABESX	1.66	1.85	1.66	1.85

Consequently, we apply another set of rules to correct for the remaining return inconsistencies. To understand how this mechanism works, consider the following example.

year	month	Ticker	Mret	totret1mo	crsp_ret	ms_ret
2002	8	UGSBX	-3.22	-3.22	-3.22	-3.22
2002	9	UGSBX	4.01	4.01	4.01	4.01
2002	10	UGSBX	0.74	1.94	0.74	1.94
2002	11	UGSBX	1.33	1.33	1.33	0.13
2002	12	UGSBX	-1.07	-1.07	-1.07	-1.07

In this case, in 10/2002, *mret* is consistent with *crsp_ret*, *totret1mo* is consistent with *ms_ret*, but

totret1mo is not consistent with *mret*. This means that any correction mechanism described so far will fail to correct this inconsistency. This also means that in 10/2002, either CRSP or Morningstar must have reported both an incorrect net asset value and an incorrect return. So instead of finding which of the two databases reported an incorrect return, we search for which one of the two reported an incorrect NAV, and from it infer which return reported is mistaken. To do so, we sort the fund's data chronologically, and look above and below the observation with the inconsistency to see which database has inaccurately reported the NAV. Is crsp_ret consistent with mret at (t-1) or (t+1)? Is ms_ret consistent with totret1mo at (t-1) or (t+1)? In the example, crsp_ret and mret are consistent but ms_ret and totret1mo are inconsistent at 11/2002 (i.e. t+1). From this we deduct that mret is accurate in 10/2002.

What if consecutive months contain errors in NAV? We need to search above and below for more than one month, until we resolve the inconsistency or we are sure that the inconsistency cannot be resolved using this method. An example of such a case is given below:

year	month	ticker	mret	totret1mo	crsp_ret	ms_ret
1999	1	TECFX	4.41	4.41	4.41	4.41
1999	2	TECFX	-1.11	-1.11	-1.11	-1.11
1999	3	TECFX	7.26	7.26	7.26	5.26
1999	4	TECFX	1.73	0.73	1.73	0.73
1999	5	TECFX	0.26	-0.77	0.26	-0.77
1999	6	TECFX	3.71	3.71	3.71	3.71
1999	7	TECFX	-6.69	-6.69	-6.69	-6.69

Note that in both 4/1999 and 5/1999, *mret* is consistent with *crsp_ret* and *totret1mo* is consistent with *ms_ret*, but *mret* is not consistent with *totret1mo*. Using the approach we just described using the earlier example, we look above and below. Using what we have in 3/1999, we judge that Morningstar made a mistake in recording its NAVs on 3/1999. Consequently, we accept that *mret* is the correct monthly return for both 4/1999 and 5/1999. Using this mechanism as illustrated in the two examples above, we were able to correct an additional 17730 return inconsistencies.

Step Three

At this stage, we still have 31782 return inconsistencies which we were not yet able to resolve. To correct these inconsistencies, we manually check observations with inconsistencies using a third source: Bloomberg. Due to time constraints, we only check those observations for which the discrepancy more than 20 basis points. While we were manually correcting for inconsistencies, we find two issues worth noting:

1. Not every fund in our database is present in Bloomberg. For all the funds we attempted to check using Bloomberg, we were only able to find approximately 65% of them.
2. Even if a fund is in Bloomberg, we notice that the fund sometimes starts in Bloomberg at a later date than it does in our database. So, some of the earlier observations are missing.

Thus, we were not able to resolve every inconsistency of more than 20 basis points. We manually collected 11720 returns from Bloomberg, and use them to correct the inconsistencies in our

database as follows:

1) If Bloomberg's return is consistent with *mret* but not consistent with *totret1mo*, we accept *mret* as the correct monthly return for that observation.
2) If Bloomberg's return is consistent with *totret1mo* but not consistent with *mret*, we accept *totret1mo* as the correct monthly return for that observation.
3) If Bloomberg's return is not consistent with either *totret1mo* or *mret*, then we say that the inconsistency still cannot be resolved at this stage.

The distributions of how many of the manually collected returns belong to each of the three categories are given in the following table:

number of Bloomberg returns attempted to collect	14267
number of Bloomberg returns collected	5149
number of Bloomberg returns consistent with mret	2621
number of Bloomberg returns consistent with totret1mo	2055
number of monthly returns on Bloomberg inconsistent with both mret and totret1mo	473

Step Four

Finally, we apply the following set of rules utilizing this extra return information we obtained from Bloomberg to do another round of automated monthly return corrections, as follows:

1. If an inconsistency that is still unresolved is within a 2-year interval of two inconsistencies that were resolved, and in both of these two cases we accep*ted mret*, then we acc*ept mret* for this unresolved observation as well. Consider the example below, in year 3/1999 and 7/1999, we initially had two inconsistencies, but in both cases, through our earlier mechanisms, we were able to resolve these and accept *the mret* as the correct one in both cases. There is one more inconsistency in 5/1999, and is not resolved by our earlier mechanism. However, we recognize that it is in between two resolved inconsistencies that are less than 24 months apart, and those two inconsistencies are both assigned to CRSP. Consequently, at this stage we acc*ept mret* in 5/1999 as well.

initially inconsistent?	resolved to?	time	ticker	mret	totret1mo	bloomberg
yes	CRSP	3/99	AMMDX	3.33	3.03	3.33
		4/99	AMMDX	6.10	6.10	
yes	No	5/99	AMMDX	1.07	0.07	
		6/99	AMMDX	2.18	2.18	
yes	CRSP	7/99	AMMDX	-3.22	-4.22	-3.22

2. Similarly, we will resolve an unresolved inconsistency and accept Morningstar return to be correct if it is sandwiched by two closely related (not apart by more than 24 months) resolved inconsistencies and in both of these cases we also accepted Morningstar.

After implementing all these steps in resolving inconsistencies in monthly returns, we resolved

about 2/3 of all incorrect returns in the entire database. We are still left with 22627 inconsistencies that we cannot resolve by any means. This is about 0.5% of total observations in the database. In such cases, we decided to accept the return reported by CRSP. The following table gives the number of inconsistencies resolved or unresolved at each stage of the correction mechanism:

step	total number of observations	inconsistency to begin with	inconsistency resolved	inconsistency still left	inconsistency eventually left
1	452508	60831	11319	49512	
2			17730	31782	
3			4676	27106	
4			4479	22627	22627

Grouping Subclasses

To group together different subclasses of the same fund, we create a new variable called *group_id*. The objective of the algorithm in this section is to assign different subclasses of the same fund the same *group_id*, while assigning subclasses of different funds different *group_id*'s. What makes things complicated is that Morningstar and CRSP do not always agree on the set of subclasses of a particular fund. In addition, Morningstar uses a numerical definition of a subclass --- it contains a field in each observation that gives the total assets of all subclasses, so that all observations with the same total assets are part of the same subclass. However, the problem with using this field is that it is only been recently populated. We therefore use the name of the fund on both data bases to identify the subclasses and use *totalassetsmm* (total assets) of all subclasses reported on Morningstar to check the accuracy of our algorithm.

We begin by first providing a summary of what we do. The name of a mutual fund observation in either CRSP or Morningstar can be divided into two segments. The first segment contains the name of the actual fund, while the second segment contains the name of the specific subclass. Consequently, subclasses of the same fund should be identical in the first segment of their names. So we can use this first segment of their names to build an algorithm to group together different subclasses of the same fund. Because some of the funds in the combined database come from CRSP alone, some from Morningstar alone, and some from both, we apply a three-step process in grouping:

1. First we group funds that belong to the original CRSP database. These are observations that either only existed in CRSP or have been successfully merged. We save the grouping result in a variable named crsp_group_id.
2. Then we group funds that only belong to the original Morningstar database. So these are observations from Morningstar not matched onto any CRSP observations in the combined database. We save the grouping result in a variable named ms_group_id.
3. Finally, we group funds in the entire combined database using the results from steps 1 and 2. That is, we look for funds with subclasses spanning both subsets of the data by finding 1-to-1 correspondences between *crsp_group_id* and *ms_group_id*. This information on final grouping is saved in the variable *group_id.*

The CRSP database uses a much more standardized format to denote subclasses in its fund name than Morningstar does. In CRSP, the subclass portion of the name is always separated from the rest of the fund name using a ";" or a "/", while for Morningstar, there is no mark that indicates where the subclass portion of the fund name is precisely located. Consequently, we first use the fund name in CRSP (we store CRSP fund name in a new variable *crsp_fundname*) to group together different subclasses of the same fund, and only use the Morningstar fund name (we store Morningstar fund name in a new variable *ms_fundname*) when *crsp_fundname* is not available.

To begin this process, we separate out the first segment of the fund name in CRSP (which we store in a variable called *crsp_mainname*), and the subclass portion of the name (which is stored in *crsp_subclass*). We use the following set of rules to separate these two parts of *crsp_fundname*:

1. If a *crsp_fundname* contains ";" and the phrase after the last ";" does not contain "/", we use the entire phrase after the last ";" as the *crsp_subclass*, and the remaining portion located prior to the last ";" as *crsp_mainname*.
2. If a *crsp_fundname* contains "/", and the entire phrase after the last "/" does not contain space (is a single word) and does not contain ";", we use the word after the last "/" as the subclass, and the remaining portion before the last "/" as *crsp_mainname*.
3. If neither condition 1 nor 2 is met, we judge that this *crsp_fundname* does not have a subclass. So for this observation crsp_subname is left empty and its entire name is used as the *crsp_mainname*.

Note that 1, 2, and 3 are mutually exclusive, preventing multiple possibilities in assigning subclasses.

After having separated the *crsp_mainname* and the *crsp_subclass* portion of the *crsp_fundname*, we can proceed to grouping together different subclasses of the same fund because they should have the same *crsp_mainname*, but different *crsp_subclass*. However, we notice that for a fund, its name can change over time, and CRSP sometimes is inconsistent in the way it names a fund, and occasionally makes mistakes in reporting *crsp_fundname*. Consequently, we have to develop a grouping algorithm to best overcome these problems. We generate a variable called *crsp_group_id* to denote the grouping for observations that appear in CRSP (this excludes observations from Morningstar alone). In our algorithm, we assign two observations to have the same *crsp_group_id* if one of the following conditions applies:

a) *crsp_fundname* of both observations are non-empty, and the *crsp_mainname* of one observation is the same as the *crsp_mainname* of the other observation.
b) The tickers of both observations are non-empty, and the *ticker* of one observation is the same as the *ticker* of the other observation.

Note that using this algorithm, we can ensure that:

1) Only funds that originally appeared in the CRSP database get grouped and assigned a *crsp_group_id* --- a fund that only appears in Morningstar (that is, those observations in Morningstar not matched to any CRSP observations) will not be grouped with other funds because its *crsp_fundname* is empty.
2) Two observations from the same fund (having the same non-empty ticker) will never get grouped into different *crsp_group_ids*.
3) Two funds are grouped together as long as any of the *crsp_mainnames* of one fund matched exactly with any of the *crsp_mainnames* of the other fund, and both *crsp_mainnames* are non-empty.
4) This algorithm is transitive. That is, if the algorithm finds that fund A belongs to the same group with fund B, and fund B belongs to the same group with fund C, then fund A is also grouped together with fund C, even if their *crsp_mainnames* never matched .

As a way to check for the accuracy of this algorithm, we used the fact that two different subclasses of the same fund should have identical *totalassetsmm* (total assets) reported by Morningstar during the same year and month. Consequently, we checked how many times *tickers*

with different, non-missing total assets reported for the same year and month been assigned the same *crsp_group_id*. We found 93 *crsp_group_ids* (<0.5%) with such an inconsistency. We then manually checked 20 of them. What we found is that none of these cases were due to a problem in the algorithm. Instead, we found that these are borderline cases where CRSP regards the tickers to be subclasses of the same fund while Morningstar regards some of these tickers to be different funds. Note that we decided to always side with CRSP when there is an unresolvable inconsistency between CRSP and Morningstar, so in these 97 cases, we stay with CRSP's definition of a subclass. This completes step one of our three-step process.

Up until now, we have not yet grouped subclass observations that only enter in Morningstar (not in CRSP). Consequently, we still have not fully identified groups that contain some subclasses in Morningstar only. Such groups can be divided into two categories:

1. Groups that have some subclasses that only enter in Morningstar's database, and some subclasses that only enter in CRSP's database.
2. Groups for which all subclasses only enter in Morningstar's database.

To proceed to step two of the mechanism, we make use of the fund name reported by *ms_fundname* (Morningstar fund name) to group these funds. As we have done with *crsp_fundname*, we need to divide *ms_fundname* into two segments, *ms_mainname* and *ms_subclass*, and perform grouping based on identical *ms_mainnames*. However, it is much harder to separate the two segments in *ms_fundname* than in *crsp_fundname*. We first attempted to check if we could take the last two words (1) or the last word (2) in *ms_fundname* as their subclass. We find that using the last two words as the subclass creates lots of inaccuracy, because *ms_fundname* tends to be very concise. For example, this strategy will mistakenly put the following different funds into subclasses of the same fund.

GS Structured SmCap Eq IR
GS Structured SmCap Gr IR

Also, we find that using the last word as the subclass also will not work. This is mainly because *ms_fundnames* sometimes omits spaces in between words, so some *ms_fundnames* have the last word containing more than the subclass information. For example:

GS Balanced StratA
GS Balanced InvA

Due to these constraints we apply a different strategy, which is outlined below:

1) Before proceeding onto the following steps, if the last two words of *ms_fundname* is "Load Waived", then remove these last two words from the name and use this truncated name as the new *ms_fundname*.
2) Before proceeding onto the following steps, if the last word of *ms_fundname* is "LW", then remove this last word from the name and use this truncated name as the new *ms_fundname*
3) If the last word of the *ms_fundname* is any of the words in table 9, we accept its last word as the subclass name. We denote the rest of the name with this last word removed as

ms_mainname, which is the name of the fund after the subclass information is removed.

4) If the last letters of the *ms_fundname* exactly match any words of the words listed in table 9, then we take these letters as the *ms_subclass* and use the rest of the *ms_fundname* with these letters removed as the *ms_mainname*.
5) For a *ms_fundname*, if we could not find that it contains subclass information using either 3) or 4), we conclude that this name does not have a subclass section.

The key words used to identify subclasses in Morningstar are:

Table 9:							
1 LETTER:	A	B	C	D	E	F	G
	H	I	J	K	L	M	N
	O	P	Q	R	S	T	U
	V	W	X	Y	Z		
2 LETTERS:	AA	A1	A2	A3	B1	B2	B3
	C1	C2	C3	S1	S2	S3	RI
	R1	R2	R3	R4	R5	II	Sv
	EC						
3 LETTERS:	AAA	Svc	Ins	SAI	(A)	(B)	(C)
	III	Inv					
4 LETTERS:	Inst						
5 LETTERS:	Instl						

We developed this table of words indicating subclasses in Morningstar using the following procedures:

1. We first make a list of unique *ms_fundnames* that exist in our database. Then we extract from them the last word. There are 2463 such words that have been used as the last word in at least one fund name.
2. We then manually look through these 2463 words to see which of them are used to denote a subclass. We isolate these, and they are given in the table above.

We have separated the *ms_fundname* into *ms_subclass* and *ms_mainname*. To group observations only in Morningstar, we generate a variable called *ms_group_id*. Now we will try to group together subclasses for observation only in Morningstar but not in CRSP, using the following steps:

1) If two observations have the same, non-empty ticker, and the ticker is used by observations only found in Morningstar but not in CRSP, then we will assign them the same *ms_group_id*
2) If two observations have the same *ms_mainname*, and both funds are only found in Morningstar, then we will assign them the same *ms_group_id*.

So far we have taken care of the following two types of groups:

a. Groups that only contains funds coming from Morningstar alone. These groups have a unique *crsp_group_id* and no *ms_group_id.*
b. Groups that only contains funds coming from CRSP. These groups have a unique *ms_group_id* and no *crsp_group_id*.

There is still one more possibility that needs to be considered. That is, fund groups that contain some subclasses that are only reported in the Morningstar database, but also some subclass from CRSP. Consider the following example, where a fund has five subclasses, two of which are in Morningstar only, one of which is in CRSP only, and the remaining two are in both CRSP and Morningstar.:

fromCRSP	crsp_fundname	ms_fundname	crsp_group_id	ms_group_id
-1		Example Fund A		
-1		Example Fund B		
0	Example Fund C	Example Fund C		
0	Example Fund D	Example Fund D		
1	Example Fund E			

In this example, *fromCRSP* = -1 indicates observations only from Morningstar, 0 indicates observations from both CRSP and Morningstar and 1 indicates observations from CRSP only. In the first step, we took the observations in CRSP (i.e. *fromCRSP* = 0 or 1) and based on *crsp_fundname* we assigned them *crsp_group_id*. Say that we assigned the Example Fund a *crsp_group_id* of C1:

fromCRSP	crsp_fundname	ms_fundname	crsp_group_id	ms_group_id
-1		Example Fund A		
-1		Example Fund B		
0	Example Fund C	Example Fund C	C1	
0	Example Fund D	Example Fund D	C1	
1	Example Fund E		C1	

In the second step, we took the observations in Morningstar (i.e. *fromCRSP* = 0 or -1) and based on *ms_fundname* we assigned them *ms_group_id*. Say that we assigned the Example Fund a *ms_group_id* of M1:

fromCRSP	crsp_fundname	ms_fundname	crsp_group_id	ms_group_id
-1		Example Fund A		M1
-1		Example Fund B		M1
0	Example Fund C	Example Fund C	C1	M1
0	Example Fund D	Example Fund D	C1	M1
1	Example Fund E		C1	

However, because for this particular Example Fund, some of its subclasses are only in Morningstar and some only in CRSP, we need to write an algorithm that recognizes the 1-to-1

correspondence between M1 and C1.

So to construct the *group_id* for the entire combined database, we apply the following rules:

1. If an observation with a non-empty *ms_group_id* has the same *year, month,* and non-empty *ms_mainname* as an observation with a non-empty *crsp_group_id*, then we say that the *ms_group_id* and the *crsp_group_id* are "linked". In the example above, because Example Fund C and Example Fund D are in both CRSP and Morningstar, we say that C1 and M1 are linked.
2. Note a *ms_group_id* can be linked to multiple *crsp_group_ids*, and a *crsp_group_id* can also be linked to multiple *ms_group_ids*.
3. If a crsp_group_id is linked to only one *ms_group_id*, and that *ms_group_id* is not linked to any other *crsp_group_ids*, we replace that *ms_group_id* with the *crsp_group_id* it is uniquely linked to.
4. If a *crsp_group_id* is linked to multiple *ms_group_ids*, but none of those *ms_group_ids* are linked to any other *crsp_group_ids*, we replace all these *ms_group_ids* with the *crsp_group_id* they are linked to.
5. Up to now, each observation still has either a *ms_group_id* or a *crsp_group_id*, but never both. So for observations with non-empty *ms_group_id*, we fill in their *group_id* variable using *ms_group_id*. For other observations, we fill in their *group_id* using *crsp_group_id*.

For the particular example we gave, because it is easy to see a 1-to-1 link between M1 and C1, we combine C1 and M1 in creating *group_id*:

fromCRSP	crsp_fundname	ms_fundname	crsp_group_id	ms_group_id	group_id
-1		Example Fund A		M1	1
-1		Example Fund B		M1	1
0	Example Fund C	Example Fund C	C1	M1	1
0	Example Fund D	Example Fund D	C1	M1	1
1	Example Fund E		C1		1

However, because Morningstar fund names are not standardized, despite our efforts it is still possible that a mistake is made in assigning *ms_group_id*. Consider the following extension to the example above:

fromCRSP	crsp_fundname	ms_fundname	crsp_group_id	ms_group_id
-1		1st Example Fund A		M1
-1		1st Example Fund B		M1
-1		2nd Example Fund A		M1
0	1st Example Fund C	1st Example Fund C	C1	M1
0	1st Example Fund D	1st Example Fund D	C1	M1
0	2nd Example Fund B	2nd Example Fund B	C2	M1
1	1st Example Fund E		C1	

In this extension to the example, we have two separate funds: 1st Example Fund and 2nd Example Fund. However, suppose that for some reason, there is a mistake in assigning

ms_group_id and the two funds are classified into subclasses of the same fund (as shown in the table above), all assigned an id of M1. But in *crsp_group_id*, the classification is correct. If we compare the *crsp_group_id* and *ms_group_id*, we can see that C2 is linked to M1 by 2nd Example Fund B, and C1 is linked to M1 by 1st Example Fund C and D. This means that M1 is linked to both C1 and C2. There is no 1-to-1 correspondence. Consequently, we take the *crsp_group_id* and fill in *ms_group_id* only when *crsp_group_id* is missing.

fromCRSP	crsp_fundname	ms_fundname	group_id
-1		1st Example Fund A	M1
-1		1st Example Fund B	M1
-1		2nd Example Fund A	M1
0	1st Example Fund C	1st Example Fund C	C1
0	1st Example Fund D	1st Example Fund D	C1
0	2nd Exmaple Fund B	2nd Example Fund B	C2
1	1st Example Fund E		C1

The intuition is that for groups with some funds in only Morningstar and some funds in CRSP, it will have a *crsp_group_id* and a *ms_group_id*. So the task of this section is to recognize and merge these *crsp_group_id-ms_group_id* pairs. However, we trust the accuracy of our *crsp_group_id* much more than we trust the accuracy of our *ms_group_ids*, because Morningstar names are much messier. Consequently, if one *ms_group_id* is linked to more than one *crsp_group_id*, we judge that this *ms_group_id* contains a mistake. This is why when one *ms_group_id* runs into multiple *crsp_group_id's*, we do not combine these *crsp_group_id's* into one. We believe CRSP to be accurate, yet we have no other alternative information than the Morningstar name to deal with funds only in Morningstar. Consequently, in such a case we leave it as it is. In this step, we found that there are 43 *ms_group_ids* that are linked to more than one *crsp_group_id*.

The distribution of the number of subclasses of funds in our database is as follows.

Funds with X number of subclasses (X:)	# of observations	% observations
1	1093197	25.85%
2-5	2535286	56.95%
6-15	678527	16.68%
16-17	16952	0.40%
18	2970	0.07%
19	1254	0.03%
20	420	0.01%
21	546	0.01%
22	176	0.00%
23	138	0.00%
29	116	0.00%

We were suspicious of the funds with over 20 subclasses and believed we may have made mistakes in our classification. We checked all funds with 18 or more subclasses in any given year. There are 9 such funds. For these 9 funds, we find 3 cases in which two separate funds are classified as subclasses of the same fund. In all 3 cases, the reason for this misclassification is

that a *ticker* changed from being the subclass of one fund to being a subclass of another fund. In our algorithm, we assume that *tickers* remain the subclasses of the same funds. The three misclassifications are given below.

Of the other 6 funds, we find that they indeed have a large number of subclasses. A list of these incorrectly classified groups which should belong to subclasses of different funds are given below:

Principal Funds, Inc: Mortgate Securities Fund
Principal Funds, Inc: Government & High Quality Bond Fund

WM Trust I: US Government Securities Fund
WM Trust II: International Growth Fund

Lord Abbet Investment Trust: Lord Abbet Total Return Fund
WM Trust I: Income Fund

After the manual correction, the final distribution of the number of subclasses of funds in our database is as follows.

Funds with X number of subclasses (X:)	# of observations	% observations
1	1092559	25.91%
2-5	2323625	55.07%
6-15	780587	18.51%
16-17	15298	0.42%
18	2782	0.06%
19	627	0.03%
20	160	0.00%

Filling in Missing Expense ratios

We recognize that all expense ratios provided by either Morningstar or CRSP are annual numbers. Consequently, if the expense ratio value is missing in CRSP and in Morningstar for some months of a fiscal year, we use other observations within the same fiscal year to fill in the missing observations. The month fiscal year begins for a fund is provided to us by CRSP.

If the fund does not have fiscal year end information, we apply the following procedure in assigning it a fiscal year end:

1) If a fund never had a fiscal year end reported, but instead has at least 5 consecutive quarters of expense ratios reported, and the month at which the expense ratio change is unique and always consistent. Then we say this month is the fiscal year end month for the corresponding fund.
2) If a subclass does not have its fiscal year end reported, but another subclass of the same fund has its fiscal year end reported, then we fill in the fiscal year end of this fund using the value reported by the other subclass.
3) If a fiscal year end still cannot be determined, we manually inputted the data by consulting the SEC online database on EDGAR.
4) If all previous steps fail and we cannot find the fund on EDGAR, we assume that the fiscal year end of the fund is December.

To fill in expense ratios manually that are still missing, we isolate funds that:

a. Have an non-empty ticker
b. Existed for 24 months or more in our database.
c. Have reached an AUM of $5 million (real)

Due to time constraints, we will only manually fill in expense ratios for funds that satisfies a)-c). The statistics for the number of missing expense ratios at each stage are given below:

number of missing expense ratios to begin with (by months)	1716117
number of missing expense ratios after filling in within fiscal year (by months)	493086
number of missing expense ratios from funds satisfying a)-c) (by months)	246754
number of missing expense ratios from funds satisfying a)-c) (by year)	24011
number of funds satisfying a)-c) with missing expense ratios	8472

We then go to the actual company filings to fill in the missing expense ratios for the funds that satisfy a)-c). These filings can be found on the EDGAR database online. We make use of the EDGAR search engine to find the forms: if the fund has a ticker, we would search for the ticker, and if the fund in our database does not have a ticker, we would search for it using some key words from its fund name. Annualized expense ratios are reported in the N-CSR and N-SAR forms. Occasionally, when a fund began its operation during the middle of the fiscal year, only semi-annualized expense ratio is reported. In such case, we double the semi-annual figure to input into our database.

For each expense ratios searched, we made the following classifications:

1. The expense ratio is found
2. The expense ratio is not found because the fund is not tracked by SEC. (case I)
3. The expense ratio is not found because although the fund is in SEC, the subclass is not. (case II)
4. The expense ratio is not found because although the subclass is in SEC, expense ratio is not reported for that year (either because the fund started in our database before it started on SEC or the fund ended in SEC before it ended in our database). (case III)

To note the different cases in our database, we denote case I with ".", case II, with "-999", and case III with "-9999". The statistics for the distribution of these four classifications are as follows:

Year	% found	% not found (case I)	% not found (case II)	% not found (case III)
Before 1995	24%	62%	6%	8%
1996-2000	63%	23%	2%	3%
2001-2005	77%	19%	2%	2%
2006-2009	74%	21%	2%	3%
Cumulative	74%	20%	2%	4%

The final statistics on missing and filled in expense ratios are:

number of observations with expense ratios filled in	182598
number of observations with expense ratios not found (total)	64156
number of observations with expense ratios not found (case I)	49481
number of observations with expense ratios not found (case II)	4344
number of observations with expense ratios not found (case III)	10331

Adding Variables to the Combined Database

In this section we summarize the variables that we have added to the combined database in the earlier sections since "Adding Variables to the Combined Database (First Addition)". A description of these variables are given below:

- *MS_ticker* - This is a variable that stores the original ticker information in the raw Morningstar database before any corrections or modifications were done to it.
- *CRSP_ticker* - This is a variable that stores the original ticker information in the raw CRSP database before any corrections or modifications were done to it.
- *fromCRSP* - This variable is assigned a -1 if the observation in the combined data is not matched and only comes from Morningstar; the variable is assigned a 0 if the observation is matched; and the variable is assigned a 1 if the observation is not matched and only comes from CRSP.
- *obsnum* - This variable takes on integer values and denotes the number of months of data we have in our database for a fund. Note that ticker is used as the identifier for fund, and so obsnum holds an empty value in any observation with an empty ticker.
- *cpiid* - This variable is the inflation index that is obtained directly from Fama_French_Inflation and is given as a monthly data. It is the value of a 1/1/2000 dollar. The purpose of this variable is used to calculate the inflation adjusted net asset value for each observation.
- *tna* - A variable denoting the total net assets for observations in the combined database. If an observation has a total net assets reported (and the reported value is not -99), we use it for tna. Otherwise, we use the Morningstar reported value. If neither database reports a valid value (or the value is -99) for total net assets, this variable is left blank.
- *rsize* - This variable holds the inflation adjusted net assets value for each observation. It is calculated using *tna*/*cpiid.*
- *under_size* - This variable is only assigned if the *ticker* is not empty. If a fund's *real_size* exceeded 5 million, then starting from the first month its *real_size* exceeded 5 million, its labeled with a 1, and it is labeled a 0 anytime before that. If a fund never reached 5 million, it is labeled a 0 for all its observations. For observation having empty tickers, *under_size* is also empty.
- *return* - A variable that stores the monthly returns for the database after correction. For inconsistency in the database that can not be resolved between *mret* and *totret1mo*, *mret* is used.
- *return_correction* - This variable is used to denote if monthly return is inconsistent and if a correction is made to the monthly returns. If *return_correction* equals 0, no change is made. If *return_correction* equals 1, a correction is made by an automatic correction procedure before using Bloomberg as the third reference source. If return_correction equals 2, a correction is made by using Bloomberg, either directly or through a automatic procedure based on the returns from Bloomberg. If *return_correction* equals 3, the inconsistency is not resolved.
- *bloomberg_ret* - stores all returns manually extracted from Bloomberg.
- *expratio* - This variable stores the annualized expense ratios. It is back and forward filled and contains all the corrections made. (I adjusted the code so that the CRSP exp_ratio does not get replaced but instead a new variable with the corrections is created. But this has not run yet)
- *expratio_add* - This variable contains all expense ratios manually collected. (again not yet

implemented but will be run in the next file uploaded)

- *fiscal_month* - This variable stores the fiscal month of the fund. If the fiscal month is missing, we assume that it is December.
- *group_id* - This variable groups together subclasses of the same fund. It means that every single *group_id* is used to denote a unique portfolio, which can contain a collection of subclasses.
- *indx* - Denotes the "index status" of a fund. If *indx* equals 0, the fund is actively managed. If *indx* equals 1, the fund is an ordinary index fund, and if *indx* equals 2, the fund is an ultra index fund.