Data Appendix for
Scale and Skill in Active Management

Lubos Pastor, Robert F. Stambaugh, Lucian A. Taylor

August, 2013

Table of Contents

1. Raw CRSP Database Cleanup and Merge

First, we merge the raw CRSP data files to create a final dataset that contains all the relevant variables for our study. The raw SAS files are downloaded directly from the WRDS server, under *wrds/crsp/sasdata/q_mutualfunds*. Table 1 lays out the structure of each raw dataset. We start the merging process from the return dataset. In each step of the merge, we merge in only observations of the year/month when return data exists. If a variable is missing, we keep the year/month observation and record a missing value for that variable.

After the merge, we follow pages 2-3 of Berk & Binsbergen's (2011) Data Appendix[1] to clean up the ticker variable. We reproduce these pages of their Appendix verbatim at the end of this document.

In the raw CRSP *fund_hdr* and *fund_hdr_hist* files, ticker is named *nasdaq*.

Next, we merge in the total US stock market capitalization for each year/month observation. We use the WRDS web interface to download this data. In particular, we download monthly *price* and *shares outstanding* data for all US stocks (from 1960 to present) with *shrcd*=10 or 11. Stock market capitalization is then calculated as the multiple of *price* and *shares outstanding*. We then sum all the stocks' market capitalization to generate the total US stock market capitalization (variable *sum_me).*

Next, we create additional variables, including one-month-lag variables of
- *mtna (total net asset),*
- *nav (net asset value per share),*
- *exp_ratio (expense ratio),*
- *turn_ratio (turnover ratio),*
- *sum_me (total US stock market capitalization),*

and

- *first_offer_dt (date the fund was first offered)*
- *fund_age (age of a fund at month end, in days)*
- *manager_tenure (tenure of a fund manager, in days),*
- *first_CRSP_dt (date when a fund first shows up in CRSP).*

Finally, we rename the variables so that each variable from CRSP dataset has "*_CRSP*" in its name. Table 2 shows the content of our final CRSP dataset. The final dataset contains 4,784,162 observations, covers 50,536 funds, and spans the time period from

[1] DOCUMENTATION - Measuring Managerial Skill in the Mutual Fund Industry, by Jonathan B. Berk and Jules H. van Binsbergen (version dated 06/09/2011).

January/1960 to June/2012.

Table 1: CRSP Raw File Structure

Raw CRSP File	**Variables**	**Variable Label**	**Dataset Duration**	**# of Obs**	**# of Funds (*crsp_fundno*)**
monthly_returns	*crsp_fundno*	Fund Identifier	01/29/1960 ~ 06/29/2012	4,784,162	50,536
	caldt	Date			
	mret	Total Return per Share as of Month End			
monthly_tna	*mtna*	Total Net Assets as of Month End	12/29/1961 ~ 06/30/2012	4,503,790	50,487
dividends	*dividend_CRSP*	Monthly Dividends, created from summing different types of dividend per fund/year/month.	12/15/1960 ~ 06/30/2012	2,271,346	46,124
monthly_nav	*mnav*	Monthly Net Asset Value Per Share	01/29/1960 ~ 06/29/2012	4,782,231	50,536
fund_style	*si_obj_cd*	Strategic Insight Objective Code	12/29/1961 ~ 06/29/2012	200,928	50,121
	wbrger_obj_cd	Wiesenberger Fund Type Code, Identifying Fund Strategy			
	lipper_class	Lipper Classification Code			
	lipper_obj_cd	Lipper Objective Code			
	lipper_class_name	Lipper Classification Name			
	policy	Type of Securities Mainly Held by Fund			
fund_summary	*per_com*	Amount of fund invested in Common Stocks	12/29/1961 ~ 06/29/2012	1,260,875	50,579
	per_pref	Amount of fund invested in Preferred Stocks			
	per_conv	Amount of fund invested in Convertible Bonds			
	per_corp	Amount of fund invested in Corporate Bonds			
	per_bond	Amount of fund invested in all Bonds			
	summary_period	Frequency			

Table 1: CRSP Raw File Structure, Continued

Raw CRSP File	**Variables**	**Variable Label**	**Dataset Duration**	**# of Obs**	**# of Funds (*crsp_fundno*)**
fund_fees	*actual_12b1*	12b-1 Fee	12/29/1961 ~ 04/30/2012	323,472	42,289
	max_12b1	Maximum 12b-1 Fee			
	exp_ratio	Expense Ratio as of Fiscal Year-End			
	mgmt_fee	Management Fee			
	turn_ratio	Fund Turnover Ratio			
	fiscal_yearend	Effective Date for Fees			
front_load	*front_load*	Max Defer & Rear Load Charges	12/29/1961 ~ 06/22/2012	93,654	19,369
rear_load	*rear_load*	Maximum Rear-End Load	12/29/1961 ~ 06/22/2012	143,223	26,761
	crsp_portno	Portfolio Identifier			
fund_hdr & fund_hdr_hist	*ncusip*	Fund CUSIP	12/29/1961 ~ 06/29/2012	410,110	50,611
	nasdaq	NASDAQ Ticker Symbol			
	first_offer_dt	Date the Fund Was First Offered			
	mgr_name	Manager Name			
	mgr_dt	Date Current Portfolio Mgr Took Control			
	fund_name	Fund Name			
	mgmt_name	Management Company Name			
	mgmt_cd	Management Company Number			
	open_to_inv	Open to Investors			
	retail_fund	Retail Fund Indicator			
	inst_fund	Institutional Fund Indicator			
	index_fund_flag	Index Fund Indicator			
	dead_flag	Dead Fund Indicator			
	delist_cd	Identifies the reason for the fund delisting			

Table 2: Final CRSP Dataset Contents

Variable	Variable Label
actual_12b1_CRSP	12b-1 Fee in CRSP
asset_lag_CRSP	One Month Lagged Total Net Assets as of Month End in CRSP
assets_CRSP	Total Net Assets as of Month End in CRSP
begdt_CRSP	Beginning Date of Fee Reporting in CRSP
crsp_fundno	Fund Identifier in CRSP
cusip_CRSP	CUSIP from CRSP in CRSP
dead_flag_CRSP	Dead Fund Indicator in CRSP
delist_cd_CRSP	Identifies the reason for the fund delisting: L=Liquidation, M=Merge in CRSP
dividend_CRSP	Monthly Dividend Distributed in CRSP
enddt_CRSP	Ending Date of Fee Reporting in CRSP
exp_ratio_CRSP	Expense Ratio as of Fiscal Year-End in CRSP
exp_ratio_lag_CRSP	One Month Lagged Expense Ratio
first_CRSP_dt	First Date a Fund Shows Up in CRSP
first_offer_dt_CRSP	Date the Fund Was First Offered in CRSP
fiscal_yearend_CRSP	Fiscal Year End of Fee Reporting in CRSP
front_load_flag_CRSP	Flag for a Fund with Positive Front Load in CRSP
fund_age_CRSP	Age of a Fund at Month End in CRSP
fund_name_CRSP	Fund Name in CRSP
index_fund_flag_CRSP	Index Fund Indicator in CRSP
inst_fund_CRSP	Institutional Fund Indicator in CRSP
lipper_class_CRSP	Lipper Classification Code in CRSP
lipper_class_name_CRSP	Lipper Classification Name in CRSP
lipper_obj_cd_CRSP	Lipper Objective Code in CRSP
logret_CRSP	Log Monthly Return in CRSP
manager_tenure_CRSP	Tenure of Manager at a Fund in CRSP
max_12b1_CRSP	Maximum 12b-1 Fee in CRSP
mgmt_cd_CRSP	Management Company Number in CRSP
mgmt_fee_CRSP	Management Fee in CRSP
mgmt_fee_lag_CRSP	One Month Lagged Management Fee in CRSP
mgmt_name_CRSP	Management Company Name in CRSP
mgr_dt_CRSP	Date Current Portfolio Mgr Took Control in CRSP
mgr_name_CRSP	Manager Name in CRSP

Table 2: Final CRSP Dataset Contents, Continued

Variable	Variable Label
month	Month in CRSP
nav_CRSP	Monthly Net Asset Value per Share in CRSP
nav_lag_CRSP	One Month Lagged Net Asset Value per Share in CRSP
open_to_inv_CRSP	Open to Investors in CRSP
per_bond_CRSP	Amount of fund invested in all Bonds in CRSP
per_com_CRSP	Amount of fund invested in Common Stocks in CRSP
per_conv_CRSP	Amount of fund invested in Convertible Bonds in CRSP
per_corp_CRSP	Amount of fund invested in Corporate Bonds in CRSP
per_pref_CRSP	Amount of fund invested in Preferred Stocks in CRSP
policy_CRSP	Type of Securities Mainly Held by Fund in CRSP
portno_CRSP	Portfolio Identifier in CRSP
rear_load_flag_CRSP	Flag for a Fund with Positive Rear Load in CRSP
retail_fund_CRSP	Retail Fund Indicator in CRSP
return_CRSP	Monthly Return in CRSP
si_obj_cd_CRSP	Strategic Insight Objective Code in CRSP
sum_me_CRSP	Monthly Aggregate US Stock Market Capitalization in CRSP
sum_me_lag_CRSP	One Month Lagged Aggregate US Stock Market Capitalization in CRSP
summary_period_CRSP	Frequency of Fund Summary Reporting in CRSP
ticker_CRSP	Ticker in CRSP
turn_ratio_CRSP	Fund Turnover Ratio in CRSP
turn_ratio_lag_CRSP	One Month Lagged Fund Turnover Ratio in CRSP
wbrger_obj_cd_CRSP	Wiesenberger Fund Type Code, Identifying Fund Strategy in CRSP
year	Year in CRSP

2. Raw Morningstar Database Clean-up and Merge

First, we assign the following six category dummies to each fund in the dataset *fund_ops*, which contains summary information for 42,575 funds. A fund can be assigned zero, one, or multiple flags. The six categories are:

- *Bond_fund_MS (bond fund)*
- *Intl_fund_MS (international fund)*
- *Sector_fund_MS (sector fund)*
- *Target_fund_MS (target fund)*
- *RealEstate_fund_MS (real estate fund)*
- *Other_nonequity_funds_MS (other non-equity fund)*

We set a flag to one if the fund's *Morningstar_category* variable's value belongs to the corresponding list below:

Bond funds	Int'l funds	Sector funds	Target funds	Real estate funds	Other non-equity
Bank Loan	China Region	Commodities Broad Basket	Target-Date 2000-2010	Global Real Estate	Currency
Convertibles	Diversified Emerging Mkts	Communications	Target-Date 2011-2015	Real Estate	Long/Short Equity
Emerging Markets Bond	Diversified Pacific/Asia	Consumer Cyclical	Target-Date 2016-2020		Managed Futures
High Yield Bond	Emerging Markets Bond	Consumer Defensive	Target-Date 2021-2025		Market Neutral
High Yield Muni	Europe Stock	Equity Energy	Target-Date 2026-2030		Multialternative
Inflation-Protected Bond	Foreign Large Blend	Equity Precious Metals	Target-Date 2031-2035		Trading-Inverse Commodities
Intermediate Government	Foreign Large Growth	Financial	Target-Date 2036-2040		Trading-Inverse Debt
Intermediate-Term Bond	Foreign Large Value	Health	Target-Date 2041-2045		Trading-Miscellaneous
Long Government	Foreign Small/Mid Blend	Industrials	Target-Date 2046-2050		
Long-Term Bond	Foreign Small/Mid Growth	Miscellaneous Sector	Target-Date 2051+		
Multisector Bond	Foreign Small/Mid Value	Natural Resources			
Muni California Intermediate	Global Real Estate	Technology			
Muni California Long	India Equity	Utilities			
Muni Massachusetts	Japan Stock				
Muni Minnesota	Latin America Stock				
Muni National Interm	Pacific/Asia ex-Japan Stk				
Muni National Long	World Allocation				
Muni National Short	World Bond				
Muni New Jersey	World Stock				
Muni New York Intermediate					
Muni New York Long					
Muni Ohio					
Muni Pennsylvania					
Muni Single State Interm					
Muni Single State Long					
Muni Single State Short					
Nontraditional Bond					
Short Government					
Short-Term Bond					
Ultrashort Bond					
World Bond					

We also make use of the variable *primary_prospectus_benchmark* in order to further assign the fund style dummy. In particular, we check if *primary_prospectus_benchmark* contains any of the words listed in the first column of the following table. If so, we assign the correspondingly style dummy (second column) to the fund.

Keywords in primary_prospectus_benchmark	*Fund Style Dummy*
bond / treasury / govt / barclay / municipal / convertible / investment-grade / consumer price index / t-bill / dollor[2]	Bond_fund_MS
ACWI / world / global / emerging markets / latin america / eafe / msci em	Intl_fund_MS
target	Target_fund_MS
property / reit	Realestate_fund_MS
commodity	Sector_fund_MS

Finally, we create a variable *num_cat* as the sum of all the six fund style dummies. A zero value of *num_cat* means a fund does not belong to any of the six fund styles.

Next, we merge the raw Morningstar data files to create a final dataset that contains all the relevant variables for our study. Table 3 lays out the structure of each raw dataset. We start the merging process by merging *fund_ops* with *returns*. In each step of the merge, we merge in only observations of the year/month when return data exists. If a variable is missing, we keep the year/month observation and record a missing value for that variable.

In file *portast*/*expenses*/*histturn*, 98.75%/100.00%[3]/98.80% of the observations are uniquely identified by the fund identifier and time period variables. The remaining few are duplicates. We take the average of the values in those cases.

We eventually use expense ratio data from CRSP rather than Morningstar. The reason is that Morningstar is ambiguous about the timing of their expense ratio data.[4] We nevertheless describe our method for processing the Morningstar expense ratio data. The *expenses* dataset in Morningstar offers only a year variable, but funds may have different fiscal year ends. In order to incorporate the fiscal year end into our merge, we use the *fiscal* variable in the *fund_ops* dataset. The *fiscal* variable specifies the last month of a fund's fiscal cycle. We only have one *fiscal* observation per fund (presumably the most recent one), and we suspect that some funds changed their fiscal calendar. We make comprehensive comparisons of expense ratio data between CRSP and Morningstar. We conclude that the correct way to assign the effective fiscal cycle for the Morningstar expenses data is a forward cycle from the *year/fiscal* time. For example, if *year*=2002, *fiscal*=9, *exp_ratio*=0.25, then the effective fiscal cycle for an expense ratio of 0.25 is October/2002 ~ September/2003[5].

---

[2] "dollor" is a typo in the raw data.

[3] Rounding to two decimal places give 100.00% but 24 cases are duplicates.

[4] Additional details are in Section 5.

[5] This forward assignment of fiscal year cycle yields over 83% of matched expense ratio between CRSP and Morningstar. A backward assignment (in the example, it would be assigned October/2001 ~ September/2002) only yields 20% of matched expense ratio between CRSP and Morningstar.

In Morningstar's monthly asset dataset, we discover that there are instances of extreme reversal patterns that likely reflect decimal-place mistakes. We perform the following procedure to remove fix these extreme reversals. First, we create a variable for the fraction change from last month to the current month,

dassets=(assets-lag_assets)/lag_assets.

Second, we create a reversal variable to capture the reversal pattern,

rev_next=(lead_assets-assets)/(assets-lag_assets).

This variable will be approximately –1 if it is a reversal (e.g. 20m, 2m, 20m).
Finally, if *abs(dassets)>=0.5, –0.75>rev_next>–1.25, and lag_assets>=10m*, then we assign missing value to both *assets* and *dassets*. As a result of this procedure, 0.026% of monthly asset observations are changed to missing.

Next, we create additional variables, including one-month-lag variables of
- *asset (total net asset),*
- *nav (net asset value per share),*
- *asset_fundid (combined total net asset of sub share classes of a fund),*
- *asset (total net asset),*
- *expense_ratio (expense ratio),*
- *turnover (turnover ratio),*

and

- *inception (inception date of a fund),*
- *fund_age (fund age since inception, in days),*
- *first_ms_dt (date the fund first appeared in Morningstar).*

Benchmark return:

The Morningstar file *cat_and_index_mapping* maps each Morningstar Category into a benchmark index. For example, the most common Morningstar Category is "Large Growth," which maps into the benchmark index "Russell 1000 Growth TR USD." We use this mapping to assign a benchmark index to each fund. The Morningstar file *idx_mth_ret* provides monthly returns on all Morningstar benchmark indexes.

Finally, we rename the variables so that each variable from Morningstar dataset has "*_MS*" in its name. Table 4 shows the content of our final Morningstar dataset. The final dataset contains 5,112,629 observations, covers 42,575 fund share classes (reflected by *secid*), 14,765 funds (reflected by *fundid*), and spans the time period from August/1924 to August/2012.

Table 3: Morningstar Raw File Structure

Raw Morningstar File	**Variables**	**Variable Label**	**Dataset Duration**	**# of Obs**	**# of Funds *(secid)***	**# of Funds *(fundid)***
fund_ops	Actual_Management_Fee	Actual Management Fee in MS	N/A	42,575	42,575	14,765
	Cusip	CUSIP in MS				
	Deferred_Load	Deferred Load in MS				
	End_Date	End Date of Fund in MS				
	Enhanced_Index	Enhanced Index (F or T) in MS				
	Family	Fund Family (e.g. 13D Management/AQR Funds/Citizens Funds) in MS				
	Fiscal	Fiscal Year End (1,2,…,12) in MS				
	Front_Load	Front Load in MS				
	Fund_of_Fund	Fund of Fund (F or T) in MS				
	Fundid	Fund Identifier (May Including Multiple Share Classes) in MS				
	Fundname	Fund Name in MS				
	Inception	Fund Inception Date in MS				
	Index	Index (F or T) in MS				
	Primary_Prospectus_Benchmark	Primary Prospectus Benchmark in MS				
	Prospectus_Objective	Prospectus Objective in MS				
	Reason	Reason (L or M) in MS				
	Redemption_Fee	Redemption Fee in MS				
	SRI	SRI (F or T) in MS				
	Secid	Security Identifier (Unique per Share Class Level) in MS				
	Share_Type	Share Type/Share Class in MS				
	Status	Fund Status (A or O) in MS				
	Ticker	Ticker in MS				
	morningstar_category	Fund Category (94 total) in MS				
	v12B_1	12b_1 Free in MS				

Table 3: Morningstar Raw File Structure, Continued

Raw Morningstar File	**Variables**	**Variable Label**	**Dataset Duration**	**# of Obs**	**# of Funds *(secid)***	**# of Funds *(fundid)***
returns	date	Date in MS	08/31/1924 ~ 08/31/2012	5,986,578	47,047	N/A
	month_yyyymm_	Month in yyyymm Format in MS				
	return	Monthly Fund Return in MS				
assets	assets	Monthly Fund Assets in MS	03/31/1953 ~ 10/03/2012	3,611,056	42,209	N/A
cat_and_index_mapping	Index_name	Index Name in MS	N/A	92	N/A	N/A
	Morningstar_Category	Fund Category (94 total) in MS				
expenses	expense_ratio	Annual Expense Ratio in MS	1972 ~ 2012	300,598	38,626	N/A
	year	Year in MS				
portast	assets	Monthly Fundid (Multiple Share Classes) Level Assets in MS	03/31/1953 ~ 09/30/2012	1,470,656	N/A	14,398
	share_classes	Number of Share Classes in MS				
histturn	turnover	Annual Turnover Ratio in MS	1908 ~ 2012	117,097	N/A	13,120
	year	Year in MS				
div	dividend	Monthly Fund Dividend in MS	10/1924 ~ 10/2010	1,483,196	33,930	N/A
	month	Year/Month in MS				
nav	nav	Monthly Net Asset Value per Share in MS	07/31/1924 ~ 09/30/2012	4,752,288	45,766	N/A
	date	Date in MS				
ratings	overall_rating	Fund Rating in MS	12/31/1985 ~ 09/30/2012	2,198,895	31,587	N/A
	date	Date in MS				

Table 4: Final Morningstar Dataset Contents

Variable	Variable Label
Actual_Management_Fee_MS	Actual Management Fee in MS
Bond_fund_MS	Bond Fund Dummy in MS
Category_MS	Fund Category (94 total) in MS
Cusip_MS	CUSIP in MS
Deferred_Load_MS	Deferred Load in MS
Enhanced_Index_MS	Enhanced Index (F or T) in MS
Family_MS	Family (Many Values, e.g. 13D Management/AQR Funds/Citizens Funds) in MS
Fiscal_MS	Fiscal Year End (Values 0,1,2,…,12, very few 0) in MS
Front_Load_MS	Front Load in MS
Fund_of_Fund_MS	Fund of Fund (F or T) in MS
Fundid	Fund Identifier (May Including Multiple Share Classes) in MS
Fundname_MS	Fund Name in MS
Inception_MS	Fund Inception Date in MS
Index_MS	Index (F or T) in MS
Index_name_MS	Index Name in MS
Intl_fund_MS	International Fund Dummy in MS
Mgr_End_Date_MS	End Date of Manager in MS
PPB_MS	Primary Prospectus Benchmark in MS
Prospectus_Obj_MS	Prospectus Objective in MS
RealEstate_fund_MS	Real Estate Fund Dummy in MS
Reason_MS	Reason (L or M) in MS
Redemption_Fee_MS	Redemption Fee in MS
SRI_MS	SRI (F or T) in MS
Secid	Security Identifier (Unique per Share Class Level) in MS
Sector_fund_MS	Sector Fund Dummy in MS
Share_Type_MS	Share Type/Class in MS
Status_MS	Fund Status (A or O) in MS
Strange_funds_MS	Strange Fund Dummy in MS
Target_fund_MS	Target Fund Dummy in MS
Ticker_MS	Ticker in MS

Table 4: Final Morningstar Dataset Contents, Continued

Variable	Variable Label
assets_MS	Monthly Fund Assets in MS
assets_fundid_MSavg	Monthly Fundid (Multiple Share Classes) Level Assets in MS
assets_fundid_MSavg_lag	1-Month Lagged Monthly Fundid (Multiple Share Classes) Level Assets in MS
assets_lag_MS	Lagged Monthly Fund Assets in MS
date	Date in MS
deferred_load_flag_MS	Flag for a Fund with Positive Deferred Load in MS (snapshot from fund_ops)
dividend_MS	Monthly Fund Dividend in MS
expense_ratio_MSavg	Annual Expense Ratio in MS
expense_ratio_MSavg_lag	1-Month Lagged Annual Expense Ratio in MS
first_MS_dt	First Date a Fund Appears in MS in MS
front_load_flag_MS	Flag for a Fund with Positive Front Load in MS (snapshot from fund_ops)
fund_age_MS	Fund Age (in days) in MS
month	Month in MS
nav_MS	Monthly Fund Net Assets in MS
nav_lag_MS	Lagged Monthly Fund Net Assets in MS
num_cat_MS	Sum of the Six Fund Style Dummy in MS
ratings_MS	Monthly Fund Ratings in MS
return_MS	Monthly Fund Return in MS
return_idx_MS	Return of Index in MS
secid_idx_MS	Secid of Index in MS
share_classes_fundid_MS	Number of Share Classes in MS
turnover_MSavg	Annual Turnover Ratio in MS
turnover_MSavg_lag	1-Month Lagged Annual Turnover Ratio in MS
v12B_1_MS	12b_1 Free in MS
year	Year in MS

3. Creating Completely Matched Sample Between CRSP and Morningstar

(A) Useful Information Concerning *Ticker*- and *CUSIP*-based Merge

(A1) Morningstar (MS).

In MS, identifier *secid* uniquely identifies a fund share class. Different share classes of the same fund have different *secid's*, as well as different *CUSIP's* and *Tickers*. Meanwhile, they all share the same *fundid*. The mapping between *secid* and *CUSIP* is one-to-one. The mapping between *secid* and *Ticker* is also one-to-one. *CUSIP* and *Ticker* may be missing; *secid* is never missing.

In particular, out of the 42,575 fund share classes in MS, 40,682 (95.6%) have non-missing *CUSIP*; only 29,244 (68.7%) have non-missing *Ticker*.

(A2) CRSP.

In CRSP, *crsp_fundno* uniquely identifies a fund share class. A *crsp_fundno* is associated with only one *Ticker*, after cleaning *Ticker* variable in CRSP data following BB documentation pages 2-3 (details above). However, in CRSP data, a *crsp_fundno* may be associated with multiple *CUSIP's* throughout a fund's life.

39,156 (77.5%) out of 50,536 fund share classes in CRSP data have at least one non-missing *Ticker*.
45,650 (90.3%) out of 50,536 funds share classes in CRSP data have at least one non-missing *CUSIP*. Out of the 45,650 fund share classes with at least one non-missing *CUSIP*, here is the distribution of number of unique *CUSIP's* per *crsp_fundno*:

1 CUSIP per crsp_fundno ~ 36,526 (80.01%)
2 CUSIP's per crsp_fundno ~ 7,248 (15.88%)
3 CUSIP's per crsp_fundno ~ 1,619 (3.55%)
4 CUSIP's per crsp_fundno ~ 239 (0.52%)
5 CUSIP's per crsp_fundno ~ 18 (0.04%)

(A3) Takeaway.

The goal of our merge is to create a one-to-one concordance between 42,575 Morningstar *secid's* and 50,536 *crsp_fundno's*. BB use *Ticker* but not *CUSIP* in their merge, suggesting they do not have *CUSIP* in their Morningstar data. We use both *Ticker* and *CUSIP* in our merge. Since *CUSIP* is more widely available than *Ticker* in both datasets, using *CUSIP* helps us make more accurate matches between databases.

(B) Algorithm for Merging CRSP & Morningstar Data

Step 1:
Merge CRSP and MS by *Ticker* at the share-class level. Both CRSP and MS datasets should have a unique *Ticker* per share class.

Step 2:
Check matching quality by comparing CRSP's and Morningstar's data on each fund's returns and total net assets. We say that a fund is "well matched" if and only if:
(1) the 60th percentile of the absolute value of the difference between CRSP and MS monthly return is less than 5 basis points, and
(2) the 60th percentile of the absolute difference between CRSP and MS monthly total net assets is less than $100,000.

If a fund is well matched, we accept its *secid—crsp_fundno* mapping as valid.
Otherwise, we proceed to the Step 3.

We find that 24,288 share classes[6] (48.1% of 50,536 CRSP fund share classes, 57.0% of 42,575 MS fund share classes) are well matched by *Ticker*.

Step 3:
Merge CRSP and MS using the CUSIP (*latest per crsp_fundno*) at the share-class level. Perform Step 2 to validate matches.

If well matched, accept the *secid—crsp_fundno* mapping.
Otherwise proceed to Step 4.

We find that 9,798 additional share classes (19.4% of 50,536 CRSP share classes, 23.0% of 42,575 MS share classes) are well matched by the CUSIP (*latest per crsp_fundno*).

Step 4:
Merge CRSP and MS using the CUSIP (*second latest*) per *crsp_fundno* at the share-class level. Perform Step 2 to validate matches.

If well matched, accept the *secid—crsp_fundno* mapping.
Otherwise, keep the remaining unmatched fund share classes till Step 6.

Only 56 additional share classes are well matched by the earlier *CUSIP*.

Step 5:

[6] BB apply thresholds of 2bp and $20,000. If we use those same thresholds, we find only 1,086 valid matches. We relax the return and asset matching threshold to 5bp/$100,000 to obtain more matches. Although looser than the BB thresholds, our thresholds are still quite demanding.

We say that a fund (identified by *fundid* in Morningstar) is "completely matched" if all the MS *secid*'s belonging to the *fundid* are well matched to *crsp_fundno*'s. We say a fund is "partially matched" if at least one share class is well matched and at least one share class (in either CRSP or MS) is not.

At this step we find that 9,711 funds (65.8% of 14,765 MS *fundid*'s) are completely matched. 1,572 funds (10.6%) are partially matched. 3,482 funds (23.6%) are not matched at all.

Step 6:
The goal of this step is to identify additional matches by comparing funds' names between CRSP and Morningstar. Only 212 additional share classes are well matched by this text-based approach.

The steps are as follows:

a. Using the full CRSP database, create variables *crsp_group_id* and *crsp_subclass* from *fund_name_CRSP* , following pages 19-20 of BB's documentation.[7] The variable *crsp_group_id* assigns a shared identifier to all share classes belonging to the same fund. The variable *crsp_subclass* assigns a share type to different share classes.
b. For each *crsp_group_id*, check whether at least one share class in the group is well matched to a MS share class. If so, assign that share class's *fundid* to all members of the *crsp_group_id*. For example, if fund share class A (well matched) and fund share class B (not yet matched) share the same *crsp_group_id*, then B will be assigned A's *fundid*.
c. Remove the following funds, which would either only yield partially matched cases or be useless for text-based matching:
   (c1) A *crsp_group_id* is associated with more funds in CRSP than its associated *fundid* indicates in MS. For example, if a *crsp_group_id* is associated with 3 *crsp_fundno*'s while the *fundid* is only associated with 2 *secid*'s. We remove this fund because we presume Morningstar is missing data on at least one share class.
   (c2) A *crsp_group_id* is associated with insufficient number of non-matched fund share classes. For example, a *crsp_group_id* is associated with 4 *crsp_fundno*'s – 3 well matched and 1 not yet matched. The associated MS *fundid* is related to 5 fund share classes. In this case, even if we match the one fund through text-based approach, we would still end up with only a partially matched case. We presume in this case that CRSP is missing data on at least one share class.
   (c3) A *crsp_group_id* is associated with only well-matched fund share classes. They offer no incremental benefits for our text-based matching.

[7] The relevant excerpt from their documentation is at the end of this document. We deviate slightly from BB when we break the fund name by "/" to assign *crsp_group_id* and *crsp_subclass*. We notice some fund names such as "Vanguard equity/bond balance fund" should not be assigned a *crsp_subclass* as "bond balance fund", but rather the whole fund name should be assigned *crsp_group_id*. After compressing the typical words (e.g. 'class', 'share') in share class, we decided to not break the fund name by "/" if what follows contains more than 11 letters. 11 is the 99th percentile for number of letters that come after "/".

(c4) A *crsp_group_id* is associated with only non-matched fund share classes. Similarly, they also offer no incremental benefits for our text-based matching.

d. Within the remaining sample of funds, continuing from example in (b), if fund share class B's *crsp_subclass* overlaps with the *share_type* of a MS fund with the same *fundid*, then we consider this pair of funds a potential match candidate. We compress *crsp_subclass* by deleting the non-essential words such as 'class' and 'share'[8]. Table 5 includes more details on processing the *crsp_subclass* in this remaining fund sample. We process *crsp_subclass* so that it takes one out of the 15 possible MS *share_type* values.
e. Perform Step 2 on these potential match candidate funds.

Step 7:
Remove from the sample any funds that are partially matched or unmatched.

The resulting dataset is a concordance file for the fund identifiers *crsp_fundno, secid, and fundid*. It contains only complete matches with 8,807 funds and 27,414 share classes. Note this is less than the 9,711 funds in step 5 because we deleted the partially matched cases described in Step 6(c1).

[8] For example, "Class A Share" is compressed to "A". The whole list of suppressed words are 'cl', 'class', 'share', 'shares', 'shar', 'shs', 'of', '-', 'series', 'trust'.

Table 5: Share Type Processing

crsp_subclass with keywords	crsp_subclass compressed	Processed to MS share_type
Ashares / (A)	1	A
BShares / (B) / B(1)	2	B
CShares / (C) / C-	3	C
DShares / (D)	4	D
MShares / (M)		M
NShares / (N)		N
SShares / (S)		S
TShares / (T)		T
Investor / Invest / Consumer /		Inv
Institutional / Instl / IB / Initial / I2 / IA / II	I / Y	Inst
Retire / R1 / R2 / R3 / R4 / R5 / R 5 / R6 / RetA / RetB	R / K	Retireme
Advisor / Adviser / Admin / Administrative / Admiral / Consultant / Service / Retail		Adv
No Load / NoLoad		NoLoad
Nav / BlackRock	E / F / H / L / W / Z	Other
All else	All else	NA

4. Merging CRSP and Morningstar Databases

Using the concordance file created in Section 3, we then perform the following three steps to merge CRSP and Morningstar data:

(1) Create a CRSP dataset that only contains completely matched funds. The resulting smaller CRSP dataset contains 2,936,897 observations compared to the 4,784,162 observations in the larger CRSP dataset created in Section 1.

(2) Create a Morningstar dataset that only contains completely matched funds. The resulting smaller Morningstar dataset contains 3,722,775 observations compared to the 5,112,629 observations in the larger Morningstar dataset created in Section 1.

(3) Merge the CRSP dataset and Morningstar dataset created in (1) and (2) by share class/year/month. We use the CRSP dataset as the master file for the merge. The resulting dataset contains 2,936,897 share class / month observations for 8,807 funds and 27,414 share classes.

5. Correcting Expense Ratios and Returns

1. Expense Ratio and Management Fee Fix.

At the share class/month level, we set the expense ratio/management fee to missing if its value is negative.

We use expense ratio data from CRSP only, because CRSP provides precise data on the timing of expense ratios, whereas timing conventions in the MS data are more ambiguous. Specifically, CRSP provides an exact start and end date for each expense ratio observation, whereas MS reports fiscal-year expense ratios along with a fiscal-year-end month. Morningstar only reports each fund's most recent fiscal-year-end month, so the timing of Morningstar data are inaccurate if a fund changed its fiscal calendar. After trying several methods of mapping MS's fiscal-year data into calendar time, we still find many (~17%) discrepancies between CRSP and MS. In the instances where CRSP and Morningstar disagree, it is difficult to find the relevant SEC filing to help determine which source is correct. Because CRSP provides exact calendar-time data on fees, we rely only on CRSP data.

Among the 2,936,897 share class/month observations, we have 2,426,178 (82.6%) observations that have non-missing expense ratios.

2. Return Fix.

Like Berk and Binsbergen (2012), we find many observations where the monthly returns reported by CRSP and Morningstar are "inconsistent," meaning they differ by more than 10 basis points. Among our 2,936,897 observations, 2,861,953 observations have both returns non-missing. Among these *comparable* observations, 88,677 (3.1%) observations have inconsistent returns that need fixing. Details on the differences can be found in Table 6 below:

Table 6: CRSP and Morningstar Return Difference

Difference	# of Observations	% of Observations
Do Not Differ	2,534,092	88.5%
1 basis point	82,507	2.9%
2~10 basis point	156,677	5.5%
11~100 basis point	74,676	2.6%
>100 basis point	14,001	0.5%

Pages 15-18 in BB's documentation describe an alternative measure for fund returns from net asset values and dividends. This method requires net asset values and dividends data. Among the 88,677 inconsistent returns, 56,060 have the required data on net asset values and dividends. We then follow pages 15-18 in BB's documentation, which we provide at the end of this

document, to fix these 56,060 inconsistent returns. We implement BB's Steps One and Two[9] but skip Step Three (Bloomberg related procedure). That is, we focus on the procedure within the scope of CRSP and Morningstar datasets. These steps reduce the number of inconsistent returns to 18,022, or 0.6% of total observations in the database.

We set the remaining, unresolved inconsistent returns to missing. We also set returns to missing for a share class/month observation if its value is missing in either CRSP or Morningstar. In other words, we only work with observations that have non-missing return data in both CRSP and Morningstar.

Finally, among the 2,936,897 share class/month observations, we have 2,799,599 (95.3%) observations that have non-missing returns.

3. Asset Fix.

We set assets under management (AUM) to missing for a share class/month observation if either CRSP's or Morningstar's value is missing. In other words, we only work with observations that have non-missing assets data in both CRSP and MS.

We then set assets to missing if CRSP and MS disagree by at least $100,000 *and* the relative disagreement is at least 5%. Otherwise, we set assets to CRSP's value.

Finally, among the 2,936,897 share class/month observations, we have 2,529,011 (86.1%) observations that have non-missing assets.

[9] In particular, we implement Step Two going back and forward by 12 months, i.e. we performed the procedure up to (t-12) and (t+12).

6. Identifying Index Funds

In order to create a dummy variable to indicate index funds, we use a simple two-step procedure:

(1) If either CRSP or MS says it is an index fund, then we call it an index fund (i.e. *index_both*=1). Otherwise, we move to step 2. CRSP's index fund variable is *index_fund_flag.* Morningstar's variables (in the *fund_ops* table) are *Index* and *Enhanced Index.*

(2) If the fund name in either CRSP or MS contains "index," we call it an index fund (i.e. *index_both*=1).

Otherwise, we call it active (i.e. *index_both*=0).

As a result of this procedure, we are able to identify 732 (8.3%) out of 8,807 funds that are index funds.

7. Grouping Subclasses

We create a dataset that is aggregated to the *fundid* level. We start our merge from a concise dataset that contains only three variables: *fundid*, *year*, and *month*. We then merge in the following lists of *fundid*-level variables:

Variable List 1
These variables take the same value across share classes of the same *fundid-month*.

index_name_MS / return_idx_MS / category_MS / sum_me_CRSP (and lag).

Variable List 2
These variables need to be summed across share classes of the same *fundid-month*.

assets_CRSP (and lag) / assets_MS (and lag) / assets_both (and lag)

Variable List 3
These variables need to be averaged (lag-asset-weighted) across share classes of the same *fundid-month*.

return_CRSP / return_MS / return_both / exp_ratio_CRSP (and lag) / expense_ratio_MSavg (and lag) / exp_ratio_both (and lag) / turn_ratio_CRSP/ turnover_MSavg

Variable List 4
We need to take the maximum of these variables across share classes of the same *fundid-month*.

enhanced_index_MS / SRI_MS / fund_of_fund_MS / index_CRSP / index_MS / index_both / realestate_fund_MS / sector_fund_MS / bond_fund_MS / strange_funds_MS / target_fund_MS / intl_fund_MS / front_load_flag_CRSP / rear_load_flag_CRSP / front_load_flag_MS / deferred_load_flag_MS

Variable List 5
We need to take the minimum of these variables across share classes of the same *fundid-month*.

first_offer_dt_CRSP / inception_MS / first_CRSP_dt / first_MS_dt

When aggregating *assets* across share classes, we set assets to missing at the fund level if any share class has missing assets data in a given month.

When aggregating *returns*, *turnover*, and *expense ratios*, and *management fees* across share classes, we take the average across all *non-missing* share classes' values. In other words, we don't set their values to missing at the fund level just because one or more share class has missing data.

Table 7 shows the content of our fund-level final dataset.

Table 7: Final Fundid Level Dataset Contents

Variable	Variable Label
Category_MS	Fund Category (94 total) in MS
Fundid	Fund Identifier (May Including Multiple Share Classes) in MS
Index_name_MS	Index Name in MS
fundid_SRI_ms	SRI Fund dummy in MS
fundid_assets_CRSP	Total Net Assets as of Month End in CRSP
fundid_assets_MS	Monthly Fund Assets in MS
fundid_assets_both	Fund Total Net Assets Validated by CRSP and MS
fundid_assets_both_lag	1-Month-Lagged Fund Total Net Assets Validated by CRSP and MS
fundid_assets_lag_CRSP	One Month Lagged Fund Assets in CRSP
fundid_assets_lag_MS	Lagged Monthly Fund Assets in MS
fundid_bond_fund_ms	Bond Fund Dummy in MS
fundid_deferred_load_flag_MS	Flag for any Sub-Share-Class with Positive Deferred Load in MS
fundid_enhanced_index_ms	Enhanced Index Fund dummy in MS
fundid_exp_ratio_CRSP	Fund Expense Ratio (Asset Weighted) in CRSP
fundid_exp_ratio_MS	One Month Lagged Fund Expense Ratio (Asset Weighted) in MS
fundid_exp_ratio_both	Corrected Fund Expense Ratio (CRSP Asset Weighted)
fundid_exp_ratio_both_lag	One Month Lagged Corrected Fund Expense Ratio (CRSP Asset Weighted)
fundid_exp_ratio_lag_CRSP	One Month Lagged Fund Expense Ratio (Asset Weighted) in CRSP
fundid_exp_ratio_lag_MS	One Month Lagged Fund Expense Ratio (Asset Weighted) in MS
fundid_first_CRSP_dt	First Date a Fund Shows Up in CRSP
fundid_first_ms_dt	First Date a Fund Appears in MS in MS
fundid_first_offer_dt_crsp	Date the Fund Was First Offered in CRSP
fundid_front_load_flag_CRSP	Flag for any Sub-Share-Class with Positive Front Load in CRSP
fundid_front_load_flag_MS	Flag for any Sub-Share-Class with Positive Front Load in MS
fundid_fund_age_CRSP	Age of a Fund at Month End in CRSP
fundid_fund_age_MS	Fund Age (in days) in MS
fundid_fund_of_fund_ms	Fund of Fund dummy in MS
fundid_inception_ms	Fund Inception Date in MS
fundid_index_CRSP	Index Fund dummy in CRSP
fundid_index_both	Index Fund dummy Constructed from both CRSP and MS
fundid_index_ms	Index Fund dummy in MS

Table 7: Final Fundid Level Dataset Contents, Continued

Variable	Variable Label
fundid_intl_fund_ms	International Fund Dummy in MS
fundid_mgmt_fee_CRSP	Fund Management Fees (Asset Weighted) in CRSP
fundid_mgmt_fee_CRSP_lag	1-Month-Lagged Fund Management Fees (Asset Weighted) in CRSP
fundid_realestate_fund_ms	Real Estate Fund Dummy in MS
fundid_rear_load_flag_CRSP	Flag for any Sub-Share-Class with Positive Rear Load in CRSP
fundid_sector_fund_ms	Sector Fund Dummy in MS
fundid_strange_funds_ms	Strange Fund Dummy in MS
fundid_target_fund_ms	Target Fund Dummy in MS
month	Month in CRSP
return_CRSP	Monthly Return (Asset Weighted) in CRSP
return_MS	Monthly Return (Asset Weighted) in MS
return_both	Monthly Return (Asset Weighted) validated in both CRSP and MS
return_idx_MS	Return of Index in MS
share_class_assets_num	Number of Share Classes in a given Year/Month that has positive validated assets
share_class_exp_num	Number of Share Classes in a given Year/Month that has non-missing validated expense ratio
share_class_mgmt_num	Number of Share Classes in a given Year/Month that has non-missing validated management fee
share_class_return_num	Number of Share Classes in a given Year/Month that has non-missing validated assets
share_classes_fundid_MS	Number of Share Classes in MS
sum_me_CRSP	Monthly Aggregate US Stock Market Capitalization in CRSP
sum_me_lag_CRSP	One Month Lagged Aggregate US Stock Market Capitalization in CRSP
year	Year in CRSP

8. Computing SectorSize

SectorSize (v.1) $_{it}$ equals the sum of fund size (in nominal dollars) across all funds j in month t that belong to fund i's sector, divided by the total market cap of CRSP stocks in fund i's sector in month t. (Our regressions used lagged SectorSize, i.e., SectorSize measured at the end of the previous month.) If the size of some fund j is missing, we use the fund's most recent size, and we update that size using interim realized fund returns and the assumption of zero flows; we do not impute fund size more than 12 months ahead. We use the 9 sectors defined by Morningstar's 3×3 StyleBox. We allocate CRSP stocks to the Morningstar 3×3 matrix using Ken French's 10×10 portfolios sorted by size and book-to-market. Note that the three Morningstar size categories (small, mid-cap, and large) are mutually exclusive, whereas the "blend" category overlaps with both growth and value. We use a 3-4-3 split for firm size and a 5-5 split for growth vs. value. Our measure takes into account the following facts:

1. Blend funds can invest in both growth and value stocks.
2. Blend funds compete with the entire portfolio of both value and growth funds.
3. Value and growth funds compete with roughly half of blend funds' portfolios.

For small-growth funds (for example), SectorSize(v.1) equals [size of small-growth funds + (1/2) size of small-blend funds] / [mkt. cap. of the 15 Fama-French portfolios in the bottom-3 size and bottom-5 B/M portfolios].

For small-blend funds, SectorSize (v.1) equals [size of all small-cap funds (blend+growth+value)]/[mkt. cap. of the 30 Fama-French portfolios in the bottom 3 size portfolios].

SectorSize (v.2) is the same as SectorSize (v.1) but only considers 3 sectors (small-, mid-, and large-cap funds). For example, SectorSize (v.2) for small-cap funds is the total TNA of all small-cap funds divided by the total market cap of all stocks in the bottom 3 Fama-French size deciles.

9. Excerpts from Berk and Binsbergen's (2011) Data Appendix

This section reproduces the parts of Berk and Binsbergen's (2011) Data Appendix that we reference in previous sections of this document. The excerpts below are reproduced verbatim from their appendix dated 06/09/2011. The page numbers below refer to page numbers in their appendix.

A. Pages 2-3:

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 they are 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 in the 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 observation. We judge that this repetition is a mistake made in the original CRSP database in trying to correct the error in fiscal year end. 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* uses more than one tickers. 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:

1. A *crsp_fundno* either corresponds to an empty *ticker*, or to a unique non-empty *ticker*, but never to both empty *ticker* and non-empty *ticker*, or a multiple of non-empty *tickers*.
2. 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* uses 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 suggest 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*, I 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, I 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 identifies an observation in this modified database.

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

Pages 15-18:

Correction of Monthly Returns

We realized that there are 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 has 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 "differing significantly" or "inconsistent" to refer to when the absolute difference in the monthly return reported by Morningstar and by CRSP differ by more 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 mechanism to these inconsistent monthly returns. Firstly, we recognize that both CRSP and Morningstar report net assets value (NAV) and sometimes also report dividend values. From these NAVs, we can compute additional two set 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:

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

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

We notice that it is often the case that the dividend value is missing. We apply the following set of rules to fill in the dividend value as best as we can:

1) If dividend is missing in one database (either CRSP or Morningstar), but not the other, then we are going to 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 to be 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 the 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

Then for a given observation with monthly return inconsistency, we apply the following set of rules (note "consistency" is defined in the beginning of this section):

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.

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

Step Two

Then we noticed that 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, lets 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

This is such a case where, 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 assets 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 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 this case, I note that crsp_ret and mret are consistent but ms_ret and totret1mo are inconsistent at 11/2002 (i.e. t+1). From this I deduct that mret is accurate in 10/2002.

What if multiple mistakes in recording NAV happen in consecutive months. 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 resolve 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 was able to correct and additional 17730 inconsistencies in returns.

B. Pages 19-20

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 1 or 2 are 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 all 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 condition 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 gets 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 .