Production of U.S. *SMB* and *HML* in the Fama-French Data Library

Eugene F. Fama
Booth School of Business, University of Chicago (Fama)

Kenneth R. French
Tuck School of Business, Dartmouth College (French)

First Draft: November 2023

Production of U.S. *SMB* and *HML* in the Fama-French Data Library

Eugene F. Fama and Kenneth R. French*

Abstract

This paper describes the construction of the size factor, *SMB* (small minus big), and the value minus growth factor, *HML* (high minus low book-to-market equity) in the Fama-French Data Library. Table 1 gives summary statistics that show how the average values of *SMB* and *HML* change because of five major CRSP data projects and the two changes in our rules that produce the factors.

After circulating an early draft of "Common Risk Factors in the Returns on Stocks and Bonds" (Fama and French, *Journal of Financial Economics* 1993), several colleagues asked for the factors of that paper. The bond factors are not ours to distribute, but we shared our stock factors (*Mkt*, *SMB*, and *HML*) by email. Publication of the paper led to many more requests, so we posted the stock factors on French's website. The resulting Data Library has grown over the last 30 years and includes data we produce for our research and series others have suggested. The Data Library includes results for markets outside the United States, but to simplify the discussion, we focus on U.S. data here. Thus, references to data should be interpreted to mean U.S. data.

The stock data we use to produce the Data Library – including prices, returns, and shares outstanding – are from the Center for Research in Security Prices (CRSP). The book equity data combine hand-collected values from the Moody's Industrial, Public Utility, Transportation, and Bank and Finance Manuals and data from Compustat. The book equity data from the Moody's Manuals are not affected by the changes we describe below. The book equity data from the

* Booth School of Business, University of Chicago (Fama) and Tuck School of Business, Dartmouth College (French). The comments of Savina Rizova are gratefully acknowledged. Fama and French are consultants to, board members of, and shareholders in Dimensional Fund Advisors.

Industrial Manuals, for example, are those used in Davis, Fama, and French (2000). All other accounting data are from Compustat.

CRSP and Compustat continually correct errors in their historical data. Like other researchers, we want the most accurate inputs for our research, so we use CRSP's most recent monthly release, with all available corrections, for our monthly data cuts. Because Compustat's data are relevant only for reforming our factor portfolios at the end of each June, we use their recent July release to produce our data cuts for July of year *t* to June of *t*+1. (To be clear, the July cut has data through July and is released in August. We use this timing convention throughout the paper and have switched to this convention on the Data Library website.)

CRSP Data Projects

CRSP has done major projects to improve its data over the last 30 years. Several are relevant for our Data Library.

1. In 2005, CRSP extended the start of its daily NYSE database from July 1962 to January 1926. Research for the project identified errors in month-end prices and dividend ex-dates in CRSP's monthly stock database. The corrections CRSP made as a result of this work affect our factor sorts and returns for 1926-1962.
2. CRSP added daily and monthly data for securities with primary listing on the Arca Exchange in July 2007. CRSP's coverage of NYSE Arca begins on March 8, 2006.
3. In December 2014, CRSP completed a review of shares outstanding data for 1925-1946. The review produced over 4000 changes to 400 PERMNOs. Most edits added shares outstanding that had been missing. The edits affect size and book-to-market equity sorts for 1925-1947.
4. CRSP corrected historical delisting codes and returns in 1998 and 1999.

5. Recent work on dividend codes with 1990-2022 ex-dates produced corrections that were made in the CRSP database from December 2021 to March 2022.

Changes in Factor Construction

The rules we use to compute the factors have changed little since 1993. We note the changes on the Data Library website when they occur. We provide more detail here.

We have changed the formula we use to compute book equity twice. The changes respond to Statements 106 and 109 issued by the Financial Accounting Standards Board (FASB).

FASB 106 – FASB issued Statement 106 in 1990. It requires U.S. firms to switch from their then current pay-as-you-go accounting approach for postretirement benefits other than pensions to an accrual approach that recognizes the expected cost of providing future benefits in the years the employee earns them. After Statement 106 was issued, firms could switch to the accrual approach immediately by recognizing the transition obligation on their next income statement and balance sheet or they could recognize the transition obligation over plan participants' future service periods, with disclosure of the unrecognized amount. We expected the transition obligation to have a substantial impact on the book equity of some firms that switched to the accrual approach immediately. To keep all firms on the same footing, we adjusted the book equity of switchers by subtracting the transition obligation they reported under FASB 106. In September 2020, however, we examined the historical impact of FASB 106 and found that our adjustment had little impact on the cross-section of book-to-market equity. Thus, starting with the August 2020 data cut, we simplify our process by eliminating the adjustment for all firms across all years. In other words, after the July 2020 data cut, we do not adjust any firm's reported book equity in any year for its promised non-pension postretirement benefits, regardless of how it accounted for the obligation during the FASB 106 transition. In effect, we undid our earlier response to FASB 106.

FASB 109 – When deciding in the early 1990s how to compute book equity, we consulted colleagues in the University of Chicago's accounting group and concluded that the present value of the payments implied by the Deferred Taxes and Investment Tax Credits on the typical firm's balance sheet was closer to zero than to the reported amount. As a result, our definition of book equity in FF (*Journal of Finance* 1992), FF (*JFE* 1993), and other papers that follow adds balance sheet Deferred Taxes and Investment Tax Credits to the book value of stockholders' equity (and subtracts the book value of preferred stock). In August 2016, a colleague pointed out that FASB 109, which was issued in 1993, improves the accounting for deferred income taxes. As a result, in files produced after July 2016 we do not add Deferred Taxes and Investment Tax Credit to BE for fiscal years ending in 1993 or later.

CRSP/Compustat Links – The links CRSP provided in the early 1990s between companies in its database and companies on Compustat were far from perfect. While doing our research on U.S. stock returns, we developed a framework and process that improved CRSP's links. The appendix describes the process used to construct our links.

We update the Data Library monthly, but for most series, including most factors, we form portfolios annually using data up to the end of June each year. Thus, from 1992 to 2021 we updated our CRSP-Compustat links after receiving CRSP's July release. CRSP improved its CRSP-Compustat links over the years. After comparing our proprietary links with CRSP's links in September 2021, we concluded that we could no longer justify the time required to update our links each year and, starting with the July 2021 data cut, we switched to CRSP's links.

To date, we have noted all changes in our processes on the Data Library webpage when they were made, including the three that affect *HML* and *SMB*. Before the July 2021 data cut, differences between our CRSP-Compustat links and CRSP's links created a problem for those

trying to replicate the returns in our Data Library. The distinction between CRSP's two permanent identifiers still causes some replicators to stumble. CRSP's Permnos identify share classes. Permcos, which are combinations of one or more Permnos, identify companies. Compustat's accounting data are for companies, so we use Permcos to form portfolios and compute their returns, book to market ratios, and other properties.

Production

We wrote the computer programs that produce the returns and other information in our Library, and for many years we ran the monthly updates. Dimensional Fund Advisors' research group began helping with the updates in 2003. Savina Rizova, who worked for French as an undergraduate research assistant for two years at Dartmouth and was Fama's Ph.D. student at the University of Chicago, is head of Dimensional's research group. We continue to determine the rules, definitions, and process used to form factor portfolios. Under our guidance, Dimensional employees produce the monthly updates, post them on a Dartmouth server, maintain the computer code, and until 2021 updated our CRSP-Compustat links.

The portfolios for most series in the Data Library are reformed at the end of each June and the first monthly returns for each year's new portfolios are for July. Data from CRSP's August release are critical inputs when we construct the new crop of portfolios, so production of our July data cut is more involved than those for other months. Before 2021, for example, we updated our CRSP/Compustat links with CRSP's August release. We also monitor the production process before releasing the July data cut by comparing the current data with the data produced 12 months earlier. In addition to checking summary statistics, when appropriate we review a few of the largest changes in the monthly returns. To the best of our recollection, all the large changes we have reviewed have been caused by corrections vendors made to the inputs.

Evolution of *SMB* and *HML*

We use our Data Library for our research, so we want the most accurate inputs and the best process when we update returns and other information. This means we change the historical values of *SMB*, *HML*, and other series when the historical data or our portfolio formation rules change. Table 1 provides specifics.

The table estimates the impact on the average monthly values of *SMB* and *HML* for 192607-202307 caused by the three changes in our portfolio formation rules described above and by four of the five major CRSP data projects. We cannot include the project to improve delisting codes and returns because CRSP corrected the database before the date of our earliest archive, 200212. We can report, however, that the average monthly premiums in FF (1993) for *SMB* (27 basis points, bps) and *HML* (40 bps), are both about 2.5 bps higher than the 200212 archive's average *SMB* and *HML* premiums for the same 196307-199112 period, 24.7 and 37.5 bps. In other words, in the roughly ten years between production of the factor premiums for FF (1993) and the 200301 update of our library, changes in the inputs and our process lower both monthly averages for 196307-199112 by 2.5 basis points.

We provide two estimates of the effect of each of the seven major changes in Table 1 on the 192607-202308 average monthly values of *SMB* and *HML*. In the first, we focus on the period in which a change can affect returns and compare the values of *SMB* and *HML* produced with and without the change. For example, while adding NYSE information for 1926-1962 to the daily database, CRSP discovered errors in prices and dividend ex-dates in the monthly data we use. They corrected the 1926-1962 errors between July 2005 and February 2006. The first line of Table 1 shows the effects of the corrections by comparing the monthly averages of *SMB* and *HML* for 192607-196206 computed using the CRSP data cuts for June 2005, before any corrections were

made, and February 2006, when the last project correction was made. CRSP may have made other improvements in the 192607-196206 data between June 2005 and February 2006, but we did not change our portfolio formation rules in this period, so we do not mind mislabeling stray data corrections in this brief window.

We also want to measure the effects of the seven changes in Table 1 on the full-period monthly averages of *SMB* and *HML*, so we translate the affected-period averages of the monthly differences to full-period averages. There are more months in the full-period averages, so the conversion from affected-period to full-period differences shrinks the monthly averages. The July 2005 to February 2006 corrections generated by the extension of CRSP's daily database, for example, increase the 192607-196206 average monthly premiums of *SMB* and *HML* by 1.90 bps and 1.52 bps. As the first line of Table 1 shows, spreading these averages for the 432 months of the affected period over the 1066 months of the full 1926-202308 period reduces them to 0.70 bps for *SMB* and 0.56 bps for *HML*. Some details describing the four major CRSP projects in Table 1 are from the release notes that accompany monthly CRSP updates. The rest are from personal correspondence with CRSP staff.

The biggest changes in the average values of *SMB* and *HML* in Table 1 are from CRSP's project to correct shares outstanding. CRSP corrected its database between June 2013 and December 2014. The changes affect the portfolios we form and the returns we compute from 192607 to 194706, reducing the average monthly value of *SMB* for that period by 3.39 bps and increasing the average monthly value of *HML* by 9.93 bps. The magnitude of the affected-period average difference in *SMB* is more than 75% larger than the next largest and the magnitude of the average difference in *HML* is more than five times the next largest. Although the affected-period changes caused by the shares outstanding project are the biggest in Table 1, their *t*-statistics, -0.90

for *SMB* and 1.75 for *HML*, are limited by the high volatility of the stock returns of 192607-194706. The magnitude of the shares outstanding corrections on the full-period averages are also the largest, 2.02 basis points for *HML* and -0.73 basis points for *SMB*.

The three changes in our rules for computing *SMB* and *HML* produce a mixed bag of small effects. Our response to FASB 106 lowers monthly average *SMB* during the affected period by -1.93 bps per month (t = -1.89) and increases average *HML* by 1.70 bps (t = 0.85). The average changes for FASB 109 are -1.70 bps (t = -1.11) for *SMB* and 1.50 bps (t = 0.40) for *HML*. For the switch to CRSP's Compustat links, the averages go the other way; the *SMB* average return increases 0.51 bps per month (t = 2.02) and averge monthly *HML* falls by -0. 42 bps (t = -1.13).

The effects of the three changes initiated by us on the full-period averages of *SMB* and *HML* are -0.40 bps and 0.35 bps for FASB 109, -0.56 bps and 0.49 bps for FASB 106, and 0.30 bps and -0.25 bps for the switch from our CRSP/Compustat links to CRSP's links. Together, the three changes reduce the 192607-202307 average monthly *SMB* by -0.66 bps and increase the monthly average *HML* by slightly less, 0.59 bps. The comparable changes for the combination of the four CRSP data projects are -0.03 bps for *SMB* and 2.57 bps for *HML*.

The combined effect of all seven changes in Table 1 on 192607-202307 average monthly premiums is a reduction of 0.69 bps for *SMB* and an increase of 3.17 bps for *HML*. Thus, changes in our rules for constructing *SMB* and *HML* are responsible for almost all the reduction in average *SMB* and less than 20% of the increase in average *HML*.

A final warning is in order. The details of factor construction are arguable, and there is no magic. After decades of experience, asset pricing research clearly recognizes that factor models, no matter how constructed, leave holes in the explanation of expected asset returns. Moreover, parameter instability and statistical estimation error combine to imply that expected return

estimates for specific assets or portfolios from asset pricing models are unreliable. The appropriate caveat is: use at your own risk.

Table 1 – Effect of Seven Changes in CRSP Data, Formulas for Book Equity, and Links between CRSP and Compustat on Average Factor Returns

The seven changes are identified in the first column and described in the text. Columns (2) and (3) show the impact of each change on the average *SMB* and *HML* premiums for 192607-202308. Column (4) is the last ("Old") data-cut before the change affects our computed returns and (5) is the first ("New") data-cut that fully incorporates the change. Columns (6)-(8) describe the months in which the change can affect *SMB* and *HML*, and (9)-(14) summarize the effects of the change on the monthly values of *SMB* and *HML* during that period. Ave and SD are the average and standard deviation of the monthly differences between the New and Old values of the premiums and *t*-stat tests whether a change's expected effect is reliably different from zero. The last row reports the cumulative impact of the seven changes on the 192607-202308 averages of *SMB* and *HML*. Returns are in percent, so for example, the first number in column (2), 0.0070, is 0.70 basis points.

	Impact on Full Period Average		Month of Cut		Affected Period			Monthly Differences in Affected Period					
								SMB			*HML*		
Change	*SMB*	*HML*	Old	New	Beg	End	Mths	Ave	SD	*t*-stat	Ave	SD	*t*-stat
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Daily NYSE 1925-62	0.0070	0.0056	200506	200602	192607	196206	432	0.0190	0.309	1.28	0.0152	0.453	0.70
Arca Exchange	0.0000	0.0000	200706	200707	200603	200706	16	0.0019	0.028	0.27	0.0019	0.023	0.33
Shares Outstanding	-0.0073	0.0202	201305	201412	192607	194706	252	-0.0339	0.598	-0.90	0.0933	0.846	1.75
FASB 109 Adjustment	-0.0040	0.0035	201606	201607	199307	201606	276	-0.0170	0.254	-1.11	0.0150	0.621	0.40
FASB 106 Adjustment	-0.0056	0.0049	202007	202008	199207	202007	337	-0.0193	0.188	-1.89	0.0170	0.369	0.85
CRSP/Compustat Links	0.0030	-0.0025	202106	202107	196307	202106	696	0.0051	0.066	2.02	-0.0042	0.098	-1.13
Quarterly Dividends	-0.0000	-0.0001	202111	202203	199001	202111	383	-0.0000	0.011	-0.05	-0.0003	0.018	-0.29
Sum of Seven Effects	-0.0069	0.0317											

Appendix – Constructing our Proprietary Links between CRSP and Compustat

We used proprietary links between CRSP and Compustat to compute *SMB* and *HML* before July 2021. Almost all Compustat data are for firms, not securities, so we organize the construction of *SMB* and *HML* around Permcos (CRSP's identifier for firms), not Permnos (CRSP's identifier for securities). The code to compute *SMB* and *HML* cycles through Permcos and before July 2021 used our links to extract the appropriate information for each from CRSP and Compustat. This appendix describes our links. Most of it was written in 2001.

Linking Permnos to Permcos

Some Permcos have more than one Permno. There are several possible reasons.

1. Multiple Share Classes – If the date ranges (MBEG-MEND) overlap for two Permnos, we combine them in the same way we would combine securities in a portfolio. The firm's return is the weighted average of its securities' returns and its market equity is the sum of its securities' market equities.
2. Sequential Listings – Several companies go off the exchange and then return. Compustat tends to treat the whole sequence as one company with one GVKey (Compustat's non-permanent identifier for a firm). CRSP often treats the second security as a new company, but sometimes it gives them the same Permco. We override this. If the securities do not overlap, we give the later Permno a fake Permco, in the range 100,000-110,000.
3. Tracking Stock – Tracking stock is stock issued by a parent company that tracks the performance of a particular division or subsidiary. In most cases the parent retains almost all tracking stock shares. Since the parent's market equity includes almost all the sub's, merging the two securities would be double counting. Moreover, the parent company's

balance sheet and income statement reflect its claim on the sub. Thus, we omit tracking stocks entirely when computing *SMB* and *HML.*

4. Separate Operating Divisions – Some companies split the firm into separate operating divisions with a tracking stock for each division and no overall or parent stock. For example, on August 8, 1999 Quantum Corp shareholders received two shares of Quantum HDD and one of Quantum DSS for each share of Quantum. Compustat has a separate record for each division, so we treat the two as separate companies and assign one a fake Permco. US West Communications / MediaOne Group is also in this category.
5. "Distinct" Companies – CRSP links some companies through Permco that Compustat separates. For example, the three Bally Permnos are linked through a common Permco. Compustat has a separate GVKey for each and their annual reports, 10K's, etc., suggest we can treat them as separate companies. In these cases, we override CRSP and assign each security (Permno) its own fake Permco.

Identifying the Permnos for a Permco

We identify Permno links by matching Permcos in the CRSP header structure. There are 400 Permcos, for example, with more than one Permno in the October 2001 monthly CRSP database. The Permnos for 250 of these overlap – at least two Permnos have data for the same month. The remaining 150 Permcos do not have overlapping Permnos. (These counts exclude Permnos that do not have at least one name record with a share code of 10 or 11 – ordinary shares.)

When developing our link procedure, we tried to identify additional Permco links using the Compustat CST_Link structure. Every additional Compustat link, however, was not in the CRSP links because at most one of the Permnos had a share code of 10 or 11. Of the approximately 20 additional links, several were REIT's, others were ADR's, and others were not even on CRSP.

Linking to Compustat

We link each Permco to at most one GVKey each year. We do, however, concatenate Compustat records. For example, we might use GVKey A for Compustat years 10-40 and GVKey B for 41-51. As a result, we have to identify not only the GVKey's that are linked to a Permco, but also the time period to use each GVKey.

We use the Compustat CST_Link structure to identify the GVKey's linked to each Permco. Starting with the Permco, there are four cases:

1. The Permco is not linked to any GVKey's.
2. The Permco is linked to only one GVKey. This may involve links through multiple Permnos but this is no problem. If the Permnos overlap, they are merged into one company portfolio, so the multiple links are irrelevant. Each non-overlapping Permno will link to the GVKey for the appropriate period.
3. The Permco is linked to multiple GVKey's through one Permno. Since there are multiple GVKey's, we have to identify the appropriate period for each. There are 220 Permcos in this category. For 97 of these, the GVKey periods do not overlap, so the appropriate periods are obvious. For the remaining 123 Permcos, we identify each GVKey's period by hand. (CRSP's apparent rule – when GVKey's overlap, use the one with the later LEND – makes many bad decisions.)
4. The Permco is linked to multiple GVKey's through multiple Permnos. Many of these are special cases, such as Bally, Quantum, and tracking stocks. The remainder are not a problem if the Permnos do not overlap. Each Permno will have its own (true or fake) Permco and each Permco should be linked to only one GVKey. Operationally, when linking GVKey's to Permcos thru CST_Link, we must confirm that the CST_Link's

Permno is in this Permco's list of Permnos. If the Permnos overlap, we must identify which GVKey to use each year.

We use the Compustat CST_Link structure to identify multiple Permcos linked to a single GVKey. There are two cases.

1. One GVKey points to multiple Permcos that do not overlap. This is not a problem. When starting with CRSP, we almost always use only the Compustat data that matches the CRSP period. Thus, each Permco will use its own section of the GVKey data. Note, to do this right, we have to be concerned with Permco-linked Permnos that overlap with the linked Permno and induce overlap with the GVKey-linked Permco. Of the 198 firms in this category, the CST_Link says that one (WebFinancial) does have multiple Permco-linked Permnos, but it does not induce overlap. To do this right, however, we must use the CRSP links.

2. One GVKey points to multiple Permcos that overlap. These 24 cases are complicated and many plagued us earlier in the linking process. In fact, in 16 of the cases, one of the Permcos points to an additional GVKey. The best way to deal with these is to use the LinkBeg and LinkEnd from the CST_Link structure.

   The case of Ventas/Vencor, in Table A1, is typical. There we link Permco 34766 to GVKEY 17239 for the whole period CRSP links them, from LinkBeg = 19980501 to LinkEnd = 19990605. We link Permco 10302 to GVKEY 17239 from 19890919 to 19980430 and we link it to GVKey 110179 after 19980430.

Table A1 – CST Links for Ventas/Vencor

GVKey 17239

	Permco	Permno	LinkBeg	LinkEnd	LBeg	LEnd	BegDat	EndDat	Delist	Compustat Name	CRSP Name
1	10302	75819	19890919	19980430	39	51	198909	200110	100	Vencor Inc	Ventas Inc
2	34766	86103	19980501	19990604	39	51	199805	199906	584	Vencor Inc	Vencor Inc New

Permco 10302

	GVKey	Permno	LinkBeg	LinkEnd	LBeg	LEnd	BegDat	EndDat	Delist	Compustat Name	CRSP Name
1	17239	75819	19890919	19980430	39	51	198909	200110	100	Vencor Inc	Ventas Inc
2	110179	75819	19980501	99999999	49	51	198909	200110	100	Ventas Inc	Ventas Inc