TECHNICAL APPENDIX

FOR "THE STATE OF PRIVATE PENSIONS: CURRENT 5500 DATA"

By Marric Buessing and Mauricio Soto*

The Center for Retirement Research at Boston College is releasing an update of the pension tabulations of the 5500 Form Annual Reports for plans with 100 or more participants. This update includes tabulations up to 2003, the latest year for which the official raw data files are available. The release includes an *issue in brief*[1] with the data highlights, a data appendix with comprehensive tabulations, and this technical appendix.

1. Data Description

Private pension plan sponsors are required to file a return, known as the Form 5500 series, with the Department of Labor (DOL). These returns are composed of a main Form — 4 pages with the basic plan information — and 11 accompanying schedules. The main Form, schedule B (actuarial information), schedule H (financial information), and schedule T (pension coverage information) are the main sources of data used in this appendix.[2] The forms detail the number and type of participants (active, retired or separated, and other beneficiaries), contain an income statement (including plan contributions and benefits paid), and a balance sheet. For defined benefit plans, the forms also include detailed actuarial information (actuarial value of liabilities, cost method, and other assumptions). The 5500 series data also include plan-level information that allows for the identification of the plan sponsor.

Governmental agencies are the primary users of these data (IRS, DOL, PBGC, and SSA). For researchers, the DOL releases two public versions of the data: the raw data with actual filings available from the DOL's Public Disclosure Room, and the "research file" — a slightly edited version that contains all plans with 100 or more participants and a 5 percent sample of small plans — available from the Office of Policy and Research of the DOL. This appendix describes the raw data filings, which allow researchers to access the most recent 5500 filings. And although this analysis is restricted to plans of 100 or more participants, it could easily be extended to account for smaller plans.

The raw data do not come in a user-friendly format. The electronic files are simple text files, with each row corresponding to one observation. A hard copy of a "File Layout" is generally included with the data to serve as a codebook for every year of data. No further instructions are included with the original data. The layout, which is also not user-friendly, contains the code name for each variable, the columns in which they are located in the data, and the item number to link the data to actual paper Form. The following is an example of the file layout:

Field Name	Form Item Number	Length	Beg Pos	End Pos
SPONS-DFE-NAME	2a	71	199	269

* Marric Buessing is a Research Associate at the Center for Retirement Research at Boston College. Mauricio Soto is a Senior Research Associate at the Center. Baris Yoruk provided research assistance in the initial stages of this project. The authors thank Daniel Beller, formerly with the Department of Labor (DOL), for important clarifications of the use and availability of the data. David McCarthy and Kevin Schutt (both with the Office of Policy and Research at the DOL) provided helpful guidance. The authors are also grateful to Anja Decressin (DOL) for pointing out clever ways to use these data in conjunction with other datasets, Vicky Kiosse for discovering duplicated observations in the latest available data, and Francis Vitagliano for explaining regulatory changes that affected the Form 5500 series since their inception. Although the staff of the DOL was extremely helpful in the development of this *brief*, the final analysis, accompanying data tables, and technical appendix are the sole responsibility of the authors and are in no way endorsed by the DOL.

This says that the variable SPONS-DFE-NAME, from the actual 5500 Form field 2a, has a length of up to 71 characters, and is located in columns 199-269 of the raw data. Users must go to the actual Form, item 2a, to learn that the variable represents the "Plan Sponsor's Name." As the 5500 Form changes over time, so does the file layout. Users must be careful to use the appropriate file layout for each corresponding year.

The 5500 series data come in plan-level format, which means that each observation represents one plan. This is the product of the filing requirements, which demand separate filings for each plan offered by the sponsors. Each plan can be uniquely identified by two variables: the "Employer Identification Number" (EIN, nine digits), and the "Plan Number" (PN, three digits). The first maps each plan to a unique sponsor; the second serves to differentiate each plan within a particular sponsoring firm. Putting together these two variables generates a 12-digit code that uniquely identifies each plan over time. Tables A1 and A2 show the total observation counts from the raw data, the gross number of pension plan-level observations, the number of unique pension-plan level observations, and the number of firm-level observations for the 1990-2003 period.

Only about a third of the observations contained in the data corresponds to pension plans — the focus of this study. The remainder corresponds to welfare plans and other types of filers (Table A1). Pension plans can easily be spotted using the binary variable "Type of Pension Benefit Indicator" which takes a value of "1" for pension plans. To control for possible miscoding, this variable is complemented by using variables that should only be filled by pension plans (i.e., "Type of Pension Benefit Indicator" for years prior to 1999).

Table A1. Number of Observations from Raw 5500 Series Data, 1990-1998

	1990	1991	1992	1993	1994	1995	1996	1997	1998
Total	142,398	150,827	176,980	186,182	191,549	201,289	206,181	212,008	215,968
Plan Level									
Gross	53,845	54,303	57,385	58,925	59,851	62,104	63,876	65,866	66,856
Net	53,229	53,484	56,819	58,261	59,211	61,324	63,057	64,993	66,339
Firm Level	36,557	37,647	39,638	41,091	42,317	44,326	46,277	48,340	50,312

Source: Authors' calculations from the raw universe 5500 data files.

It is evident from Table A1 that there are repeated plan-level observations, as defined by the combination of EIN and PN. The criterion used to select a unique observation from each set of repeated plan-level observations was to keep the latest filing that contains non-missing information. This can easily be accomplished for recent years using the date of filing variables. For years prior to 1999, the criterion was to keep plans with End of Year Assets greater than zero, and of those remaining, keep the last entry found in the raw data for that combination of EIN and PN.

The inclusion of the EIN, a nine-digit federal tax identification number that uniquely identifies the plan sponsor, allows for approximating the data to a firm-level dataset.[3] This can be useful for researchers interested in merging the 5500 data with other firm-level datasets.

In 1999, the Form 5500 was modified in several ways. One of the changes was to include small plans — those with less than 100 participants — in the main 5500 Form. Prior to that year, small plans were required to fill out a special version of the Form, known as the 5500C/R. Table A2 shows the effect of this change in the raw data. The total number of observations more than triples relative to previous years because the 1999-2003 period includes small pension plans and welfare plans. The presence of Schedule H is used to identify plans with 100 or more participants, since only large plans file this schedule. Table A2 shows the number of plans with 100 or more participants and the corresponding firm-level measure from the raw data.

Table A2. Number of Observations from Raw 5500 Series Data, 1999-2003

	1999	2000	2001	2002	2003
Total	668,618	771,748	995,752	720,824	889,247
Plan Level					
Gross	53,765	57,189	106,066	106,073	92,747
Net	46,581	53,679	70,641	70,603	67,569
Firm Level	37,419	43,596	56,281	56,723	55,403

Source: Authors' calculations from the raw universe 5500 data files.

Note that in 1999 and 2000, the plan-level and firm-level observation counts are significantly lower than in previous years. In 1999, the Form changed — which naturally increased the number of errors — and a new private vendor was made responsible for the 5500 Form processing — a procedure previously done by the IRS. As a result, a large number of the 1999 and 2000 filings was missing. For 2003, the data are preliminary, and some filings could be missing. This produces lower aggregate values for these years. (See the section "Imputations" for a suggestion of how to deal with these three years to estimate values for missing plans).

2. Identifying Plans by Type: DB, DC, 401(k) and Cash Balance Plans

The coding in the raw data follows closely the actual paper form coding. For the period 1990-2003, the data include variables for the "Type of Pension Benefit Plan" or "Plan Characteristics Codes" which serve to identify the type of pension plan for each observation. Table A3 summarizes the variables that classify pension plans by type.

Table A3. Coding for the Type of Pension Plan, 1990-2003

		Type-Pension-Benefit-Ind		Type-Pension-Benefit-Code
	Type of Plan	1990-1991	1992-1998	1999-2003
DB	Defined Benefit	1	1	1A, 1B, 1D, 1E, 1G, 1H
	Cash Balance			1C
DC	Profit-Sharing	2, DC Type A	2	2E
	Stock Bonus	2, DC Type B	3	2I
	Target Benefit	2, DC Type C	4	2B
	Money Purchase	2, DC Type D	5	2C
	Other	2, DC Type E	6	2A, 2D, 2F, 2G, 2K, 2P, 2O
Other	414(k)	3	7	1F
	403(b)(1)	4	8	2L
	403(b)(7)	5	9	2M
	408	6	0	2N
	Other	7		

Source: Authors' calculations from raw universe 5500 data files.

For 1990 and 1991, the "Defined Contribution Type" field includes a one-character code (A-E) for defined contribution plans. For 1992-1998, the codes are straightforward — one-digit codes 0-9 identify the type of pension plan. These codes are used in conjunction with a string search to determine the type of plan for each observation. Table A4 tabulates the type of pension plan for 1990-1998.

Table A4. Pension Plans by Type of Pension, 1990-1998

	1990	1991	1992	1993	1994	1995	1996	1997	1998
Defined Benefit	20,385	19,681	19,135	18,464	17,732	17,253	16,494	15,793	14,915
Defined Contribution									
Profit-Sharing	25,190	17,853	29,039	31,292	33,418	36,078	38,752	41,707	44,205
Stock Bonus	1,453	344	1,486	1,537	1,578	1,609	1,631	1,606	1,571
Target Benefit	167	74	185	188	164	155	145	139	124
Money Purchase	4,115	1,902	4,219	4,250	4,211	4,227	4,182	4,094	4,041
Other	1,919	13,630	2,755	2,530	2,108	2,002	1,853	1,654	1,483
Other	0	0	0	0	0	0	0	0	0
Total	53,229	53,484	56,819	58,261	59,211	61,324	63,057	64,993	66,339

Source: Authors' calculations from the raw universe 5500 data files.

Note the large number of defined contribution plans classified as "Other" for 1991. The version of the 1991 raw data used for this appendix does not include the "Defined Contribution Type" field, so it is difficult to pinpoint the specific class of defined contribution plan because the alphabetical identifier is missing (i.e., it is virtually impossible to distinguish between a profit-sharing plan and a money purchase). The section "Imputations" includes a suggestion of how to impute the type of plan for 1991.

For 1999-2003, changes in the Form allow plan administrators to report up to ten plan characteristics, with each characteristic defined by a two-digit code. The Form changes implemented in 1999 required a more complex coding for DB and DC plans, because the information was spread across ten different variables.[4] To identify the type of pension plan, the first two-digit code is used. For plans that can not be categorized, the next two digits are used. This sequence continues for plans that can not be categorized until all ten plan characteristics have been checked. The procedure is then complemented with a string search for each type of plan in the "Plan Name" field to account for plans that can not otherwise be identified. The tabulations for these years are presented in Table A5.

Table A5. Pension Plans by Type of Pension, 1999-2003

	1999	2000	2001	2002	2003
Defined Benefit	9,778	10,050	12,892	12,263	11,240
Defined Contribution					
Profit-Sharing	31,712	38,336	50,941	51,854	50,700
Stock Bonus	813	916	1,090	1,088	1,003
Target Benefit	75	76	86	85	71
Money Purchase	2,635	3,089	4,047	3,655	3,012
Other	1,568	1,212	1,585	1,657	1,543
Total	46,581	53,679	70,641	70,603	67,569

Source: Authors' calculations from the raw universe 5500 series data.

401(k) plans are a subset of defined contribution plans. Before 1992, 401(k) plans can be identified by the field "Cash Deferred Arrangement" (1 for 401(k) plans, 0 otherwise); between 1992 and 1998, "Pension Feature Code" is used to flag the 401(k) plans. This variable is generally an 8-character long variable, which includes the letter "G" for 401(k) plans. To ensure that all 401(k) plans are identified, these flags are complemented with a string search in the plan name to mark plans that contain "401(K)," "K401," "401PW," "401-PW," or "401 K."

Some cash balance plans can also be identified for the period 1990-1998. For years before that, the only way to identify cash balance plans is by a string search, which is likely to produce a lower count than the actual number of cash balance plans for those years. Table A6 shows the counts for 401(k) and Cash Balance Plans for 1990-1998.

Table A6. Pension Plans: Cash Balance and 401(k) Plans, 1990-1998

	1990	1991	1992	1993	1994	1995	1996	1997	1998
Defined Benefit									
Traditional	20,365	19,655	19,103	18,425	17,683	17,186	16,416	15,706	14,814
Cash Balance Plans	20	26	32	39	49	67	77	87	101
Defined Contribution									
401(k)	18,456	6,935	20,703	25,296	28,015	31,198	34,191	37,512	40,357
Other	14,388	26,868	16,981	14,501	13,464	12,873	12,373	11,688	11,067

Source: Authors' calculations from the raw universe 5500 data files.

Two issues are evident from table A6. First, counts for 401(k) plans for 1991 are significantly lower. This is because the "Cash Deferred Arrangement" field is not available in the raw 1991 data used in these calculations. Second, although less noticeable, a small but significant number of defined contribution plans have missing values for the "Cash Deferred Arrangement" field in 1992. The section "Imputations" contains suggestions on how to deal with these years.

For the years 1999-2003, 401(k) plans are identified as those with the code "2J" in the "Type of Pension Benefit Indicator" field. These years also allow for the easy identification of Cash Balance plans, which are those with "Type of Pension Benefit Code" equal to "1C." Table A7 presents the raw counts of 401(k) and Cash Balance Plans for 1999-2003.

Table A7. Pension Plans: Cash Balance and 401(k) Plans, 1999-2003

	1999	2000	2001	2002	2003
Defined Benefit					
Traditional	9,203	9,479	12,016	11,333	10,267
Cash Balance Plans	575	571	873	927	973
Defined Contribution					
401(k)	29,868	35,449	47,662	48,982	48,258
Other	6,941	8,182	10,090	9,361	8,079

Source: Authors' calculations from the raw universe 5500 data files.

3. Participant Counts

Each plan-level observation has a precise count of the number of participants and active participants in the plan. These could be easily aggregated to obtain a total participant count from the raw data. There are, however, two problems with these numbers, which would tend to overestimate participation in pension plans. First, the active participant count might include non-vested employees and 401(k)-eligible employees who do not participate in their plans. Second, these numbers do not control for dual coverage — individuals that are covered by more than one pension plan, even within the same firm.[5] The following procedures intend to replicate the adjustments done by the DOL to obtain their official participation numbers that appear in Table E4 of the U.S. Department of Labor (2004).

a. Adjusted Active Participants

The basic idea is to replace the active participant count from the raw data with the number of employees that actually benefit under the plan — a figure generally reported for plans which are not collectively bargained and which include highly compensated employees. For defined contributions, active participants are further adjusted to ensure that only participants with non-zero balances are counted as active.

For 1990-1998, the procedure is the following. For defined contribution plans, the active participant count was first replaced by the number of defined contribution plan participants with account balances net of separated and retired participants both receiving and planning to receive benefits. This is done for plans that meet the following conditions: A) the new resulting active participant count is greater than 80% of the original active participant count; B) the count of participants with account balances is less than the number of active participants; and C) the count of participants with account balances is greater than the sum of the fully and partially vested active participants and the separated and retired participants receiving and planning to receive benefits.

The resulting number for defined contribution active participants and those for defined benefit plans are then replaced by the number of employees benefiting under the plan, if three conditions are met: D) the number of employees benefiting under the plan is less than the active participant count; E) the number of employees is greater than 80 percent of the active participant count; and F) the number of employees benefiting is more than the sum of fully and partially vested active participants.

These adjustments can only be made to plans which are not collectively bargained and include highly compensated employees. For all other plans, as well as for those that remained unchanged after the first round of adjustments, the active participant counts are adjusted to mimic the average change from plans that were adjusted. This is done separately for defined contribution and defined benefit plans. The unadjusted and adjusted active participant counts are presented in Table A8.

Table A8. Adjustments to Active Paricipant Counts, 1990-1998

	1990	1991	1992	1993	1994	1995	1996	1997	1998
Active Participants	55,387	55,685	57,343	57,956	57,738	59,499	60,817	64,399	66,197
Unadjusted	55,387	55,685	57,343	57,956	57,738	59,499	60,817	64,399	66,197
Adjusted	54,083	51,768	52,666	56,303	55,827	57,457	58,613	61,928	63,305

Source: Authors' calculations from the raw universe 5500 data files.

For 1999-2003, the line items for "Fully Vested" and "Partially Vested" active participants were removed, and the "Employee Benefiting Under the Plan" line was moved to Schedule T. The adjustments to these years differ from those for previous years in two ways. First, condition C changes to: C) the count of participants with account balances is greater than the sum of the active participants and the separated and retired participants receiving and planning to receive benefits. (Since the number of active participants is always greater than or equal to the sum of the fully and partially vested active participants, the new condition C is more binding than before. Second, condition F is relaxed. The resulting unadjusted and adjusted active participant counts are shown in Table A9.

Table A9. Adjustments to Active Paricipant Counts, 1999-2003

	1999	2000	2001	2002	2003
Active Participants					
Unadjusted	45,769	48,876	68,652	69,646	66,490
Adjusted	41,721	44,106	63,873	64,830	59,997

Source: Authors' calculations from the raw universe 5500 data files.

After these adjustments have been done, the number of active participants for each plan can be added across all plans to obtain the aggregate active participant number, which replicates table E10 of the U.S. Department of Labor (2004). Although these adjusted numbers provide a more accurate count of the active participants of each plan, the aggregate numbers still include double counting of individuals that might be covered by more than one plan. The following section details a strategy to adjust the active participation numbers to avoid double-counting due to dual coverage.

b. Dual Coverage

To control for dual coverage, it is necessary to transform the data from plan-level to firm level. This transformation is simple for the dollar amounts and the plan counts, but requires some assumptions to aggregate plan participants and active participants. The first step is to obtain the counts of defined benefit and defined contribution active participants by firm:

Assumption 1. For defined benefit plans, the main assumption is that each plan offered by the sponsor covers different groups of workers, so that the total defined benefit participant count for a particular employer is the sum of the participants of each plan. The exception is those plans that include the terms "supplemental" or "past service" in their pension plan name or description. The participants in these plans are excluded from the participant count. The following example shows the participant count for the firm XYZ, with 3 defined benefit plans, one of them supplemental. The total count is 250.

Plan Level (Raw data)

Plan Number	EIN	Participants	Plan Name
1	XYZ	100	DB1
2	XYZ	150	DB2
3	XYZ	60	DB3 Supplemental

Firm Level

	Participants			
	DB1	DB2	DB3	Net of Dual Coverage
XYZ	100	150	60	250

Assumption 2. For defined contribution plans, the assumption is that plans of the same type cover different groups of workers, but plans of a different type offer dual coverage. This means that the participants in defined contribution plans are obtained by first aggregating the participants by each type of defined contribution plan (profit sharing, stock bonus, target benefit, and money purchase), and then selecting the maximum number of participants out of these aggregates. The following example shows the participant count for firm ABC, with 5 defined contribution plans (2 profit sharing, 2 stock bonus, and 1 money purchase). The defined contribution participant count is 180.

Plan Level (Raw data)

Plan Number	EIN	Participants	Plan Name
1	ABC	100	Profit Sharing 1
2	ABC	80	Profit Sharing 2
3	ABC	60	Stock Bonus 1
4	ABC	40	Stock Bonus 2
5	ABC	75	Money Purchase

Firm Level

	Participants			
	Profit Sharing	Stock Bonus	Money Purchase	Net of Dual Coverage
ABC	180	100	75	180

For firms that offer only defined benefit plans or only defined contribution plans, these are all the required adjustments. But since some firms offer some form of defined benefit plans combined with some form of defined contribution plans, these tabulations still include double counting.

Assumption 3. For firms that offer defined contribution plans and defined benefit plans, assumptions 1 and 2 are used to obtain participant counts for each type of plan. Then, the assumption is that these plans usually cover the same group of workers. Defined benefit plans are generally considered to be the primary plan, and defined contribution plans are considered to be supplemental. The active participants of supplemental plans for firms that offer both defined benefit and defined contribution plans are assigned as covered by both defined benefit and defined contribution. The exception is when the number of participants in defined contribution plans is more than four times the number of participants in defined benefit plans. In that case, it is assumed that these plans cover different groups of workers, and the participants are assigned accordingly. The following example illustrates the participant count for firm DEF. The counts are: "defined benefit only": 0, "defined contribution only": 0 and "both": 180. (The exception means that if the number of defined contribution participants were above 1000, then the coverage by both would be zero, and the counts for "defined benefit only" and "defined contribution only" would correspond to the participant count for each type of pension).

Plan Level (Raw data)

Plan Number	EIN	Participants	Plan Name
1	DEF	100	DB1
2	DEF	150	DB2
3	DEF	60	DB3 Supplemental
4	DEF	100	Profit Sharing 1
5	DEF	80	Profit Sharing 2
6	DEF	60	Stock Bonus 1
7	DEF	40	Stock Bonus 2
8	DEF	75	Money Purchase

Firm Level

	Participants				
DEF	DB Participants	DC Participants	DB Only	DC Only	Both
	250	180	0	0	180

Using these definitions replicates Table E4a from U.S. Department of Labor (2003). The resulting participant counts from the raw data are presented in Table A10. Note that the raw data used in this appendix includes plans with 100 or more participants only, which explains why primary defined benefit participants follow the official tabulation closely but defined contribution numbers are significantly lower — almost all of the defined benefit participants are in large plans while many defined contribution participants are in small plans.

Table A10. Adjustments to Active Participant Counts, 1990-2003†
Thousands

	DOL Tabulations			Authors' Calculations		
	Workers Covered by a			Workers Covered by a		
Year	Primary Defined Benefit Plan	Primary Defined Contribution Plan	Supplemental Defined Contribution Plan(s)	Primary Defined Benefit Plan	Primary Defined Contribution Plan	Supplemental Defined Contribution Plan(s)
1990	26,323	16,116	15,671	25,691	10,437	17,278
1991*	25,701	17,133	15,287	25,073	11,294	17,241
1992*	25,318	19,474	16,300	24,647	12,342	18,298
1993	25,091	19,780	16,621	23,730	13,194	18,799
1994	24,591	20,948	16,516	23,338	14,083	17,806
1995	23,531	23,038	16,482	22,397	16,007	18,398
1996	23,262	24,173	17,199	22,049	16,906	18,952
1997	22,724	27,045	18,531	21,568	20,820	18,874
1998	22,972	29,139	18,526	22,084	21,800	18,876
1999*				20,952	22,484	18,743
2000*				19,379	25,024	18,482
2001				19,040	26,342	18,252
2002				19,399	26,624	18,589
2003				18,626	25,725	17,246

Source: U.S. Department of Labor (2003); and authors' calculations from raw universe 5500 data files.

* 1991, 1992, 1999, 2000, and 2003 include imputations.
†DOL tabulations are for all plans; Authors' calculations are for plans with 100 or more participants.

Table A11 shows the raw numbers used to replicate Table E4 from the U.S. Department of Labor (2004). The authors' calculations use plans with 100 or more participants; the official tabulations include all plans. To create a consistent series that can be connected to the official numbers — shown in Table E4 of the Data Appendix included with this release — the percentage changes from the calculations from the raw data are applied to the official tabulations.

Table A11. Pension Plan Participation, by Type of Coverage, 1999-2003*
Thousands

	DOL Tabulations			Authors' Calculations		
Year	Defined benefit plan only	Defined contribution plan only	Both	Defined benefit plan only	Defined contribution plan only	Both
1990	12,273	16,023	13,932	9,611	11,541	13,659
1991	12,233	17,024	13,370	9,326	12,124	14,571
1992	11,557	19,340	13,665	8,456	13,154	14,865
1993	10,449	19,632	14,537	7,895	14,199	15,220
1994	9,929	20,781	14,551	8,189	15,064	14,464
1995	8,978	22,734	14,417	7,555	17,542	14,445
1996	7,830	23,954	15,303	7,436	18,897	14,533
1997	6,768	26,785	15,851	7,266	22,776	14,559
1998	7,061	28,839	15,802	7,783	24,091	14,448
1999				7,123	24,902	14,282
2000				6,259	27,792	13,944
2001				6,200	29,112	13,649
2002				6,696	29,338	14,276
2003				6,579	30,061	14,943

Source: U.S. Department of Labor (2004); and authors' calculations from raw universe 5500 data files.

*Note: DOL tabulations are for all plans; Authors' calculations are for plans with 100 or more participants.

4. Imputations

Imputations are necessary to generate the aggregate values for the 1991, 1992, 1999, 2000, and 2003 datasets. For the years 1991 and 1992, only minor imputations are needed. The datasets seem to have the full universe of plans, but one or two variables appear to be miscoded; for 1999 and 2000, significant imputations are required, because the raw data are missing a large number of plans from the universe; for 2003, almost all of the plans in the universe are available from the raw datasets, so only a few observations need to be imputed. The data appendix presents the imputed numbers along with the aggregate numbers from the raw datasets.

a. 1991 and 1992

For 1991, it is necessary to impute two fields: the "Defined Contribution Type" field, which indicates the type of pension benefit of each defined contribution plan, and the "Cash Deferred Arrangement" field, which serves to identify 401(k) plans; for 1992, the "Cash Deferred Arrangement" field seems to be missing for some plans (see tables A4 and A7).

The year 1990 is used as the baseline to impute 1991 and 1992. The strategy is simple: 1) identify the plans that are in both 1990 and the year to be imputed; 2) for defined contribution plans that are classified as "Other" in 1991, but are classified as one of the defined contribution types in 1990 (profit-sharing, stock bonus, target benefit, or money purchase), replace the type of defined contribution pension from "Other" with the type reported in 1990; 3) for 1991 and 1992, indicate defined contribution plans as 401(k)-type if these plans were identified as 401(k) in 1990. The results of these imputations are reported in table A12.[6]

TABLE A12. IMPUTATIONS FOR 1991 AND 1992

	Actual				Imputed	
	1990	1991	1992	1993	1991	1992
Defined Benefit	20,385	19,681	19,135	18,464	19,681	19,135
Defined Contribution						
Profit-Sharing	25,190	17,853	29,039	31,292	25,540	29,039
Stock Bonus	1,453	344	1,486	1,537	1,013	1,486
Target Bonus	167	74	185	188	136	185
Money Purchase	4,115	1,902	4,219	4,250	3,559	4,219
Other	1,919	13,630	2,755	2,530	3,555	2,755
401(k)	18,456	6,935	20,703	25,296	18,067	23,011
Total	53,229	53,484	56,819	58,261	53,484	56,819

Source: Authors' calculations from raw universe 5500 data files.

Note that no actual plans have been imputed for these two years, as only the defined contribution classifications are changed by these imputations. These adjustments are necessary to control for dual coverage (Table E4 of the data appendix) and to disaggregate the results by 401(k) and other types of defined contribution plans (Table D6). But these imputations do not affect the aggregate level of assets, benefits, contributions, or participants.

b. 1999 and 2000

The 1999 and 2000 raw data seem to have a significant number of plan-level observations missing — about 30 and 20 percent respectively. For these years, the entire record for a large number of pension plans requires imputations. The strategy is to 1) identify which plans are missing from the 1999 and 2000 raw datasets; and 2) impute the values for these plans for 1999 and 2000.

For 1999, the detailed methodology is as follows:

1. Identify plans that are present in 1) both the 1998 and 2000 raw datasets; or 2) both the 1998 and 2001 raw datasets (using the two adjacent datasets insures that plans were not terminated in the year prior to imputation). Of these, keep only the plans that are not in the 1999 data. The resulting data, made up of 16,995 plan-level observations, are the plans to be imputed.
2. Then, the 1999 data contain 46,594 plans from the raw data (call this Sample A) and 16,995 plans to be imputed (call this Sample B). For the portion of plans in Sample A that are in both 1998 and 1999, measure the average percent change for each variable (assets, participants, contributions, and benefits) by type of plan (defined contribution or defined benefit). Then apply these percent changes to the 1998 values of Sample B to impute the variable values for 1999.
3. To get a more accurate estimate of assets, use the beginning of year asset values in the year 2000 for those plans that are in Sample B and are also present in 2000.

For 2000, the detailed methodology is as follows:

1. Identify plans that are present in 1) both the 1999 and 2001 raw datasets; or 2) both the 1998 and 2001 raw datasets. Of these, keep only the plans that are not in the 2000 data. The resulting data, made up of 12,018 plan-level observations, are the plans to be imputed.

2. Then, the 2000 data contain 53,687 plans from the raw data (call this Sample C) and 12,018 plans to be imputed (call this Sample D). For the portion of plans in Sample C that are in both 1999 and 2000, measure the average percent change for each variable (assets, participants, contributions, and benefits) by type of plan (defined contribution or defined benefit). Then apply these percent changes to the 1999 values of Sample D to impute the variable values for 2000.

3. To get a more accurate estimate of assets, use the beginning of year asset values in the year 2001 for those plans that are in Sample D and are also present in 2001.

Table A13 shows the basic statistics for these imputations. The resulting aggregate values are provided in the Data Appendix.

Table A13. Imputations for 1999 and 2000

	Actual				Imputed	
	1998	1999	2000	2001	1999	2000
Defined Benefit	14,915	9,779	10,054	12,892	13,719	13,016
Defined Contribution	51,424	36,815	43,633	57,749	49,870	52,689
Total	66,339	46,594	53,687	70,641	63,589	65,705

Source: Authors' calculations from raw universe 5500 data files.

c. 2003

Less than 5 percent of the plan-level observations seem to be missing from the 2003 raw data used in this appendix. The strategy is similar to that used to impute 1999 and 2000: 1) identify the plans that are missing in the 2003 raw data set; and 2) impute the values for these observations. To impute this year, the latest available year of complete data (2002) is used as a baseline. Plans that have less than 150 participants in 2002 and those that checked the "Final Return" box in the 2002 Form are not considered for imputation (this is to exclude small plans and to control for plan terminations, respectively).

For 2003, the detailed methodology is as follows:

1. Identify plans that are present in 2002 but are not in 2003. Of these, keep only the plans with more than 150 participants in 2002 that did not check the "Final Return" box in the 2002 main 5500 Form. The resulting data, made up of 3,900 plan-level observations, are the plans to be imputed.

2. Then, the 2003 data contain 67,569 plans from the raw data (call this Sample E) and 3,900 plans to be imputed (call this Sample F). For the portion of plans in Sample E that are in both 2002 and 2003, measure the average percent change for each variable (assets, participants, contributions, and benefits) by type of plan (defined contribution or defined benefit). Then apply these percent changes to the 2002 values of Sample F to impute the variable values for 2003.

Table A14 shows the basic statistics for the 2003 imputations. The resulting aggregate values are provided in the Data Appendix.

Table A14. Imputations for 1999 and 2000

	Actual		Imputed
	2002	2003	2003
Defined Benefit	12,265	11,241	11,990
Defined Contribution	58,338	56,328	59,479
Total	70,603	67,569	71,469

Source: Authors' calculations from raw universe 5500 data files.

Endnotes

1 Buessing and Soto (2006).

2 For instructions and more details on the main Form and schedules see: http://www.dol.gov/ebsa/5500main.html.

3 More precisely, the resulting data is EIN-level data because it is possible that some firms are using multiple EINs (see Decressin, Lane, McCue, and Stinson 2003).

4 A string search of "CASH BAL" in the plan name identifies a few cash balance plans for years before 1999.

5 For example, a firm that offers both a DB and a DC plan to 1,000 of its employees could generate a participant count of 2,000, even if the firm only has 1,000 workers.

6 Comparing defined contribution plans that are in both 1990 and 1993 serves to validate the assumption that the type of defined contribution plan remains unchanged between 1990 and 1992. In 1993, more than 93 percent of the plans have the same classification that they had in 1990.

References

Buessing, Marric and Mauricio Soto. 2006. "The State of Private Pensions: Current 5500 Data." *Issue in Brief* # 42. Chestnut Hill, MA: Center for Retirement Research at Boston College. [Available at: http://www.bc.edu/crr/issues/ib_42.pdf].

Decressin, Anja, Julia Lane, Kristin McCue, and Martha Stinson. 2003. "Employer-Provided Benefit Plans, Workforce Composition, and Firm Outcomes." *Technical Paper No. TP-2003-06*. Washington DC: U.S. Census Bureau. [Available at: http://lehd.dsd.census.gov/led/library/techpapers/tp-2003-06.pdf].

U.S. Department of Labor, Employee Benefits Security Administration, Office of Policy and Research. 2003. "Abstract of 1998 Form 5500 Annual Reports." Private Pension Plan Bulletin 11. Washington, DC: U.S. Government Printing Office. [Available at: http://www.dol.gov/ebsa/PDF/1998pensionplanbulletin.pdf].

U.S. Department of Labor, Employee Benefits Security Administration, Office of Policy and Research. 2004. "Abstract of 1999 Form 5500 Annual Reports." *Private Pension Plan Bulletin 12*. Washington DC: U.S. Government Printing Office. [Available at: http://www.dol.gov/ebsa/PDF/1999pensionplanbulletin.pdf].