A. APPENDIX

SAMPLING

Table A1 shows the way in which the population of self-insured employers was divided into strata, and the number sampled in each strata. We sampled 150 private self-insured employers. We wanted the database to be a sample of claims and not of firms, which implied that larger firms would be more likely to be selected. At the same time, we wanted to have representation from all firm sizes. Therefore, we stratified by firm size, as shown in Table A1. In addition, we stratified by whether the employer had switched third-party administrator, on the assumption that we would be more likely to be able to receive data from employers who had continuously been with the same TPA.

Table A1

Sampling Plan for Self-Insured Private Employers

Changed TPA	Number of Cases	Number of Employers	Sample Size
No	0 - 99	38	3
	100 - 199	27	5
	200 - 299	24	7
	300 - 499	17	8
	500 - 699	21	14
	700 - 999	18	16
	1000+	72	72
Total		217	125
Yes	0 - 99	13	1
	100 - 199	22	1
	200 - 299	22	1
	300 - 499	39	1
	500 - 699	29	1
	700 - 999	35	2
	1000-1499	24	2
	1500-1999	9	1
	2000-2999	21	3
	3000-3999	11	2
	4000-5999	14	4
	6000+	10	6
Total		249	25

DATA CLEANING AND CHECKING

We received our data from the self-insured from June through August of 1998, with most of the data received during June and July. In several cases, problems were immediately identified at RAND when the data arrived, and the employers or the TPA sent new files. During this period, a RAND staff member fielded daily calls from employers and TPAs with questions, logged the arrival of incoming data, and organized the files on a secure computer so that the programmer could begin to process them.

The data files came on many different media, including email attachments, on floppy disks, and on tapes. Many were in excel format. A few were in dBase format. The others were text format in columns, comma delimited, or in multi-line report format. Many came without documentation or column headings. The data had to be converted into a cohesive analysis file.

Each file type had to be handled differently when converting the file. Excel and dBase files were converted using special software. The conversions needed to be checked and the variable names and formats needed to be made consistent. The data that came without headings or documentation were particularly problematic. Many included additional variables, or were missing requested variables, and there was no automatic way to determine the identity of missing or additional variables. The programmer examined each variable in these files, looking for patterns across variables, and called the data providers whenever uncertainty remained.

Here's a sampling of the types of problems faced:

- Dates in different formats (e.g. 03/95, 03-15-95, 03-15-1995, 950315).
- SSN in different formats (e.g. 123456789, character or numeric, 123-45-6789).
- Dollar amounts in different formats (e.g. 123456.78, 123,456.78, $123,456.78)
- Use of character symbols for missing values

In each case the programmer would need to identify the approach adopted by the data provider and write code to convert the variable to a format that is consistent with the other files.

The dollar variables presented additional problems. Not only do they need to be in a consistent format, but also we must be sure the definitions are comparable across the different files. While the programmer was processing the data, other staff members called the employers or TPAs to verify whether they sent us incurred amounts, paid amounts, and any other dollar fields such as salary continuance. We must emphasize how helpful and cooperative the self-insured community was during this process providing RAND with explanations and corrections as needed.

CHECKING DATA QUALITY AND CONSTRUCTING ANALYSIS FILE

The second task was to examine the dollar variables in each of these files more closely. Data quality checks that are performed by the WCIRB need to be repeated. For instance, if a field of data is reported, are there numbers in the field? Are the numbers sensible? Do paid amounts equal incurred amounts for closed claims? Inconsistencies in data definitions across entities would not be surprising, but these inconsistencies need to be identified to avoid misleading results. In addition, inevitably some of the data providers make mistakes when they created files to send to RAND. This is the nature of administrative data. While this task is performed with every new data set that is analyzed, the process was more arduous in this case because there were 167 different data files (including data from the public self-insured, not included in this report). The programmer has recontacted many of these entities/TPAs to verify what the dollar variables really mean.

In total, we received 103,416 claims (all indemnity, and some medical-only) from 68 companies representing 80,229 persons. The analysis sample was to be permanent partial disability claims (PPD) only, with matched wage records. The following steps were taken to reach the ultimate analysis sample:

248 accident records have SSN < 001000000: deleted.

2,248 accidents are before 1991 or after 1996: deleted.

21,704 accidents have no indemnity: deleted (medical-only claims)

1,701 claims have no wage records: deleted

394 claims have no wages prior to accident: deleted

357 claims have no wages in quarter of injury or 4 quarters prior: deleted

88 claims have 3+ names in one quarter on wage file[1]: deleted

993 claims match have two claims, same employer, same quarter: combined (indemnity added)

44,401 person-quarter records not ppd: deleted

137 person-quarter records no wages 8 prior quarters: deleted

335 person-quarter records do not match any employer account number: deleted

30,774 person-quarter claims remain on analysis file

We then decided to focus on second quarter 1991 (due to data problems with the first quarter from EDD) through fourth quarter 1995. We also dropped 1,449 claims for the same

[1] These claims are likely to be from Social Security numbers that are in error and therefore match to more than one individual with wages reported by employers in California.

person in later quarters.[2] After dropping 1996 and 1991-1 claims as well as later claims, 23,171 observations remained. We then dropped 343 with wages in the quarter of injury but for whom the self-insured employer was not the main (highest wage) employer, 696 with indemnity=0, 55 with wages greater than $100,000 in one quarter or more, and 255 without controls. This left the final data set of 21,852.

SELECTING CONTROLS FROM EDD DATA

The self-insured entities selected for our sample are identified by Federal Employer Identification Number (FEIN). Wage records from EDD are based on employer account number (EAN). FEIN is often a higher level identifier than EAN. For example, a holding company may have one FEIN, but each subsidiary has its own EAN. We needed wage records for the EANs where the injureds worked in order to select controls. After we received the wage records for the injured persons, it was possible to construct a crosswalk between FEIN and EAN.

To get the cleanest mapping between FEIN and EAN, we dropped all persons who had wage records at more than one company in the quarter of injury. This left us with a file containing one record with FEIN and EAN for most claims. For each FEIN, we checked on the frequency of EANs for each FEIN, and wrote one record per FEIN-EAN containing the number of claims, percent of claims, and cumulative percent of claims. In the example below (see Table A2), the TAXID 123456789 is the same as EAN 111111. Many TAXIDs have more than one EAN. After studying the output, we devised the following rules to select the records for the crosswalk: Keep record if it is the first record in the set, or CUMPCT < 99, or COUNT > 2. We dropped 2 companies that had a small number of accidents and a large number of EANS. We did not feel confident that we could identify a useful crosswalk for these companies.

We then sent the resulting EANs to EDD who provided us with de-identified data on all workers at these EANs.

---

[2] Essentially, we assume that permanent disability claims for the same person are independent events. This is a strong assumption, but given data limitations (only observing later claims when employees were retained by their employer, only observing later claims within the time period 91-95, and not observing claims for either injured workers or controls at other employers, we adopted this choice. We also note that in some cases, we suspected that subsequent claims were not new claims but rather corrections of old claims without deleting the old record (since for instance, the total benefit amounts were the same). As a check, we estimated earnings losses on the sample without subsequent claims and found that earnings losses were the same.

Table A2

Example of FEIN-EAN Crosswalk

FEIN = 123456789 CONAME = COMPANY ONE

EAN	COUNT	PERCENT	CUMPCT	EMPNAME1
111111	171	97.7143	97.714	COMPANY 1
222222	2	1.1429	98.857	LA COMPANY
333333	1	0.5714	99.429	SF COMPANY
444444	1	0.5714	100.000	SD COMPANY

RESPONSE RATE

Table A3 reports a regression of a dummy variable for whether the firm is one of the 68 included firms. The regression is weighted by the inverse of the sampling probability for the firm (constructed from table A1). The standard errors are heteroskedasticity consistent.

CONSTRUCTION OF NONRESPONSE WEIGHTS

Using a logistic regression with the same specification as in table A3, we obtained predicted probabilities of response (phat), and combined them into five response probability bins.

Bin1: 0<=phat<0.17038

Bin2: 0.17038<=phat<0.34091

Bin3: 0.34091<=phat<0.55689

Bin4: 0.55689<=phat<0.71718

Bin5: phat<=0.71718

Within each bin, we calculated Wr = Sum(sampling weights for respondents) and Wnr = Sum(sampling weights for nonrespondents). We therefore defined the nonresponse weight as (Wr + Wnr)/Wnr. See Little and Rubin (1987).

NUMBER OF CONTROLS PER INJURED WORKER

Table A4 reports the number of controls per injured worker. Table A5 reports the number of controls per injured worker for the self-insured sample. The large firms from which the self-insured sample was drawn allowed for a far higher probability of controls for each injured worker.

CONSTRUCTION OF TOTAL INDEMNITY FOR PERIODS LESS THAN 10 YEARS.

When constructing replacement rates, most of the data received from the self-insured was incurred (*i.e.* including predicted future) amounts reported at about five years. For replacement

rates of three, four, and five years, we did not want to count indemnity not yet paid. Even in the paid data, we did not want to count the full amount of future indemnity included in settlements but preferred to spread it out as though it was paid out according to the schedule. To do this, we simulated the stream of benefits using the WCIRB data to three, four and five years. For all individuals with benefits still being paid according to the simulation at three, four, and five years, we calculated the total benefit paid at that point. We then capped the total benefits in both the self-insured and the insured data at the average of the amount received in that time period for all those still receiving benefits at the end of the period.

Table A3

Response Rates Regression for Construction of Weights

Variable	Parameter Estimate	Standard Error	Adjusted T-stat
Intercept	0.744	1.416	0.526
Cases (in logs)	-0.060	0.027	-2.195
Cases (in logs) Squared	-0.062	0.023	-2.687
Number of Employees (in logs)	0.065	0.030	2.196
Number of Employees (in logs) Squared	0.063	0.028	2.273
Number of Administrative Changes	-0.095	0.085	-1.108
SIC 0 Agriculture, Forestry and Fishing	0.051	0.287	0.179
SIC1 Mining and Construction	0.017	0.403	0.043
SIC 2 Manufacturing	0.166	0.145	1.143
SIC 3 Manufacturing	-0.089	0.142	-0.625
SIC 4 Transportation	-0.213	0.162	-1.309
SIC 4 Communication, Gas, Electric, Water	0.722	0.140	5.151
SIC 6-7 Banks, Insurance, Hotels, Entertainment	-0.097	0.138	-0.706
SIC 8 Health Care Services	0.201	0.114	1.756
Self-administered	-0.312	0.096	-3.246
Combination Administered	0.150	0.152	0.990
Southern California Headquarters	0.190	0.091	2.095
Outside California Headquarters	0.398	0.101	3.947
Payroll per Employee (in thousands)	0.003	3.820	0.777
Total Indemnity per Employee (in thousands)	-0.009	3.629	-0.247

R-squared = 0.3449
Omitted SIC category 5: retail, wholesale trade

The simulation proceeded as follows. We assume that TTD benefits commence during the quarter of injury, followed by VRMA benefits. The WCIRB data do not report the duration of either temporary total benefits or vocational rehabilitation benefits. We calculate a weekly benefit for each using the average weekly wage reported in the WCIRB data. We derive the number of weeks of benefits of TD from the weekly benefit and the total TD benefits incurred. Similarly, we calculate the number of weeks of benefits of VRMA from the average wage and the total VRMA incurred. When both VRMA and TTD are exhausted, we assume that the payment

of permanent partial disability benefits begins. We use the last observed WCIRB disability rating and the WCIRB average weekly wage to derive the weekly benefits paid and the number of weeks of benefits using the benefit schedule.

Table A4

Number of Controls Per Injured Worker, Insured, 1991-1995

	Year of Injury				
Number of Controls	1991	1992	1993	1994	1995
1	1515	1291	1736	2010	673
2	1201	1162	1399	1504	586
3	1070	999	1309	1364	458
4	847	796	1148	1091	388
5	617	674	844	922	267
6	496	467	639	683	205
7	313	314	439	478	130
8	153	191	254	267	74
9	80	93	118	129	35
10	29	23	51	55	9
Total	6321	6010	7937	8503	2825

For example, the maximum for 1993 benefits at three years is 29,242.32, and 26.6% of claims are capped at this amount. At four years, it is 38,046.23, and 19.9% of claims are capped. At five years, it is 46,597.38, and 15.2% are capped.

Table A5

Number of Controls Per Injured Worker, Self-Insured, 1991-1995

	Year of Injury				
Number of Controls	1991	1992	1993	1994	1995
1	30	70	64	74	81
2	34	88	89	92	75
3	43	76	79	96	80
4	55	112	98	81	70
5	3875	4376	4056	4182	3876
Total	4037	4722	4386	4525	4182

TAX SIMULATION

We simulated taxes given earnings for every individual in the sample. However, we do not have the information necessary to actually calculate taxes, including marital status, number of dependents, nonlabor income, etc. We used the following approach:

From the Congressional Budget Office (1998), we obtained the information on average tax rates in table A6. This is tax rates after all deductions, and including federal income taxes and

social insurance (Social Security, Medicare), which we converted to quarterly amounts. From Ettlinger et al (1996), we obtained the California average income tax information in table A6, after deductions (and adjusted to account for the federal income tax deduction for California state taxes), calculated using California income tax data.

Table A6

Average Tax Rates

Quarterly Wages	Federal Tax Rate
0- 2500	.031
2500-5000	.08
5000-7500	.136
7500-10000	.169
10000-18750	.212
18750-25000	.231
25000-50000	.236
50,000+	.260
	CA State Tax Rate
0-5750	.001
5750-10000	.004
10000-14250	.012
14250-20000	.016
20000-36500	.022
36500-93500	.03
93500+	.047

Since these data refer to household income (for income tax purposes), we also converted individual income into household income using the March 1996 Current Population Survey (CPS). Using data on the civilian adult population in California, aged 16-65, we regressed total family income on individual income using a spline with five nodes, weighted by hours worked last year to insure that a working population is more heavily weighted in the regression. We then predicted family income from individual income. For each family income then, we used the tax information to estimate taxes paid, and this information is used to construct after-tax replacement rates.

Figure A1 shows the average tax used in the calculations for every level of quarterly earnings. The higher line illustrates the impact of the adjustment to tax rates to reflect the increased probability that low-wage workers have other income in the household. The lower line illustrates the tax rates that result from the data in table A6. The higher line, using imputed family earnings to estimate tax rates, represents the approach used in the report.

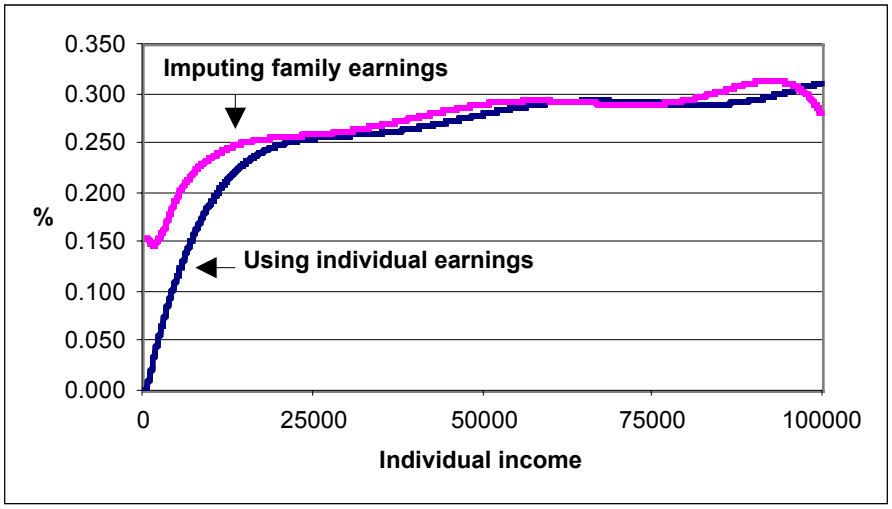

Figure A1— Estimated Average Tax Rates

INDEMNITY QUINTILE AND PERMANENT DISABILITY RATING QUINTILE

For 1993 injuries, the correlation between disability rating and total indemnity incurred in the WCIRB data was 0.68. In Table A7, we report a cross-tabulation between quintile of disability rating and the quintile of indemnity.

Table A7

Indemnity Quintile by Permanent Disability Rating Quintile, Insured 1993

		Rating				
Total Indemnity		1-6	7-12	13-20	21-32	32+
1-3558	number	1078	432	61	15	1
	(proportion)	(67.93)	(27.22)	(3.84)	(0.95)	(0.06)
3559-8255	number	291	812	435	47	2
	(proportion)	(18.34)	(51.17)	(27.41)	(2. 96)	(0.13)
8256-17352	number	73	256	708	537	13
	(proportion)	(4.60)	(16.13)	(44.61)	(33.84)	(0.82)
17353-33792	number	29	90	297	801	370
	(proportion)	(1.83)	(5.67)	(18.71)	(50.47)	(23.31)
33792+	number	4	20	66	296	1202
	(proportion)	(0.25)	(1.26)	(4.16)	(18.64)	(75.69)

PROPORTIONAL WAGE LOSS REGRESSIONS

We estimated descriptive regressions of individual proportional wage loss on a dummy for insured, and other characteristics in Table A8. We multiplied both sides of the equation by potential uninjured earnings, which means that we regressed total individual (12-quarter) losses on potential uninjured earnings, and potential uninjured earnings interacted with all the other

Table A8

Proportional Wage Loss Regressions

	1994 Injuries, 12 Quarters After Injury			1993-1995 Injuries, 12 Quarters After Injury		
Variable	1	2	3	1	2	3
Potential uninj. earnings	0.211*	0.358*	0.604*	0.210*	0.342*	0.567*
(Control's 3-yr earnings)	(0.005)	(0.021)	(0.025)	(0.003)	(0.149)	(0.017)
Insured	0.111*	0.048*	0.006	0.133*	0.084*	0.051*
	(0.007)	(0.011)	(0.014)	(0.005)	(0.007)	(0.009)
Log Preinjury quarterly			-0.120*			-0.112*
earnings			(0.007)			(0.004)
Log Number of employees			-0.011*			-0.009*
			(0.003)			(0.002)
Industry (SIC Code)						
Agriculture, Forestry and		0.045	0.029		0.023	0.001
Fishing (SIC-0)		(0.035)	(0.035)		(0.231)	(0.023)
Mining and Construction		0.006	0.035		0.003	0.027*
(SIC-1)		(0.330)	(0.187)		(0.013)	(0.012)
Manufacturing (SIC-2)		-0.047*	-0.037		-0.065*	-0.052*
		(0.019)	(0.019)		(0.012)	(0.012)
Manufacturing (SIC-3)		-0.105*	-0.058*		-0.111*	-0.065*
		(0.015)	(0.016)		(0.010)	(0.010)
Transportation (SIC-4)		-0.029	0.002		-0.063*	-0.027*
		(0.022)	(0.104)		(0.013)	(0.013)
Communication, Power,		-0.124*	-0.063*		-0.118*	-0.056*
Water (SIC-4)		(0.015)	(0.015)		(0.009)	(0.009)
Financial, Hotels, Enter-		-0.041*	-0.010		-0.043*	-0.011
tainment (SIC-6-7)		(0.016)	(0.016)		(0.010)	(0.011)
Health Care Services		-0.033*	0.001		-0.027*	0.003
(SIC-8)		(0.014)	(0.045)		(0.009)	(0.009)
Location of At-injury Employer						
Bay area		-0.077*	-0.050*		-0.059*	-0.030*
		(0.017)	(0.017)		(0.011)	(0.011)
Southern California		-0.037*	-0.024		-0.019	-0.004
		(0.016)	(0.015)		(0.011)	(0.010)
Multiple locations		-0.054*	-0.011		-0.027*	0.014
		(0.015)	0.017)		(0.010)	(0.011)

Note: Dependent Variable: Total 3-year losses. Regressions include dummy variables for quarter of injury. Omitted SIC category: retail and wholesale trade (SIC-5). All variables (except Potential earnings) are multiplied by potential earnings to obtain impact of variable on proportional earnings loss. The 1994 regression has 12824 observations. The 1993-1995 regression has 31,948 observations.

variables in the regression. The table shows the estimates for 1994 injuries at three years, and for 1993-95 injuries pooled, also at three years.

The regression estimates only proportional wage loss, with no controls in column 1. In column 2, controls for industry are added, In column 3, controls for the log preinjury quarterly earnings and log of the number of employees are added. For the 1994 estimates, inclusion of

Table A9

Proportional Wage Loss Regressions, 1993, 1995

	1993 Injuries, 12 Quarters After Injury			1995 Injuries, 12 Quarters After Injury		
Variable	1A	2A	3A	1B	2B	3B
Potential uninj. earnings	0.215*	0.296*	0.463*	0.191*	0.317*	0.577*
(Control's 3-yr earnings)	(0.05)	(0.021)	(0.026)	(0.005)	(0.032)	(0.037)
Insured	0.144*	0.115*	0.093	0.144*	0.103*	0.057*
	(0.007)	(0.011)	(0.013)	(0.011)	(0.154)	(0.021)
Log Preinjury quarterly			-0.084*			-0.137*
earnings			(0.007)			(0.010)
Log Number of employees			-0.006*			-0.014*
			(0.003)			(0.005)
Industry (SIC Code)						
Agriculture, Forestry and		0.057	0.032		-0.115	-0.136
Fishing (SIC-0)		(0.035)	(0.035)		(0.062)	(0.061)
Mining and Construction		-0.074*	-0.030*		0.003	0.027*
(SIC-1)		(0.028)	(0.028)		(0.032)	(0.032)
Manufacturing (SIC-2)		-0.032	-0.005		-0.124*	-0.092*
		(0.028)	(0.028)		(0.031)	(0.030)
Manufacturing (SIC-3)		-0.014	0.048		-0.165*	-0.090*
		(0.027)	(0.028)		(0.021)	(0.022)
Transportation (SIC-4)		-0.091*	-0.061		-0.043*	0.003*
		(0.197)	(0.020)		(0.031)	(0.031)
Communication, Power,		-0.083*	-0.038*		-0.154*	-0.063*
Water (SIC-4)		(0.015)	(0.016)		(0.019)	(0.020)
Financial, Hotels, Enter-		-0.050*	-0.028		-0.032*	0.019
tainment (SIC-6-7)		(0.163)	(0.017)		(0.024)	(0.024)
Health Care Services		-0.024*	-0.006		-0.026*	0.017
(SIC-8)		(0.015)	(0.015)		(0.020)	(0.020)
Location of At-injury Employer						
Bay area		-0.031*	-0.007*			
		(0.017)	(0.017)			
Southern California		0.006*	0.018			
		(0.016)	(0.016)			
Multiple locations		-0.008*	0.023			
		(0.016)	0.017)			

Note: Dependent Variable: Total 3-yr losses. Regressions include dummy variables for quarter of injury. Omitted SIC category: retail and wholesale trade (SIC-5). All variables (except Potential earnings) are multiplied by potential earnings to obtain impact of variable on proportional earnings loss. The 1994 regression has 12824 observations, and the 1995 regression has 6942 observations.

preinjury earnings and the log of the number of employees renders the insured dummy insignificant. This suggests that the differences in proportional wage losses, in 1994, between insured and self-insured claimants, can be explained by preinjury earnings and the number of employees at the firm. For 1993-95 pooled claimants, the insured variable remains significant and positive but falls by more than half. This estimate implies that proportional wage losses are

higher at insured firms, even after controlling for industry, preinjury earnings, number of employees, and the part of the state. Table A9 reports the results for 1993 and 1995.