6. Survey versus Administrative Data: Analysis by Account-type and Regression Analysis

1.   Comparison by asset types

In this document we compare survey and administrative data for subsets of wealth–IRA, employer sponsored plan and taxable assets.[1] All comparisons are made conditioning on survey-singles. Table 6-1 shows results for IRAs. Survey and administrative data line up very well. The median difference is very small, while the interquartile range does not go over ±5%.

One source of error is misreports of whether the asset is at Vanguard. Note that the comparison in Panel A of Table 6-1 is done conditioning on that the respondent has IRA either in survey or administrative data. In the following two panels, we show the cross-tabulations of indicators of having Vanguard IRA wealth in the survey and the administrative data. The correlation between the two indicators is very high – if they do (not) have IRA wealth in administrative data, with more than 90% likelihood, in the survey they answered they do (not) have IRA wealth.   But the correlation is not perfect, so some of the discrepancy arises from misclassification of assets at Vanguard or not.

Table 6-2 shows the same comparisons for employer-sponsored plan, while Table 6-3 is on taxable assets. All the patterns are fairly similar to what we found with IRA, though these items turn out to be somewhat noisier.

2.   Comparison conditioning on other variables.

---

[1] Since some of the administrative data were pulled after the survey, we do not have administrative data on respondents who died or otherwise left Vanguard between the administration of survey 1 and when this analysis commenced.   As a result, we do not have administrative data on 58 respondents who took Survey 1. Going forward, the VRI will have administrative data up to the point that a client leaves Vanguard for any reason.

The administrative data we used so far in the comparison is the snapshot of the account holder's portfolio at the beginning of months, in addition to basic demographic information available in the administrative records. The administrative data contains other information useful for analyzing survey responses. We highlight several measures—summarized in Table 6-4— that are useful for understanding corrections as well as having independent interest.

Vanguard tracks usage of its Web site. From these data, we can determine about 6% of respondents logged in while taking the survey.

The data also contain the start date of their web registration (potentially earlier than 2003), which can be used as a proxy for their length of relationship with Vanguard. The median length of relationship is 10 years.

The administrative data also allow us to track volatility of balances. The interquartile range of mean monthly percentage change is from 2.6% to 5.1%.[2]

These variables might be related to discrepancies between the survey and administrative data. Table 6-5 reports the conditional distributions of the difference between the wealth measures of the survey and the administrative data (singles only). Logging into the account during the survey reduces the interquartile range for the individual client sample, but the number of respondents who logged in is too small for a meaningful comparison. The interquartile range is smaller when they have been web-registered for more than 10 years. Finally, the difference tends to be larger for the respondents whose historical balance has been more volatile. The interquartile range is about 50% wider for the group whose historical volatility has been larger

---

[2] The volatility of historical data is calculated in the following way. First, the percentage difference of total balances between any two consecutive months is calculated using a formula very similar to the one explained in footnote 8 of the main paper. The only difference is that the numerator is now calculated as absolute value difference, since we are interested in the magnitude of changes in balance not the direction here. Then the historical average percentage difference is calculated as the mean of these monthly percentage differences.

than the median. There are two possible explanations for this. On the one hand, respondents with volatile balances might find it harder to keep track of their financial records, and as a result their survey response error can be larger. On the other hand, when the volatility measured as monthly changes in the balance is high, it is also likely that their intra-month volatility is also high. If that is the case, the difference between the survey measure, which is measured at the moment of the survey, and the administrative measure, which is a snap shot at the end of months, can be accounted for by the timing issue.

3.   Regression analysis of deviation of survey and administrative data

Table 6-6 shows the results from the regression analyses where the dependent variable is the absolute percentage value of discrepancy between the survey and the administrative wealth measures and the explanatory variables capture all the variables we considered so far. The discrepancies are smaller if they are younger than 75, if they checked for the record for majority (more than 80%) of their accounts, if they logged in during the survey, if their history of balance at Vanguard has been less volatile, and if they had longer relationship with Vanguard.

Table 6-1. Survey versus Administrative Data: IRA

A. Balances (Singles)

	Mean	Percentiles				
		10	25	50	75	90
Survey	202,498	3,000	32,078	101,272	269,000	519,000
Administrative	214,047	16,422	40,989	114,408	277,162	519,706
Difference	-11,549	-51,640	-4,918	-116	1,866	17,809
% Difference	-14.60%	-171.66%	-4.41%	-0.13%	1.44%	18.03%

Note: Figures on this table condition on having positive holding in either survey or administrative data. 1,914 observations.

B. Have or Not (Singles, Employer-sponsored)

		Survey measure		N
		Have	Do not have	
Administrative	Have	NA	NA	0
Measure	Do not have	4.62%	95.38%	585

C. Have or Not (Singles, Individual client)

		Survey measure		N
		Have	Do not have	
Administrative	Have	90.70%	9.30%	1,860
Measure	Do not have	5.32%	94.68%	489

Table 6-2. Survey versus Administrative Data: Employer-sponsored plan

A. Balances (Singles)

	Mean	Percentiles				
		10	25	50	75	90
Survey	214,663	12,000	35,000	111,000	275,000	518,581
Administrative	195,204	0	24,402	91,425	247,532	500,999
Difference	19,460	-41,710	-5,276	476	11,018	98,652
% Difference	15.17%	-63.77%	-3.70%	0.44%	12.19%	200.00%

Note: Figures on this table condition on having positive holding in either survey or administrative data. 760 observations.

B. Have or Not (Singles, Employer-sponsored)

		Survey measure		N
		have	do not have	
Administrative	Have	91.42%	8.58%	583
Measure	do not have	0%	100%	3

C. Have or Not (Singles, Individual client)

		Survey measure		N
		have	do not have	
Administrative	Have	88.52%	11.48%	61
Measure	do not have	5.07%	94.93%	2,288

Table 6-3. Survey versus Administrative Data: Taxable assets

A. Balances (Singles)

	Mean	Percentiles				
		10	25	50	75	90
Survey	189,852	1,900	20,000	74,000	229,770	500,000
Administrative	147,860	0	11,971	55,333	179,005	417,028
Difference	41,991	-27,334	-2,623	94	13,800	130,508
% Difference	16.31%	-96.17%	-3.70%	0.23%	25.74%	200.00%

Note: Figures on this table condition on having positive holding in either survey or administrative data. 1,629 observations.

B. Have or Not (Singles, Employer-sponsored)

		Survey measure		N
		Have	Do not have	
Administrative	Have	NA	NA	0
Measure	Do not have	2.90%	97.10%	586

C. Have or Not (Singles, Individual client)

		Survey measure		N
		Have	Do not have	
Administrative	Have	91.44%	8.56%	1,414
measure	Do not have	21.07%	78.93%	935

Table 6-4. Other Administrative Variables: Summary Statistics

Variable	Fraction	25 percentile	median	75 percentile
Logged in during survey	5.77%			
Length of relationship (years)		5	10	13
Mean volatility of balance history		2.59%	3.53%	5.08%

Table 6-5. Effects of logging in during survey, length of relationship and volatility of historical balances: Percentage deviation, Survey versus administrative wealth. (Singles only)

A. Logged in during survey

Sample	Logged in	N	Mean	Percentiles				
				10	25	50	75	90
Employer-sponsored	Yes	4	33.79%	0.27%	0.51%	2.53%	67.07%	129.84%
	No	581	1.83%	-22.06%	-3.07%	0.33%	6.21%	29.56%
Individual Client	Yes	177	4.91%	-9.75%	-0.64%	0.08%	1.52%	30.08%
	No	2,172	-1.10%	-23.93%	-3.26%	-0.04%	2.20%	23.49%

B. Length of relationship

Sample	Length of relationship	N	Mean	Percentiles				
				10	25	50	75	90
Employer-sponsored	< 5 years	188	0.73%	-37.64%	-5.60%	-0.10%	5.97%	66.33%
	5-10 years	230	2.96%	-22.80%	-3.51%	0.41%	7.49%	31.71%
	> 10 years	167	2.28%	-6.64%	-0.82%	0.79%	5.25%	18.75%
Individual Client	< 5 years	524	-0.86%	-27.40%	-3.50%	-0.08%	2.36%	32.06%
	5-10 years	724	-2.47%	-26.09%	-3.73%	-0.05%	2.10%	20.45%
	> 10 years	1,101	2.60%	-16.09%	-2.30%	0.01%	2.02%	24.38%

C. Volatility of historical balances

Sample	Volatility	N	Mean	Percentiles				
				10	25	50	75	90
Employer-sponsored	< 3.5%	250	1.61%	-14.61%	-2.00%	0.28%	4.70%	26.91%
	> 3.5%	335	2.38%	-27.26%	-3.69%	0.44%	7.49%	37.56%
Individual Client	< 3.5%	1,188	-2.76%	-20.39%	-2.50%	-0.09%	1.52%	14.62%
	> 3.5%	1,161	1.52%	-23.87%	-3.52%	0.05%	2.87%	31.36%

Table 6-6. Regression analysis of survey – administrative wealth measure discrepancy (Singles only)

Variable/Category	coefficient	Standard error
Intercept	0.215	0.045
Male	-0.005	0.017
Employer-sponsored sample	0.009	0.023
Age (55-64)	-0.067	0.024
Age (65-74)	-0.082	0.023
Logged in during survey	-0.052	0.035
Checked records (> 80% of accounts)	-0.032	0.017
Number of accounts	0.004	0.002
Length of relationship (<5yrs)	0.062	0.021
Length of relationship (5-10yrs)	0.016	0.020
Volatility of historical balance (>3.5%)	0.050	0.017

Note: N = 8,948. Dependent variable is absolute value of percentage difference defined in footnote 10. Omitted categories for independent variables are individual client sector, the oldest, married, female, checked for record for less than 80% of accounts, did not log in during the survey, and having relationship with Vanguard for more than 10 years.