CORPORATION LIFE CYCLES: EXAMINING ATTRITION TRENDS AND RETURN CHARACTERISTICS IN STATISTICS OF INCOME CROSS-SECTIONAL 1120 SAMPLES

Matthew L. Scoffic, Internal Revenue Service

EVERY YEAR THE STATISTICS OF INCOME (SOI) Division of the IRS produces a cross-sectional study of 1120 series corporation tax returns based on a weighted sample of the population of certain Forms 1120. These data are used by the Department of the Treasury and others to study tax policy and tax administration issues. Aggregate tabulations of the data are released to the public.

While these data provide an excellent source for annual financial tabulations and for developing an understanding of the implications of tax policy for the taxpaying public, there is less focus on the implicit longitudinal characteristics of the SOI sample or the changing population of 1120 filers from which SOI draws its sample. This paper examines the extent to which business entities in the SOI sample survive, perish, or appear inconsistently, and to what extent returns from these three categories differ in certain financial characteristics. Examining these issues can provide insight into what types of business entities tend to survive and perish over a period of time and can provide users of SOI tabular data with insight into whether estimates are based on the same entities over time, or a sample that changes with regularity.

THE SOI 1120 SAMPLE

Before examining the performance of the SOI sample over a period of years, it is first useful to understand the structure of the cross-sectional SOI sample itself. The SOI study's target population consists of all for-profit corporations that are required to file an 1120 series tax return that is included in the SOI study. SOI studies Forms 1120, 1120-A, 1120-S, 1120-L, 1120-RIC, 1120-REIT, 1120-PC, and 1120-F. The survey population consists of those returns that are selected for the SOI sample and are processed on IRS computer systems. SOI has been using a sample of 1120 series returns to estimate population values for over 50 years. The first SOI sample was implemented for Tax Year 1951, when 41.5 percent of the 1120 filing population was sampled. In 1951 the total number of Forms 1120 filed was 687,000 and SOI selected 285,000 returns for its study. The sample size as a percentage of the population has fluctuated over time, and in the last Tax Year for which data are available, 2003, the SOI sample was 2.4 percent of the total population of over 5.8 million 1120 returns, or 141,678 returns. In the 10 years that are the focus of this paper, the SOI sample size has increased from 91,687 returns in 1993 to 141,678 returns in 2003 (see Figure 1).

The sample is stratified by form type, size of total assets, and income, or in some cases form type and size of total assets alone. Returns in different strata are sampled at different rates, ranging from a fraction of 1 percent to 100 percent. Generally, the sampling rates increase as size increase for all form types. Over the 10 years studied, sampling rates have tended to increase for most size classes and form types, but rates for some strata have declined.[1]

To determine whether an individual return is to be sampled, an algorithm is used to transform the Employer Identification Number (EIN) of the tax return to produce a Transform Taxpayer Identification Number (TTIN). This TTIN can be characterized as a pseudo-random number; the same algorithm is used to produce the TTIN every year, so the same algorithm applied to the same EIN will produce the same TTIN in any study year. This implies that with no change in the selection probability of the applicable stratum and no change in the stratum into which the return falls, a return selected in year one should be selected in year two, providing it is present in the population (and providing it has not changed its EIN).

Each stratum is associated with a sampling rate. The sampling rate is multiplied by 10,000 to create a 4-digit number between 0000 and 9999. If the last four digits of the TTIN for a given return are less than or equal to this number, the return is selected for the SOI study. If the last four digits of the TTIN are greater than this product, the return is not selected. The rate at which returns are sampled depends on their size (measured in income and/or total assets) and form type.

Figure 1: **Sample and Population Size for SOI 1120 Study 1993-2003**

YEAR	SAMPLE SIZE	POPULATION SIZE	SAMPLE AS PERCENT OF POPULATION
1993	91,687	4,340,688	2.11
1994	95,021	4,700,268	2.02
1995	97,461	4,852,305	2.01
1996	94,172	4,968,490	1.90
1997	98,204	5,102,958	1.92
1998	137,600	5,204,810	2.64
1999	140,984	5,315,461	2.65
2000	144,917	5,429,473	2.67
2001	146,479	5,563,781	2.63
2002	145,353	5,701,024	2.55
2003	141,678	5,845,672	2.42

This selection process takes place over a 24-month window of time. Typically more than 15 percent of corporations file tax returns based on a non-calendar year accounting period. Therefore a selection window of July through the following June is necessary for any given study year. The time necessary is extended further due to optional extensions of the filing deadline which are used by many corporations, and by administrative processing delays on the part of the IRS. A study for Tax Year X is therefore composed of returns selected from July of Year X through June of Year X+2. Some returns can also be added after this time if their presence in the SOI study is deemed critical.[2] Returns that would meet the sampling criteria may not be selected because they have been filed later than SOI's deadline for selection, because the returns were not available to the SOI Division while being held by another IRS function, or because data processing errors caused the returns to fall into an incorrect stratum.[3]

DATA DESCRIPTION

In order to study the behavior of returns in the SOI sample, I compiled 11 years of selected data from SOI's cross-sectional 1120 study, Tax Years 1993 to 2003. To create the data set, I first identified all unique EIN's in the Tax Year 1993 study. There were 86,632 records in this dataset. I used this file as the "base year" to which I compared SOI studies from other years to determine the presence or absence of the base-year returns in subsequent years. For the subsequent 10 years of SOI studies from 1994 through 2003, I compiled 10 data sets containing the EIN's of base-year returns that were selected again in the subsequent years, and 10 data sets containing the EIN's of base-year returns not selected in the subsequent SOI study years.[4]

This allowed me to determine the presence or absence of each base-year EIN for each study year and compiled an inventory data set which represents the life cycles of each base-year EIN throughout the 10 years. This data set contained all EIN's from the base year and an observation for each subsequent study year, 1994 – 2003. The observation could take on a value of "0" if the return was not present in the study year, or "1" if the return was present in the study year. The dataset also contained a data item representing the life cycle of the return. This data item was a concatenation of all the study year observations ("0" or "1") and represented the 10-year pattern of presence or absence for each base-year return.

I then used the inventory data set to group the base-year returns into three categories based on a characterization of their life cycles over the 10 years studied: Consistent, Inconsistent, and Terminal. I defined a Consistent return as one that was present in at least 8 out of the 10 years analyzed

but was not absent from the sample in the last 2 years, 2002 and 2003.[5] I defined an Inconsistent return as one that was present in less than 8 years of SOI studies and was not categorized as a Terminal return. I defined Terminal as returns whose life-cycle pattern matched 1 of 9 specific patterns that indicate the returns left the sample and never returned. Figure 2 shows the patterns used to characterize Terminal returns. A "1" indicates the return is present for the year and a "0" indicates the return is absent. Each of the 10 characters comprising the life-cycle pattern represents a study year, 1994 – 2003.

Because returns can be either present in the SOI study and present in the population, absent from the SOI study and absent from the population, or absent from the SOI study but present in the population, I matched files of base-year returns not present in each subsequent year to administrative IRS population files to examine the ultimate status of the returns.[6] In some cases it could be shown that although base-year returns were missing from the SOI sample for a subsequent year, they were present in the population of 1120 filers. These returns are in general presumed to have not met the SOI selection criteria for the study year, subject to the limitations of the selection process described previously. In other cases, it could be shown that a base-year return not selected for a subsequent SOI study was not selected because it was no longer present in the population of 1120 filers. It is of use to determine which nonselected base-year returns remained in the population and available for selection to demonstrate whether a return has simply failed to meet SOI sampling criteria or is in fact no longer required to file a nonconsolidated 1120 series tax return.[7]

In order to determine whether Consistent, Inconsistent, and Terminal returns differed qualitatively in terms of their financial characteristics or other characteristics, I compiled these three groups of returns and determined the means of four key financial data items and the age of the entity using financial and age data from the base-year returns. I compared the 1993 means of the data items and the ages in each category and tested the differences to determine statistical significance. Differences in the means of these items in the base year may

Figure 2: **Criteria for Terminal Return Definition**

Life-Cycle Patterns Characterizing Terminal Returns
0000000000
1000000000
1100000000
1110000000
1111000000
1111100000
1111110000
1111111000
1111111100

From left to right, each character represents an SOI study year 1994 – 2003.
A "0" indicates absence from the SOI study for the year.
A "1" indicates presence in the SOI study for the year.

indicate that returns with certain characteristics are more likely to survive or perish over time. The four financial items compared were Total Receipts, Net Income, Total Assets, and Net Worth.[8] The age of the entity is the number of years between the date of incorporation and the base year, 1993.[9]

DATA ANALYSIS

Figure 3 presents the count of base-year returns present in each subsequent SOI study and filing population from 1994 – 2003 as well as the percent of base-year returns present in the sample and population in subsequent years. The same data are represented graphically in Figure 4.

In the base year of 1993, 86,632 returns were selected for the SOI study. The number of base-year returns remaining in the SOI study declined steadily over the 10 years analyzed. The number of base-year returns available to be selected from the population declined in a very similar fashion.

The difference in the counts and percentages of base-year returns in the sample and population can be attributed to a number of factors. Returns which exhibit a year-to-year change in total assets and/or income may qualify for a sampling rate different than that applied in a prior year in which the returns were selected for the SOI study. Similarly, a change to the sampling rates for a stratum may cause returns that were selected in that stratum

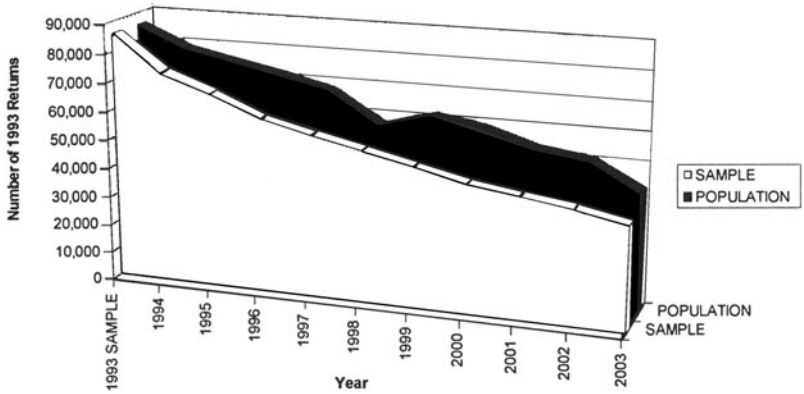
Figure 3: **Presence of Base-Year Returns in SOI Sample and Population**

SOI STUDY YEAR	BASE-YEAR RETURNS IN SAMPLE	BASE-YEAR RETURNS IN POPULATION	BASE-YEAR % IN SAMPLE[1]	BASE-YEAR % IN POPULATION[2]
1993	86,632	86,632	100	100
1994	74,303	79,243	85.8	91.5
1995	68,122	75,965	78.6	87.7
1996	60,948	72,585	70.4	83.8
1997	56,465	68,633	65.2	79.2
1998	52,750	57,734	60.9	66.6
1999	48,842	62,674	56.4	72.3
2000	44,728	59,257	51.6	68.4
2001	42,154	53,743	48.7	62.0
2002	39,998	51,683	46.2	59.7
2003	36,159	42,414	41.7	49.0

[1]Percentage of base-year returns remaining in sample

[2]Percentage of base-year returns remaining in population.

Figure 4: **Presence of Returns from BaseYear**



previously to no longer qualify for sample selection based on the values of their TTIN's. There are other administrative and processing reasons that may prevent a negligible number of returns from being included in the SOI study. These reasons include rejection by tax examiners from the SOI study, improper coding or processing, unavailability of returns, or late filing of desired returns.[10]

Since the difference between the base-year returns present in the sample and population is small and stable throughout the 10-year period, it can be concluded that the majority of returns which leave the SOI study have also left the population of 1120 filers. For example, in 1994, only 5.7 percent (4,940) of base-year returns were absent from the sample but present in the population. In 2003, this percentage had increased to only 7.3 percent (6,255). Although the SOI sample size has increased over the 10-year period studied, sampling rates for various strata have fluctuated. This means that in addition to any base-year returns with changes in total assets and/or income becoming ineligible for sampling at prevailing rates, changes to the sampling rates in individual strata may make previously eligible returns ineligible. This helps explain why the percentage of base-year returns in the population but not the sample has increased slightly over the ten years observed. Since larger returns are sampled at a 100 percent rate, decreases in sampling rates tend to affect strata where smaller returns are located. Decreases in sampling rates help account for a loss of base-year returns, but only if they are still available in the target population. However, since Figures 3 and 4 indicate that the majority of the base-year returns leaving the sample have also left the population, it appears that most of the missing base-year returns have not survived as individual 1120 return filers. They may no longer exist, they may file a non-1120 tax return, or they may be included in the consolidated return of another 1120 filer.

When returns from the base year were grouped into categories based on their life-cycle patterns, 37,614 returns were observed to be consistently present in the SOI study from 1993 – 2003. This category of returns was called "Consistent." The number of Inconsistent Returns totaled only 9,482, showing that a relatively small number of returns appeared sporadically. The Terminal return category contained a total of 39,536 returns (see Figure 5).[11]

A pronounced and statistically significant difference in the means of all the data items was observed among the various categories of returns. Figures 6, 7, and 8 summarize the means of the various categories. The statistical significance of the differences of the means was determined by performing a t-test. The results showed statistical significance above the 99 percent level for comparison of all means across all categories.

The means presented in Figures 6, 7, and 8 clearly show that Consistent returns appear on average to be larger in terms of financial characteristics than either returns that appear in the SOI study only inconsistently or returns that have dropped out of the SOI sample and most likely the population as well. When financial items from Consistent returns

Figure 5: **By Type of Return**

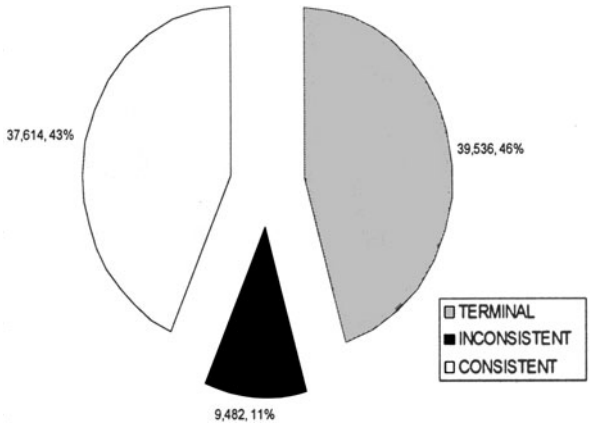


Figure 6: **Consistent Returns**

Variable	N	Mean	Standard Deviation
Total Receipts	37,744	$136,238,155	$1,498,106,574
Net Income	37,744	$8,215,763	$96,288,521
Total Assets	37,744	$304,742,101	$3,776,946,351
Net Worth	37,744	$109,835,169	$902,754,411
Age	37,744	19.4	21.0

Figure 7: **Inconsistent Returns**

Variable	N	Mean	Standard Deviation
Total Receipts	9,459	$25,796,330	$238,476,363
Net Income	9,459	$220,453	$14,196,113
Total Assets	9,459	$37,207,485	$444,127,898
Net Worth	9,459	$6,618,853	$70,868,775
Age	9,459	14.8	16.6

Figure 8: **Terminal Returns**

Variable	N	Mean	Standard Deviation
Total Receipts	39,926	$77,461,225	$814,956,006
Net Income	39,926	$3,222,766	$58,191,247
Total Assets	39,926	$205,827,618	$3,493,116,498
Net Worth	39,926	$43,992,315	$583,865,566
Age	39,926	15.7	19.6

are compared to those of Terminal returns, all items are larger for Consistent returns by significant margins. The largest differences in the averages are between Consistent and Inconsistent returns. Clearly, the returns that are consistently selected for the SOI sample have higher average levels of assets and income. Although this may seem intuitive since larger returns fall into strata with higher sampling rates, in fact, the design of the sample leads to the same returns being selected each year in each stratum. Therefore, barring changes to the sampling rates of the relevant strata, a small base-year return exhibiting no drop in assets or income and no change in form type would be expected in the sample again, as would a large return in a stratum with a 100 percent selection rate. In practice, sampling rates for certain strata have declined at times. Most base-year returns that are not selected are demonstrably not in the population, but for those smaller base-year returns that are in the population and are not selected, sampling rate changes are the major explanation. Other reasons for attrition include the processing limitations discussed previously.

To conduct a more detailed analysis of the three categories of returns, I created another data item called "Size." This data item was determined by the size of total assets of the return. Returns with less than $10,000,000 in total assets were defined as "small," returns with between $10,000,000 and $249,999,999 in total assets "medium," and returns with $250,000,000 or more in total assets "large." I then grouped each of the three "consistency" categories into subgroups of small, medium, and large returns to analyze differences in mean financial characteristics and mean age by both consistency and size.

After segmenting returns as described, it was observed that large returns made up a considerably higher percentage of Consistent returns than they did Inconsistent or Terminal. Conversely, small returns tended to make up a much larger percentage of Inconsistent and Terminal returns, as indicated by Figure 9. The attrition rate was defined as the percentage of returns within each size category-- small, medium, and large-- which was ultimately classified as Terminal. Large returns had the lowest attrition rate, followed by medium-sized returns. Small returns had the highest attrition rate at 49.5 percent. This may partially be due to the fluctuating sampling rates for smaller returns, but since most nonselected returns were also not present in the population, most of these taxpayers did not file individually.[12]

Examining Figure 9 can provide insight as to why the averages of selected financial items tend to be higher for Consistent returns than the other categories. The averages for Consistent returns are based on a higher proportion of large returns than are the other categories. As a function of the definition of large returns, these financial items will tend to be greater on returns with more assets, thus averages based on a higher proportion of large returns will be greater. All means and standard deviations of financial items and ages by consistency and size are reported in the appendix.

In addition to being on average larger in terms of these selected financial items, this comparison indicates that Consistent returns tend to be older than Inconsistent or Terminal returns. Age was defined in years as the base year (1993) minus the year of incorporation. The average age of returns consistently in the SOI study is 19.7 years. The average ages of both Inconsistent and Terminal returns are

Figure 9: **Return Counts by Size and Consistency with Attrition Rate**

	Consistent	Inconsistent	Terminal	Attrition Rate
Small	19,041 (50.4%)	6,959 (73.6%)	25,479 (63.8%)	49.5%
Medium	14,719 (39.0%)	2,322 (24.5%)	11,789 (29.5%)	40.9%
Large	3,984 (10.6%)	178 (1.9%)	2,658 (6.7%)	39.0%

Small returns are those with less than $10,000,000 in assets, Medium with $10,000,000 to $249,999,999 in assets, and Large with $250,000,000 or more.

Percentages following counts indicate the percent of the total count for the group of Consistent, Inconsistent, or Terminal.

Attrition rate is the percentage of the total number of base-year returns in this size category which were categorized as Terminal returns.

lower at 14.6 years, and 15.9 years, respectively. With most of the base-year returns missing from the SOI study also missing from the population of 1120 filers, the analysis indicates that on average, business entities that were older in the base year tended to survive longer.[13] Younger returns were more likely to be Inconsistent or Terminal.

Of particular interest is the difference in mean ages of large Consistent, Inconsistent, and Terminal returns. The mean age of large Consistent returns is 20.6 years, while the mean ages of large Inconsistent and Terminal returns are 22.4 years and 24.8 years, respectively. The difference between large Consistent and large Inconsistent returns is not statistically significant, but the difference between large Consistent and large Terminal returns is significant at the 99 percent level. Although returns of all sizes exhibit higher mean ages for Consistent returns than for Inconsistent or Terminal returns, breakouts by size showed that large Consistent returns were younger on average than large Terminal returns.

CONCLUSIONS AND FURTHER RESEARCH

The analysis showed that the majority of base-year returns that left the SOI sample also left the population of 1120 filers, indicating that the SOI sample largely selects the same entities from year to year when those entities are available in the population. Therefore even though a small number of returns exited the SOI study due to changes in sampling rates, the conclusions drawn from analysis of the SOI studies largely apply to the population of 1120 filers as well as to the sample. After analyzing 10 years of data from SOI samples and 10 years of population data from IRS computer files, 41.7 percent of the base-year returns were shown to be present in the latest SOI study and 49.0 percent of base-year returns present in the filing population. With the lowest attrition rate of all groups, large business entities are more likely than smaller business entities to remain in the SOI sample and in the filing population. The group of returns defined as Consistent exhibited a larger proportion of returns with $250,000,000 or more in total assets than the other two categories of returns and large returns made up the smallest proportion of Terminal returns at 5.5 percent. The surviving business entities also tended to be older on average than business entities that fell out of the population or were not selected for SOI studies. This relationship was not true for the group of large returns however, where Consistent returns were slightly younger on average than Terminal returns.

The next steps in corporation life-cycle research will be to define specific reasons for attrition from the SOI sample and population and more fully explain attrition based on these reasons. This research should include the assembly of corporate family structures capable of accounting for previously individual returns that become part of consolidated groups. A predictive model could be implemented to determine if financial relationships are predictive of presence in the SOI sample or population.

Notes

1. For a complete history of sampling rates for all sizes and form types, see SOI's annual Publication 16, various years.
2. For an explanation of critical returns see SOI's annual Publication 16, various years.
3. For a more detailed description of SOI's sampling process and studies, please see the most recent version of SOI's Publication 16, various years. For this section of the paper, also see IRS (2006).
4. For data sets where the returns were not present in the SOI sample, the data items were populated with values from the most recent SOI study in which the returns were available.
5. A return that was missing from the population in 2002 and 2003 would otherwise qualify as Consistent if it was present in all earlier years because the sum of all presence observations would total eight. A classification of Terminal is more desirable because the return is not present for the latest two years and will presumably not return.
6. SOI maintains a file of return transaction data extracted annually from the Business Master File (BMF). This file contains a code that indicates whether an 1120 return was processed on the BMF for a given EIN anytime in the "Processing Year," roughly equivalent to a calendar year. The file also contains a tax period indicating the year to which the transaction relates.
7. The entity formerly filing its own 1120 return may no longer due so because it is included in the consolidated filing of another return or group of returns with a different EIN.
8. For SOI's definition of financial items, please see Publication 16, various years.
9. Age was calculated and carried through the analysis as of the base year rather than recomputed each year because increasing appearances in SOI studies would correlate directly with increasing age.
10. For descriptions and counts of unavailable returns, please see SOI's Publication 16, various years.

[11] The sum of Consistent, Inconsistent, and Terminal returns does not equal the total of the base-year returns due to legitimate "duplicate" records. Duplicate records can be present in one study when part-year returns are selected in addition to full-year returns.

[12] These entities may be filing a non-1120 type return or may be included in the consolidation of another return or group of returns.

[13] Entities counted as not surviving may be filing a non-1120 type return or may be included in the consolidation of another return or group of returns.

References

U.S. Department of the Treasury. Internal Revenue Service.

Statistics of Income. Corporation Income Tax Returns. Washington, D.C., various years.

Statistics of Income Bulletin. Summer 2006. Washington, DC 2005.

APPENDIX

Consistent Returns

Size	Data Item	Mean	Standard Deviation
Small	Total Receipts[1]	$6,371,580.79	$57,384,713.78
	Net Income	$120,879.88	$4,079,558.5
	Total Assets	$1,807,835.87	$2,312,005.37
	Net Worth	$639,986.34	$ 4,270,068.29
	Age	16.4479282	16.6014683
Medium	Total Receipts	$53,895,910.61	$106,779,628
	Net Income	$2,511,693.13	$7,407,540.48
	Total Assets	$69,825,074.13	$57,974,136.63
	Net Worth	$29,494,265.47	$44,890,136.91
	Age	22.7388410	24.0182814
Large	Total Receipts[2 3]	$1,061,133,974	$4,499,784,062
	Net Income[4]	$67,978,026.03	$289,082,191
	Total Assets[2 3]	$2,620,483,834	$11,364,833,471
	Net Worth	$928,540,800	$2,638,900,731
	Age[2]	21.5155622	25.4626241

Inconsistent Returns

Size	Data Item	Mean	Standard Deviation
Small	Total Receipts[5]	$4,077,602.06	$15,518,169.88
	Net Income[5]	-$34,503.10	$1,936,312.34
	Total Assets	$1,479,486.82	$2,162,763.78
	Net Worth[5]	$200,645.81	$4,779,648.44
	Age[5]	13.2152608	14.5542741
Medium	Total Receipts	$41,511,957.43	$79,428,394.05
	Net Income	$-598,765.04	$13,179,286.11
	Total Assets	$43,880,737.74	$44,024,985.24
	Net Worth	$8,721,769.96	$62,242,205.94
	Age[6]	18.8165375	20.1862701
Large	Total Receipts[5]	$669,891,521	$1,583,578,000
	Net Income[5]	$20,874,759.10	$88,900,726.62
	Total Assets[5]	$1,346,959,444	$2,956,099,587
	Net Worth[5]	$230,109,460	$405,911,755
	Age[5]	24.9157303	25.7444784

Terminal Returns

Size	Data Item	Mean	Standard Deviation
Small	Total Receipts	$4,952,880.42	$70,038,460.90
	Net Income	-$71,616.51	$6,520,985.17
	Total Assets	$1,382,087.57	$2,069,756.45
	Net Worth	$133,487.37	$5,351,577.13
	Age	12.9184034	14.8322453
Medium	Total Receipts	$47,605,901.58	$95,661,811.13
	Net Income	$1,147,350.28	$9,267,561.22
	Total Assets	$67,945,915.83	$57,212,181.19
	Net Worth	$17,690,263.35	$59,872,085.44
	Age	20.0385105	24.1205414
Large	Total Receipts	$904,927,191	$3,025,364,570
	Net Income	$44,007,051.15	$219,787,529
	Total Assets	$2,777,142,544	$13,275,372,904
	Net Worth	$580,019,080	$2,190,282,973
	Age	23.2558315	29.4368933

Difference across means statistically significant at the 99 percent level unless otherwise noted.

[1] Difference between Consistent and Terminal statistically significant only at the 97 percent level.

[2] Difference between Consistent and Inconsistent not statistically significant.

[3] Difference between Consistent and Terminal not statistically significant.

[4] Difference between Consistent and Inconsistent statistically significant at only the 97 percent level.

[5] Difference between Inconsistent and Terminal not statistically significant.

[6] Difference between Inconsistent and Terminal statistically significant at only the 97 percent level