Guided Tour Through the Section 15 Maze

ATHANASSIOS K. BLADIKAS and CHARLES PAPADIMITRIOU

ABSTRACT

Before the first year's Section 15 data were released hopes were high among academics, researchers, and policy makers. It was believed that this wealth of new, detailed, consistent, and accurate information would help answer conclusively questions about transit productivity and performance and about whether subsidies contribute to better performance or are simply wasted in inefficiencies and wage increases. In addition, the data base was eagerly awaited as a tool that would assist in making "peer" comparisons among transit systems, determine the reasons for performance variations, and possibly help in shaping future federal and state subsidy allocation formulas. After 4 consecutive years of data collection, however, Section 15 proved to be far from what was originally envisioned. Although the quality of data has been improving, for all practical purposes, a uniform reporting system that includes all transit properties receiving federal assistance does not yet exist because of numerous problems that may be classified broadly into four categories: (a) access and structural problems, (b) erroneous and missing data, (c) inconsistencies and definitional ambiguities, and (d) exclusion of important data elements. The problems that were encountered in these areas when using the first 4 years' Section 15 data are presented. Suggestions are also made about how users may solve some of these problems and how future editions of Section 15 data may be improved.

Section 15 of the Urban Mass Transportation Act of 1974 as amended requires transit operators receiving federal operating assistance to annually file auditable reports on their system's operations and finances, create an internal system of accounts and records that can provide accurate and detailed information, and improve overall budgeting and operating management. To minimize data collection burdens a "Uniform System of Accounts and Records and Reporting System" was established with a minimum set of mandatory reporting requirements as well as three possible levels (C, B, and A) of voluntary, more detailed reporting. The Transportation Systems Center (TSC) is responsible for editing, tabulating, storing, and releasing to the public data that transit operators supply annually. So far (January 1985) 4 consecutive years of Section 15 reports are available (FYs ending June 30, 1979, 1980, 1981, and 1982, which are also referred to as Years 1, 2, 3, and 4). All information supplied by transit system operators is available on magnetic tape. Most information supplied at the required level is also available in hard copy form in Annual Section 15 Reports (1).

Before the first year's Section 15 data were released, hopes were high among academics, researchers, and policy makers. It was believed that the wealth of this new, detailed, consistent, and accurate information that was about to become available would be used to answer conclusively questions about transit performance and thus enable researchers to determine the impact of subsidies on performance, formulate or restructure federal or state subsidy allocation policies, make transit system "peer" comparisons, and use the results to train future transportation professionals better. Unfortunately, Section 15 proved to be a disappointment. Missing, inaccurate, and badly structured data made the use of Section 15 information a frustrating experience. For all practical purposes, a uniform reporting system that includes all transit properties receiving federal assistance does not yet exist because of a variety of problems that are discussed in this paper.

ACCESS AND STRUCTURAL PROBLEMS

File WDSPSC, which contains the Weekday Service Period Schedule, is a good, but certainly not the only, example of the structural problems of Section 15 data. In all 4 years of existing data, the first 14 columns of the file provide system identification, fiscal year, and mode and day codes. The remaining columns provide information on 11 more variables, as the TSC documentation given in Table 1 indicates. However, this detailed file description is misleading, because variables MYSRB, PMPSB, NTSRB, and NTSRE are actually five and not the four columns wide that the column description from the Year 2 documentation on the table's left side indicates. TSC attempted to correct the documentation

TABLE 1 File WDSPSC Variable Length Specifications for Year 2 (left) and Years 3 and 4 (right)[a]

Column	Name	Type	Description	
15-18	AMSRB	Real	A.M. service begins	15-18
19-22	AMPSB	Real	A.M. peak service begins	19-22
23-26	MYSRB	Real	Midday service begins	23-27
27-30	PMPSB	Real	P.M. peak service begins	28-32
31-34	NTSRB	Real	Night service begins	33-37
35-38	NTSRE	Real	Night service ends	38-42
39-42	AMPRD	Real	A.M. peak period	43-46
43-46	MDYPD	Real	Midday period	47-50
47-50	PMPRD	Real	P.M. peak period	51-54
51-54	NGTRD	Real	Night period	55-58
55-58	TOTHR	Real	Total hours	59-62

[a]File = 39: (DSNAME=WDSPSC.XMI); from area: OPERA.

for Years 3 and 4 and came up with the column widths on the right side of Table 1, but, again, this is wrong because the record is 60 and not 62 characters long. TOTHR is two not four columns wide as indicated. In addition to the incorrectly specified variable lengths, there are considerable discrepancies between the documentation describing the file and its actual structure, and, to make things worse, file structures vary from year to year in terms of record lengths and block sizes. Fortunately, after some changes in the first 3 years, the third and fourth year structure is identical. Furthermore, structural problems are not major, provided that the researcher remembers that block lengths produced by TSC's computers are multiples of four and, from the second year on, record lengths are also multiples of four to avoid the first year practice of adding blanks in the last record of each block. These problems made trial-and-error the only appropriate technique for accessing some of the files, at least during the first 3 years, and occasionally the discrepancies forced users to abandon a file entirely (2).

ERRONEOUS AND MISSING DATA

One major problem that Section 15 data users have to deal with is the distinction between missing data and legitimate zero entries. For some variables this distinction is trivial. Zero vehicle-miles (VM) for a time period a system operates obviously means that the system failed to report. However, a zero for a minor expense item may mean either that the system spent nothing for that item or that it failed to report. Fielding et al. (3) have covered extensively the problem of missing data, which are fortunately progressively decreasing in consecutive years of Section 15 releases, and it is hoped that this problem can be completely corrected in the future. But, even when nonzero entries are provided, on many occasions they are erroneous.

Often erroneous data cannot be detected by inspection. A bus system may report 3 million vehicle-miles accumulated on its vehicles during the period, and that it used 100,000 gallons of fuel. These figures appear perfectly legitimate individually. However, their ratio indicates an average fuel consumption of 30 MPG, which is impossible for buses. Ratios of other pairs of variables from all 4 years of data produce many more surprises. Average operator salaries exceed $100,000, and vehicle-hours per operator exceed the total hours available in a year. These errors do not exist only in the machine readable data but in the raw data and ratios contained in the annual reports as well. According to the second year annual report, taking the ratio of revenue-vehicle-miles to revenue-vehicle-hours, which can be interpreted as average speed of revenue service, the trains in Boston provided service at 153.9 MPH.

This is the most serious problem with the Section 15 data, because users may unsuspectingly use erroneous data. For example, a person who wants to investigate what determines the observed variations in expenses per vehicle-hour may read only operating expenses, vehicle-hours, fleet size, unlinked trips, and hours of system operations and attempt various regressions. Because errors in the data become apparent only when ratios are formed, this person will not suspect the validity of the data and will include in the regression all nonzero values. However, if even a few of the provided entries are twice or three times (and occasionally 1,000 times) larger or smaller than they should have been, the regression fits will become completely meaningless.

All 4 years of Section 15 data contain .XXXXXXEYY entries (Xs and Ys are integers), mainly because the columns allocated to a variable are not enough to accommodate its digits. In Year 1, for example, the four files that contain information on service supplied, service consumed, and service personnel for rail (Form 407) and nonrail (Form 406) systems by hour of day and day of week (files NRSDWK, RSDWK, NRSTDY, and RSTDY) contain a number of entries (56 to be exact) of the type .XXXXXXE69. In later years, the number of .XXXXXXE69 entries was reduced and was replaced by entries of the form .XXXXXXE11. This could be interpreted as 10 to the 11th and it is correct in some instances. But, on other occasions, the power should be 9 or 10. However, whether the power is correct or not, the letter E thrown in unexpectedly will stop execution by causing conversion from decimal to character errors no matter what the access language or package is.

INCONSISTENCIES AND DEFINITIONAL AMBIGUITIES

For the purposes of this paper, a Section 15 data element is inconsistent if it can be read or derived from more than one data file and the alternative derivations produce values that differ by more than round-off errors would warrant. Inconsistent data exist in all 4 years of Section 15 data. However, unless otherwise indicated, Year 3 examples involving single-mode, motor bus systems will be presented here due to lack of space. There is a total of 187 such systems in the third year of data.

Data on Service Supplied and Consumed

Annual vehicle-miles (VM) is an important variable because it is often used to produce a number of performance indicators. It can be computed by multiplying average weekday, Saturday, and Sunday VM by 253, 53, and 59, respectively (from file NRSDWK containing Form 406 information), or by summing the miles accumulated over the reporting period on all classes of vehicles that a system operates (file RVINV containing information on Form 408, Revenue Vehicle Subsidiary Schedule). Taking the ratio of VM from Form 406 to VM from Form 408 should produce numbers probably in the range of 0.98 to 1.02, because all systems may not operate on holiday schedule the same number of days and the multipliers used (253, 53, and 59) may be inappropriate. However, the ratios are distributed as given in Table 2. Ratios that are

TABLE 2 Distribution of Form 406 to Form 408 Vehicle-Mile Ratio

No. of Systems	Ratio Range
5	More than 1,000
5	31.13-889.19
26	1.11-2.65
50	1.01-1.09
14	Exactly 1.00
62	0.99-0.90
17	0.89-0.56
8	0.00

close to 1,000 are easy to explain. Forms often indicate that figures should be reported in thousands. The line where the number has to be entered may read "Total vehicle miles (000)," although UMTA reversed itself and advised in the first year to report on Form 406 in whole numbers because the small systems could not report meaningfully in thousands. Some operators obviously became confused in the beginning and remained confused in subsequent years. But if the ratios that are close to 1,000 can be explained, a ratio of 2.65 or 0.6 is an unexplainable and inexcusable variation. Similar discrepancies, although not as large, exist if the ratio of vehicle-hours (VH) from Form 406 to platform-hours (time drivers spend operating a vehicle) from Form 321, Operators Wages Subsidiary Schedule, are calculated.

Although VM are accumulated every time a vehicle moves, revenue-vehicle-miles (RVM) are accrued only when a vehicle is serving the public (i.e., RVM is MV minus deadheading miles). Therefore, RVM should always be less than VM, and, similarly, revenue-vehicle-hours (RVH) should always be less than VH. However, using only Form 406 information, the ratio of VM to RVM is identically 1.00 for 42 systems, or 22 percent of those investigated and VH to RVH is 1.00 for 55 systems or 29 percent of the total. On the other hand, some systems appear to have enormous deadheading miles. There are 12 that have a VM-to-RVM ratio greater than 1.2, and 42 for which the ratio exceeds 1.1.

The definition of RVM and RVH is rather confusing and it is certainly a factor that contributes to the inconsistencies. In defining Service Supplied, the original Volume III of the Uniform System of Accounts and Records that contains the required level forms provides the following instruction for Form 406:

> Revenue miles (line 04) and Revenue Hours (line 05) should exclude charter and school bus miles and hours respectively.

A later version of Volume III (Revised July 1982) changes the instruction to read:

> Revenue miles (line 04) and Revenue Hours (line 05) should exclude charter and deadhead miles and hours.

And finally, Volume II of the Uniform System of Accounts and Records that contains definitions provides on page 8.7-1 yet a third version stating that what should be excluded is miles and hours traveled ". . . to and from storage facilities and other deadhead." In summary, deadhead travel is excluded by the second and third definition but not by the first. Charter service is excluded by the first and second definitions but not by the third, and school bus service is excluded only by the first definition. This is certainly sufficient to confuse and discourage even the most conscientious and well-meaning form preparer.

The existence of two alternative derivations for VM and VH implies that speed (an explanatory variable for many performance measures) can be computed four different ways as follows:

(a) VM/VH (Form 406)
(b) VM (Form 406)/platform hours (Form 431)
(c) VM (Form 408)/VH (Form 406)
(d) VM (Form 408)/platform hours (Form 431)

Using these four alternative derivations to compute system speeds, values that differ among themselves by up to a factor of 2 can be obtained as the small sample given in Table 3 indicates.

Data on service consumed are derived from samples (4) and their accuracy depends on how rigorously the sampling instructions are followed. Some, and fortunately few, operators apparently fill the forms by copying the figures from the previous year's report. There are three systems that reported exactly the same unlinked trips in Years 3 and 4, one system that gave the same figures for Years 2 and 3, and finally one that reported identical trips through the first 3 years. The accuracy of data on service supplied (e.g., vehicle-hours and vehicle-miles) depends on the systems' record keeping diligence. Judging from the identical entries for vehicle-hours that exist throughout the 4 years of data, operators are copying data on service supplied from one report to the next more frequently than they are copying on service consumed. There are 10 identical vehicle-hour entries between Years 1 and 2, 10 between Years 2 and 3, and 8 between Years 3 and 4. In addition, there are five systems that report the same vehicle-hours continuously from the first to the third year, and five more that give the same figures for Years 2 through 4. Finally, there is one system that reports exactly the same vehicle-hours for all 4 years.

TABLE 3 Sample of System Speed Variations Depending on Calculation Method

	Speed Derivation Method			
Case	a	b	c	d
1	11.15	8.73	12.10	9.47
2	8.13	6.30	9.36	7.26
3	7.53	6.96	10.61	9.80
4	12.95	13.54	14.67	15.35
5	11.83	13.74	10.38	12.05
6	16.65	17.36	12.50	13.04
7	12.52	13.77	6.35	6.98
8	13.52	19.97	10.74	15.91
9	13.67	11.29	15.50	12.80
10	6.63	5.63	9.63	8.18

Service-Profile Data

Variables that describe a transit system's peaking characteristics provide valuable information for the analysis of efficiency indicators. Information on the peaking characteristics of supplied and consumed service can be obtained from Form 406 (file NRSTDY) and Form 401 (Transit System Service Period Schedule contained in files WDSPSC and WESPSC). A number of indicators can be derived from this information such as peak-to-base ratio in terms of vehicles and VH, shoulder-to-shoulder time (start of a.m. peak to end of p.m. peak), and interpeak hours. This wealth of information should satisfy even the most demanding researcher. However, a detailed examination of these indicators and the raw Section 15 data that produce them causes serious doubts about their validity. First of all, the beginning and ending of each period is not defined clearly. Volume II instructions simply say that the a.m. or p.m. peak begins when "additional service is provided to handle higher passenger volumes . . . when scheduled headways are reduced" and ends when "headways return to normal." The problem with this definition is that it is not concise and it can be interpreted differently by the reporting transit systems. A typical service profile may look like the one shown in Figure 1.

When does a.m. peak service really begin in Figure 1? If the Section 15 definition is followed, just about 4 a.m. (the beginning of a.m. service). A few operators reported just that, but most of them reported some later time, which they felt was more appropriate. This leads to a variety of starts, ends, and durations of service periods that are superficial. The service periods of 37 of the 187 systems investigated cannot even be determined because they have zero entries. This may be a problem introduced by the instructions, because operators are provided with an example on page 6.3-4 of the

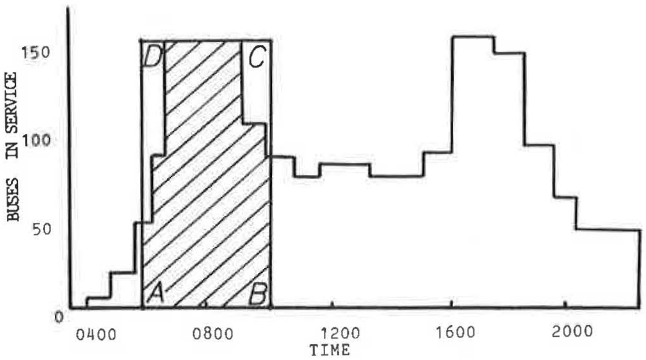

FIGURE 1 Typical transit service profile.

revised Volume III where it is shown that entries should be left blank if service does not fluctuate by time of day. It is therefore impossible to distinguish between systems that did not report and systems that should have zeros due to lack of peaks. Of the remaining 150 systems that report times for their various service periods, 40 report the same time as the start of a.m. service and start of a.m. peak service, and some have the peculiarities given in Table 4. Some of these entries may be legitimate. Systems may really start their a.m. peak before 6:00 a.m. But an a.m. peak duration of 13.8 hr or a p.m. peak longer than 7 hr makes the researcher wonder how valid the rest of the service-period data really are.

TABLE 4 Suspicious Service Period Figures

No. of Systems	Reported Variable	Value
29	Start of a.m. peak	Before 6:00 a.m.
3	Start of a.m. peak	Before 5:00 a.m.
7	End of a.m. peak	11:00 a.m. or later
21	Duration of a.m. peak	4 hr or more
1	Duration of a.m. peak	13.8 hr
40	Duration of p.m. peak	4 hr or more
10	Duration of p.m. peak	5 hr or more
2	Duration of p.m. peak	More than 7 hr

The problem is compounded if an examination is made of the VH reporting during each period. It would be expected that the ratio of (VH during period)/{(vehicles during period)•(duration of period)} would not exceed 1.0 for the peaks and would roughly be in the 0.80 to 0.99 range. This is obvious from Figure 1. Vehicles during the period are, according to the definitions, the maximum number of vehicles. Therefore, the ratio for the a.m. peak is graphically the rectangle ABCD over the cross-hatched area under the service profile. If the midday period is considered, the ratio should be expected to exceed 1.0 for the same reason. However, when this ratio is computed, it is over 1.0 for 39 systems in the a.m. peak and 34 systems in the p.m. peak.

Vehicle Data

Fleet size is critical because it is used in the denominator of many performance indicators. If the number of vehicles is incorrect, particularly if fleet sizes are small, the performance ratios can be over- or underestimated by rather wide margins. There are actually three files containing information on vehicles. File TRSYS gives the total number of revenue vehicles, file TRSVEH gives the total number of revenue vehicles by type, and file RVINV describes revenue vehicles in more detail and groups them according to make and model year. The distinction between active and revenue vehicles is also made in the last file, and, according to the definitions, the number of revenue vehicles should be larger than the number of active vehicles.

However, the ratio of revenue to active vehicles is less than 1.0 for 32 systems. In addition, the ratio of total revenue vehicles from file RVINV to revenue vehicles from file TRSYS is greater than 1.1 for 24 systems and greater than 1.5 for 8 systems, which are not small, and therefore the discrepancy does not involve just a few vehicles.

Examining the ratio of revenue vehicles to peak vehicles in order to find out the percentage of reserve buses a system has (spare ratio), often produces surprises as the sample of some of the worst cases given in Table 5 indicates. The revenue vehicles for this table were taken from file TRASVEH, and those are also the figures that TSC provides in the third year annual report. Peak vehicles are the vehicles serving the highest of the two peak periods. A revenue-to-peak vehicle ratio of up to 1.5 may be believable. However, it is extremely difficult to think of a reason why a system would have three or four times the maximum number of vehicles it needs. A closer examination of the data given in Table 5 together with the data in file RVINV reveals that the system of the third case does not have 102 revenue vehicles; it owns only 52 outright and leases 4 more. Similarly, the system of the fourth case owns only 59 and leases from related parties 10 vehicles, and the system in the fifth case owns just 34 vehicles.

TABLE 5 Revenue-to-Peak Vehicle Ratios for Some Systems

Case	Revenue Vehicles	Peak Vehicles	Revenue-to-Peak Vehicles
1	3,362	1,948	1.73
2	392	206	1.90
3	102	24	4.25
4	107	41	2.61
5	47	17	2.76
6	25	11	2.27

Even if a system's fleet size can be determined, calculating the fleet's capacity is not an easy task. There are two ways to compute average capacity. The first is to divide revenue capacity miles (RCM) by RVM from file NRSDWK. A second figure can be obtained directly by multiplying the number of vehicles times the sum of standing and seating capacity for every vehicle class in file RVINV and dividing by the total number of active vehicles. Table 6 gives 10 of the worst discrepancies between the two alternative derivations of average capacity. The figures differ between 12 percent and more than 100 percent. The inconsistencies in Table 6 are serious because the cases were taken from Year 4 data and from a reduced set of 87 systems that had reasonably clean data in Year 3. This implies that, even if a researcher works hard to select a set of systems that have reasonable data for a set of variables in a given year, data for the same systems cannot be extracted blindly from another year of data either for the same or for different variables. The same painstaking clean-up and cross-checking efforts have to be undertaken for each year and for every variable.

TABLE 6 Average Fleet Capacity Discrepancies

	Average Fleet Capacity	
Case	RCM/RVM (File NRSDWK)	Direct Computation (File RVINV)
1	64	73.5
2	54	63.2
3	45	58.3
4	60	53.8
5	62	51.4
6	50	59.1
7	80	90.2
8	41	61.0
9	45	99.9
10	42	53.4

Along with fleet age, the variable average cumulative miles per vehicle could be used for the prediction of maintenance efficiency variables such as road cells per vehicle-mile. Average cumulative miles per vehicle can be obtained from Form 408 (file RVINV). The data in Table 7 indicate that about one-third of the 187 systems in the sample have provided absolutely useless information by reporting unbelievably high or low figures. The figures that are in the millions are obviously operator reporting errors. Some saw the column heading "Average Accumulated Miles per Vehicle (000)" and instead of reporting in thousands, they multiplied the actual figure times one thousand, thus creating the 1-million-mile discrepancy.

TABLE 7 Erroneous Mileages on Vehicles

No. of Systems	Average Cumulative Miles per Active Vehicle (range in thousands)
6	157,892-630,000
2	18,903-32,751
6	1,487-3,122
11	503-955
6	Less than 18
27	0.0

Employee Data

Serious inconsistencies exist mainly in the files containing information on operating statistics (400 series Forms). Some of them have already been presented. Another common inconsistency exists between total labor-hours for inspection and maintenance (I/M) from file MNPENC (Form 402) and the sum of employees in the maintenance function contained in file EMPSCH (Form 404). These employees are grouped in five personnel classes as follows:

- Class 21 Maintenance executive, professional, and supervisory personnel;
- Class 22 Maintenance support personnel;
- Class 23 Revenue vehicle maintenance mechanics;
- Class 24 Other maintenance mechanics; and
- Class 25 Vehicle service personnel.

The ratio of total labor-hours for I/M to the sum of employees in these five classes should be 2,000, because this is the annual person-hour equivalent for one employee according to the instructions (although TSC appears to use 2,080 to produce the hard copy annual reports). However, the ratio ranges from the teens to the hundred thousands. This is another example of the confusing reporting instructions. The original Volume III contained the following Form 402 instruction on page 6-5 concerning total labor-hours for I/M:

> See definition Volume II Section 8.5, Page 8.5-1. NOTE: You should include all hours worked by employees in Employee Record hours to nearest hour.

However, on page 8.5-1 of Volume II labor-hours for I/M are defined as: "The labor-hours of transit system maintenance personnel working on revenue vehicles for the period." Examining the definitions of Employee Elements and Classification on pages 8.4-2 to 8.4-7 of Volume II, it seems that the only personnel working on revenue vehicles for the period are the revenue vehicle maintenance mechanics and the vehicle service personnel (Classes 23 and 24) and not all classes of maintenance employees as the Volume III instructions indicate. The revised Volume III changed the instruction to read:

> See definition Volume II Section 8.5, Page 8.5-1. NOTE: You should include all hours worked by employees whose labor expenses were charged to Function 061 INSPECTION AND MAINTENANCE OF REVENUE VEHICLES. (See definition on pages 7.4-29 to 7.4-31 of Volume II).

Some operators interpreted this as: "include only Class 23." In addition, Function 061 pertains only to the most detailed, voluntary level of reporting (A). Taking the ratio of employee-hours to employees using the latest interpretation, a range of about 10 to 100,000 is again obtained. Apparently operators became so confused that they simply threw in numbers, and 14 did not even bother to report.

A useful indicator for the examination of operator productivity is revenue vehicles per operator. The value of this ratio is determined by the number of additional operators a system wishes to have during a period either because of its standby policy or because of nonoperating duty assignments. Form 406 and file NRSTDY contain information on vehicles operated during each time period and the number of full- and part-time operators during the same period. Taking the ratio of vehicles over full- plus part-time operators, a researcher would expect values roughly in the 0.85 to 0.99 range. A ratio higher than 1.0 is impossible unless the vehicles are automatically controlled, and 15 percent or more extra operators would be an extreme waste of resources. However, the sample of cases given in Table 8 produces unreasonable ratios that are another example of sources of errors that produce useless Sec-

TABLE 8 Erroneous Data on Operators

Case	Period	Vehicles	Operators (full + part time)	Vehicles per Full-Time + Part-Time Operators
1	X	113	12.8	8.83
	Y	46	4.6	10.00
2	X	8	2	4.00
3	X	32	52	0.62
	Y	18	38	0.47
4	X	281	340	0.83
	Y	264	358	0.74
	Z	101	200	0.50
5	X	7	13	0.54
6	X	30	56	0.54
7	X	16	27	0.59
8	X	27	46	0.52

tion 15 data. The first and most illogical case can be corrected by moving the decimal one place to the right in the number of operators; this is obviously a transcription error. The up to 100 extra operators of the fourth case are probably the result of a confusing definition. Operators are defined to be those scheduled to operate vehicles. Thus it is conceivable that if during a time period, for example midday from 10:00 a.m. to 4:00 p.m., the number of vehicles in operation were 100, the system might be inclined to report the number of operators as 210 because 105 operators worked until 2:00 p.m., and at that time another 105 replaced them.

The conflicting instructions and misunderstandings presented so far are certainly not the only ones. Holec et al. (5) mention more in their discussion of Year 1 data. However, Section 15 data inconsistencies cannot always be blamed on conflicting or unclear instructions. The revised Volume III instructions for Form 321 state clearly that the figure to be placed in the dollar column for total operating and nonoperating time ". . . must balance to the dollar amount reported in Object Class 501.01--Operators' Salaries and Wages on the appropriate Section 15 expense reporting forms." The appropriate form suggested in the instructions is Form 301 (Expenses Classified by Function), and the total for Object Class 501.01 can be found in file X0. Although the instructions are perfectly clear, some operators report figures that differ between -3 and 22 percent.

Financial Data

Financial and expense-related data that do not come from samples or inconsistent collection procedures should be expected to be more accurate, and generally they are, although occasionally some expense figures are questionable. Average salaries by function and for the entire system can be computed by taking the ratio of salaries over the number of employees. Five such ratios are possible in the most aggregate function level (i.e., operations, vehicle maintenance, nonvehicle maintenance, general administration, and total), and they would not be expected to fall outside the approximate range of $10,000 to $30,000. However, some systems appear to pay large amounts for some employees, whereas other systems do not even pay the minimum wage. There are eight systems that report average salaries of more than $60,000 and as high as $800,000 and 16 systems that appear to be paying less than $7,000 and as low as $436 per year.

Major discrepancies and inconsistencies exist in the financial data reported in series 100, 200, and 300 Forms. Balance sheets and revenue summaries were reproduced for all systems and for each of the reporting years and checks were performed on the additions and to see whether assets were equal to liabilities and capital. If discrepancies existed, efforts were made to resolve and correct them.

Year 1 data contain nine resolvable discrepancies most of which arise from the omission of the minus sign in depreciation items ranging from $2,000 to $48 million. Year 2 contains 28 resolvable discrepancies most of which involve the reporting of items 10 times as large as they should have been, and the omission of the minus sign from accumulated losses, which appear as gains. Year 3 data contain 30 resolvable discrepancies about half of which involve the improper addition of revenues. Finally, Year 4 contains 12 resolvable discrepancies produced by all of the previously mentioned causes. Although the errors in the financial data are small in number, they are serious when they involve large systems. For example, the Washington Metropolitan Area Transit Authority's assets were overestimated by $21 billion in the second year. This figure should be compared with the approximately $15 billion in assets, $6 billion in liabilities, $9 billion in capital, and $7 billion of revenues for all Year 4 systems together. Thus the errors of one system can make the calculation of totals or averages completely useless. This is also true for variables that represent service supplied and consumed.

Although most major errors in the financial data can be resolved, there are a large number of unresolved discrepancies that are quite minor (less than $10 in most cases). Table 9 gives a summary of the unresolved discrepancies. Year 3 is the best of all. Sometimes the dollar amounts are smaller than the number of systems, because the real deviations between totals and the sum of their component parts were used. Thus, if a total is larger than its parts by $2 in one system and smaller by $2 in another, the discrepancies will cancel each other out. The large dollar amounts that appear in Table 9 are contributed mostly by a single system in each case.

The errors in the financial data provide an indication of what portion of the Section 15 inaccuracies can be blamed on transcription errors because there are no ambiguities involved in the preparation of a balance sheet. However, revenue inaccuracies may arise from the confusion of cash accounting, which most systems use, and accrual accounting that is required for Section 15 reporting (5,6). If there are about 30 major (more than $10 that can arise from round-off errors) balance sheet-related errors in each year of Section 15 data, it might be safe to assume that about 10 percent of the reporting systems have errors in nonfinancial data as well because of transcription errors and poor quality control.

EXCLUSION OF IMPORTANT DATA ELEMENTS

On the basis of the structure of the Section 15 system, it appears that those who originally conceived the breakdown of the data into nine areas and 62 files had intended to create a system that grouped the reported information better than the forms them-

TABLE 9 Unresolved Discrepancies Between Totals and Component Parts in Financial Variables

	Year 1		Year 2		Year 3		Year 4	
Discrepancies	Systems	$	Systems	$	Systems	$	Systems	$
Assets	41	329	37	2,479	9	4	12	13,859
Liabilities	6	3	10	896,664	6	8	9	1,006
Capital	27	172	20	17,975	12	6	8	2
Revenue	7	22,702	9	79,501	6	38	8	24
Total	81		76		33		37	

selves. This would allow the Section 15 data user to go easily to a file or set of files and obtain the pertinent information no matter what his interest was. Whether the data were to be used to evaluate issues pertaining to transit financing, performance evaluation, safety, or the cost-effective provision of service, the analyst would have a small number of files to access and could perform the intended task efficiently. The noble effort was also undertaken to provide users with even more data than the operators supplied. File UAREA is the produce of such an effort, and it is supposed to contain two variables, the square miles of area (USQMI) and the urban population (UPOP) that a system serves. First of all, variable USQMI is always zero, so half of the file is useless. The values of UPOP are for all practical purposes useless too. The New York City transit system with more than 10,000 vehicles is shown to serve 15.59 million people, and so do 35 more systems, including ones like the Resort Bus Lines of Yonkers, New York, with all of its 8 vehicles. Apparently, the system's address was the only determinant of population served. Area and population served are important elements in the analysis of a system's effectiveness, and for all practical purposes they are missing from the data. An attempt was made to include the variable UPOP in models that analyze bus systems' performance, and it failed to enter any of the equations with even a minimal degree of explanatory power (see paper by Bladikas and Papadimitriou in this Record.) Obtaining data for USQMI and UPOP is certainly not an easy task. UMTA could obtain this information with a research grant, or at production cost from the Bureau of the Census.

The usefulness of the Section 15 data would improve significantly if some additional information were collected. Most of this information could be collected by simply expanding Form 001 (Transit System Identification Schedule) to two pages or by replacing Form 332 (Pension Plan Questionnaire), required for systems with 25 or more vehicles and apparently ignored by everybody, because only a handful fill it out every year. The additional information could be useful not only because it would provide more explanatory variables for research in the area of performance evaluation, as others have already suggested (7), but because it would also make possible additional consistency checks. The minimum additional information should be

- The system's fare structure by mode;
- Entry, average, and maximum salary by function;
- The system's organizational and management structure (city agency, independent authority, whether it is managed by an independent company, and so forth);
- Union contract data (whether employees are unionized and some key contract provisions such as split shifts, spread times);
- Vehicle retirements and purchases during the reporting period; and
- Revenues (at least transportation revenues) broken down by mode.

SUGGESTIONS FOR IMPROVEMENT

There are three parties responsible for the introduction of errors in Section 15 and, therefore, the situation will be corrected only if all three make efforts to improve. UMTA has to work on the four volumes of the Uniform System of Accounts and Records and Reporting System to clear up the definitional ambiguities and conflicting instructions that the various volumes contain. In addition it should become stricter and reject operator reports that are obviously erroneous or inconsistent. Operators certainly see Section 15 as an additional burden imposed on them by UMTA and appear to be the major cause of errors. Unfortunately, the lack of clear instructions on the forms that they are filling out does not make that burden easier. TSC can improve its quality control and the accuracy of the published information, as it is capable of doing if judged on the basis of the excellent quality of other, more voluminous than the Section 15, data that it releases.

As far as the machine readable data are concerned, the field lengths for the variables that do not fit in them should increase so that the EYY entries can be eliminated. The detailed file descriptions should be changed to include the blanks (when they exist) at the end of every record, so that the record lengths match those described in the access JCL. The last set of files (LOOKUP) is practically redundant and useless because it contains the codes that exist on the reporting form anyway. The information contained in the 17 data files should be placed on just 7 typewritten pages to provide the user with an easy reference. Mileage-weighted fleet capacities, ages, and capacity miles would be more valuable indicators in the annual reports, because it is most likely that the largest, smallest, oldest, or newest vehicles are used only during certain periods of time and the entire fleet is not in service from the beginning to the end of the service period.

There is a set of about 40 flags in the machine readable Section 15 data. Only eight of the files have from one to twelve flags each. Apparently when the files were initially created these flags were supposed to act as checks for the values of every variable. However, soon this effort was abandoned, leaving only eight files with flags. Unfortunately, even the existing flags are of little value because they do not appear to be flagging anything. Flags should be placed on all files and put to use. It is possible to devise a flag system as follows:

Flag Value	Meaning that Item Is
0	Correct
1	Not reported (missing)
2	Out of a preset range
3	Inconsistent

Of course, if such a flag system is to be implemented, TSC has to develop software that will perform range checks and look for all possible inconsistencies in the reports. When a report is received, it should be analyzed and returned by UMTA to the operator for corrections. Data should be flagged if, even after the corrections, they are still erroneous or inconsistent.

Looking at Section 15 data for all 4 years suggests that there is some improvement in their quality with every subsequent edition. However, the improvements have to be speeded up by a factor greater than the normal learning process for all involved parties is bringing about. This is needed particularly because (a) the existence of a few years' data will soon allow the analysis of time series, (b) states are using Section 15 data to develop their own performance measures [e.g., Michigan (8)], and (c) future legislation may include in the allocation formulas even more performance measures than does the Surface Transportation Assistance Act of 1982.

REFERENCES

1. National Urban Mass Transportation Statistics. Report UMTA-MA-06-0107-81-1. UMTA, U.S. Department of Transportation, May 1981.

2. G.J. Fielding, M.E. Brenner, and O. de la Rocha. Using Section 15 Data for Transit Performance Analysis. Interim Report, UMTA-CA-11-0026-1. UMTA, U.S. Department of Transportation, Jan. 1983.
3. G.J. Fielding, M.E. Brenner, and O. de la Rocha. Using Section 15 Data: Adopting the Magnetic Tape Version for Statistical Analysis. In Transportation Research Record 961, TRB, National Research Council, Washington, D.C., 1984, pp. 28-35.
4. Circular UMTA C 2710.1. UMTA, U.S. Department of Transportation, Feb. 22, 1978.
5. J.M. Holec, Jr., D.S. Schwager, and M.J. Gallagher. Improving the Usefulness of Section 15 Data for Public Transit. In Transportation Research Record 835, TRB, National Research Council, Washington, D.C., 1981, pp. 9-15.
6. Circular UMTA C 2710.5A. UMTA, U.S. Department of Transportation, Oct. 1, 1981.
7. S.C. Anderson, G.J. Fielding, and D. Methe. Comparison of Supply, Demand and Cost Models Using UMTA Section 15 Data. UCI-ITS-WP-81-4. Institute of Transportation Studies, University of California, Irvine, Dec. 1981.
8. J.M. Holec, Jr., D.S. Schwager, and A. Fandialan. Use of Federal Section 15 Data in Transit Performance Evaluation: Michigan Program. In Transportation Research Record 746, TRB, National Research Council, Washington, D.C., 1980, pp. 36-38.

Research Implications of Proposed Changes in the UMTA Section 15 Reporting System

JOEL E. MARKOWITZ

ABSTRACT

An intensive effort started in 1983 to review the Urban Mass Transportation Act (UMTA) Section 15 reporting system for transit statistics. Although many transit industry professionals have been involved, few researchers are aware of the ways in which proposed changes would alter the national data base. A summary of the efforts to date is presented, and the implications of the proposed changes for those who have been routinely relying on Section 15 data for the conduct of research on U.S. transit systems are highlighted.

In 1974 Section 15 of the Urban Mass Transportation Act of 1964 was amended to require that transit agencies receiving federal formula grant funds submit a uniform report on their financial and operational characteristics each year (1). The requirement grew out of a large-scale study that examined transit industry accounting practices in detail. The result was a series of forms and manuals documenting accounting definitions that would be used for the required annual reports (2). The standards laid the framework for upgrading the management information systems in the industry as a whole. The nonfinancial data did not receive as much careful study in the early days and have continued to cause some problems, especially now that certain of those data have been incorporated into the new Section 9 transit block grant formula program (3).

Although transit industry representatives were actively involved in the work leading to the adoption of the Section 15 standards, some problems in the reporting system appeared only after the first few years of implementation (FY 1978-1979, 1979-1980 and 1980-1981). A massive amount of information is involved, from a few hundred data elements for the lowest level of required reporting to a few thousand elements for the larger multimodal systems. Inaccuracies in reporting, misunderstanding of definitions, inconsistencies within reports, and difficulties in quality control joined with some instances of outright refusal to cooperate. The result was a national data base with serious limitations. At the 1984 Transportation Research Board Annual Meeting, several presentations were made on transit performance analysis using Section 15 data. All illustrated the many problems inherent in the data that required either elaborate cleaning procedures or simply the exclusion of whole sets of agency reports. Fielding et al. (4,5) reported on the difficulties in organizing the magnetic tape version of the FY 1980 data for statistical analysis. Of 304 agencies that reported that year, 106 had missing data that prevented their being used in the performance analysis work. Vaziri and Deacon (6) similarly used the FY 1980 data base and had to work around problems caused by missing data. Hobeika et al. (7) found so much missing data on the items of