3. D. Young. And Budd Stays In. Mass Transit, Vol. 6, No. 3, March 1979, p. 7.
4. The United States and the International Market for Rail Equipment. Richard J. Barber and Associates, Washington, D.C., March 1978.
5. Pub. L. 88-365 (1964), Urban Mass Transportation Act of 1964, as Amended Through December 1978. 49 U.S.C. 1601 et seq.
6. A. Dzingelis. (New York City Transit Authority). R-62 Passenger Car Procurement. Presented at American Public Transit Association Rapid Transit Conference, Cleveland, Ohio, June 15, 1982.

Publication of this paper sponsored by Committee on Rail Transit Systems.

Automated and Passenger-Based Transit Performance Measures

KELVIN BUNEMAN

ABSTRACT

Operational performance measures of a transit system are often best expressed in terms of the passenger. These include measures of productivity per passenger trip or per passenger mile, measures of crowding or seat capacity, and measures of on-time performance or schedule adherence. At the Bay Area Rapid Transit, detailed operational data are available daily as a by-product of automation, which saves considerable manual data collection. A computer program to use these data to produce operational performance measures is described. The program runs daily to allow schedulers to allocate vehicles accurately, which yields energy and maintenance savings, and for ongoing analysis of delays. Passenger-based delay measures are produced that have hitherto been unavailable.

Measurement of crowding or seat occupancy is relevant in two ways. First, during the peak periods, most transit systems are subject to a fleet size constraint. There are only so many buses or rail cars available, and some passengers have to stand. It is important to balance capacity among all routes to avoid an excess of standees on any particular route. Passenger demand varies within the peak period; for example, it may be highest between 7:30 and 8:00 a.m. but still heavy at 8:30 a.m. Schedulers must also balance capacity between the different times within the peak. Second, during the off-peak periods, it is important to operate no more vehicles than the minimum needed to guarantee most passengers a seat. This can be done in train systems by changing headways or train lengths. Headways may be mandated by externally imposed minimum service requirements, but train lengths are adjustable subject only to physical constraints such as the minimum indivisible train set length. An excess of vehicles will incur unnecessary energy and maintenance costs because of extra vehicle miles traveled. An undersupply of vehicles causes crowding, which may drive passengers away.

Measurement of on-time performance is generally done in terms of the number of late or missed bus runs or train runs. However, it is useful to produce statistical information on train delays as experienced from the passenger's point of view. Passenger-based on-time performance measurement puts in proper perspective the magnitude of system delays, differentiating, for example, between a 30-min peak-period delay versus a 30-min off-peak delay. It can assist a transit system in better allocating its resources toward various planned improvements and also provide a better measure of understanding of the actual impact of these improvements on passenger service.

The major distinction between the standard train on-time reporting of passenger service and passenger-based service measures is the weighting or importance given to a delay event. In train on-time reporting, as much value is given to an empty 3-car train delayed for 10 min as to a packed 10-car train delayed for the same 10 min. Obviously, the packed 10-car train carries a greater impact on passenger service than the empty 3-car train and should be so measured. Passenger-based delay measures reflect this proper weighting by basing calculations on passenger trips that are on time rather than trains that are on time.

A report prepared for UMTA in 1978 (1) addressed service availability at length, where service availability is defined as the impingement of failures on passenger-perceived service. The report is for automated guideway transit (AGT) systems, but the concepts discussed in the report are applicable to any transit system where there is substantial control over the vehicle and right-of-way, whether automated or not. In the report it is found that passenger-based service availability measures are most desirable but that these require more data than are collected by most operating systems.

In 1979 Heimann (2,pp.314-322) developed a pas-

senger-based dependability measure. He modeled several hypothetical train delay incidents, showing train-to-train delay interactions and passenger-to-train delay interactions. He then computed delays for several cohorts of passengers traveling between several different stations for each of the train delay incidents. Though Heimann's work was based on the Massachusetts Bay Transportation Authority (MBTA) Red Line in Boston, a real system, he intended to exemplify a process rather than actually compute the daily dependability of the Red Line.

PRODUCTION OF PERFORMANCE MEASURES

Data Collection

In many transit systems, ride checks and point checks are necessary for measurement. These manually collected data must be coded for analysis, which requires a good deal of labor in the field as well as clerical effort at the central office. Budget limitations enforce limited checking and sampling. San Francisco's Bay Area Rapid Transit (BART) is fortunate to have both central computer train control and automated ticketing, which allow production of both train and passenger movement data. Central train control makes a record each time a train opens or closes its doors and of every train action.

BART pioneered the stored-fare magnetically encoded ticket system, which has since been adopted by the Washington (D.C.) Metropolitan Area Transit Authority (WMATA). A complete description is inappropriate here; the key point is that the proper fare is determined at the exit fare gate by reference to a table indexed by station of entry. Thus, origin information for each passenger is available at the exit station. Obviously, destination information--the fare gate's own location--is available and the time of exit can be recorded. This allows accurate production of passenger counts each day by origin and destination and time of day--a travel demand modeler's dream.

Computation of Performance Measures

The steps from raw data on passenger and train movement to the final performance measures are shown broadly in Figure 1. Structured systems analysis notation is used (3).

The passenger flow model (PFM) system within Figure 1 (so called because one part of the system includes a simulation model) is the focus of this paper. Elements of the data flows into it are defined as follows:

- Train action is one record identifying an event such as an open door or a closed door, including time of day, location, train identification, and train length.
- Exit counts by origin and destination are passenger exit counts for each of the 34^2 possible origin-destination (OD) pairs plus time of day.
- Detail timetable for each scheduled train run on each route gives scheduled door closing times for the first and all intermediate stops. It gives door opening times for the final stop. In addition, it gives expected final arrival times at other stations for passengers who will transfer. Like the public timetable, this timetable is offset 60 sec earlier than the central control internal train schedule. This allows trains to get up to 60 sec ahead of the internal schedule without penalty.

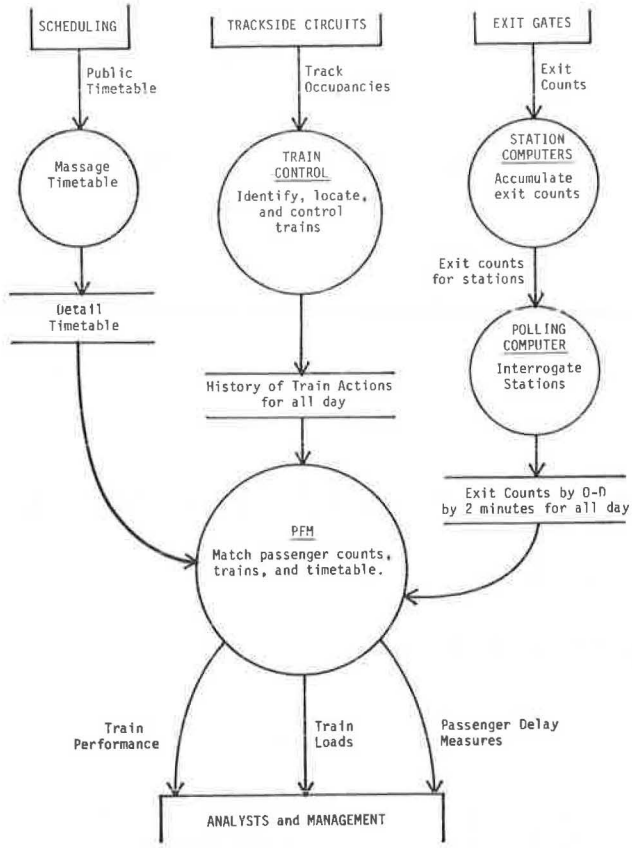

FIGURE 1 Data flow for performance monitoring.

The PFM system consists of computer programs written in FORTRAN. The following detailed descriptions of the programs are intended to accompany Figure 2.

Program 2.1: Automated Edit of Train Actions

Program 2.1 checks the train actions for consistency against an internal map showing stations and routes. It assembles train actions into train runs. For these purposes, a train run is a sequence of train actions from train reversal to train reversal (generally from one end of the line to the other end of the line). Each action in a train run is tagged with the starting and ending stations for that run.

Program 2.2: Matching Trains with Timetable and Generation of Train Performance Measures

For each train run, Program 2.2 finds the slot in the detail timetable that most closely precedes the train's actual departure time. It then checks every actual station stop along the run against the scheduled time for that stop. Because time is measured to 1-sec rather than 1-min accuracy, it is important to distinguish between when the door is opened (arrival) and when the door is closed (departure). The PFM system adopts the convention that the door closing time is the more important time for the first and all intermediate stops on a train run. The door opening time must be used at the final stop. By checking at each station, this program can annunciate each train delay by location, which allows accurate analysis of delays. Four types of delay event are possible. First, a train may be dispatched

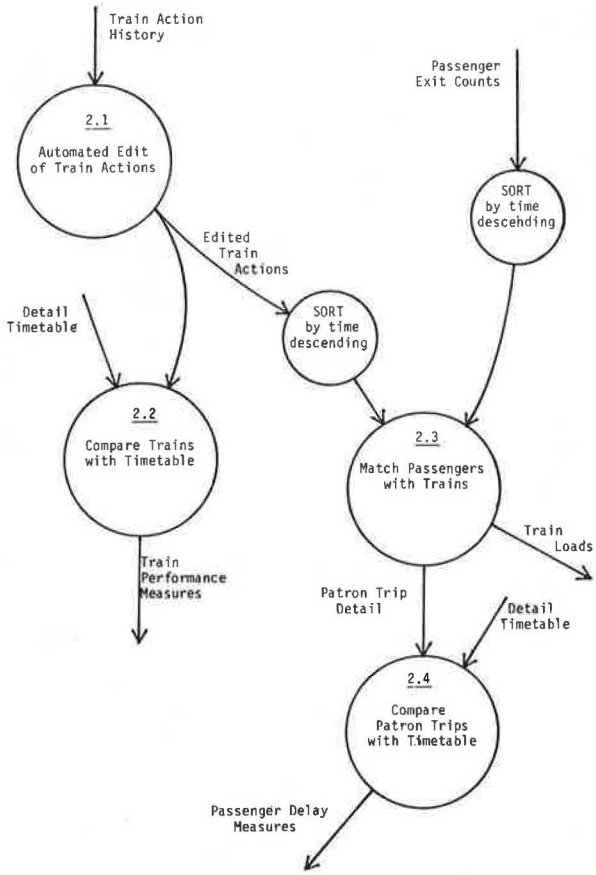

FIGURE 2 Data flow within PFM process.

late. Second, a train may be delayed en route between adjacent stations. Third, a train may be delayed at a station with its doors open--an excessive dwell time. Fourth, a delay may extend across several stations--a slowly moving train.

Program 2.2 can also generate standard train performance measures such as the number of train runs on time and end to end, the total number of successful train runs, and the number of cancelled or incomplete train runs. The program also has access to a table of distances between stations. Using these distances and the actual train actions and train lengths, it generates total revenue vehicle miles.

Program 2.3: Matching Actual Patrons with Actual Train Actions

Program 2.3 is the most intriguing program in the PFM system. Train actions consist of all door openings and closings with time, station, route, and train identification. Patrons are recorded at their exit by time, exit station, and entry station. Because only patron exit data are available, this program is a time-reverse simulation, hence the nickname, BACKWARDS. Unlike stochastic simulations for experimental use, this is deterministic. The simulation may be best described by following one patron's trip backward from the exit:

1. Hold patron in exit station;
2. Load patron on the previous train that served his entry station; set patron arrival time to train door opening time;
3. Follow train back toward patron's entry station; if a transfer is needed, unload patron into transfer station; hold the patron at the transfer station until the previous train that served his entry station;
4. Set patron departure time to train door closing time and unload patron into entry station;
5. Hold patron in entry station until previous service between this station pair; and
6. Set previous departure time to time of previous train door closing; record patron trip (steps 5 and 6 allow recording of the headway to deduce patron waiting time).

Although the foregoing description is for only one patron, Program 2.3 actually works on all patrons and trains in parallel. It is driven by the train actions. Exit stations, transfer stations, entry stations, and trains are simple data structures containing counts of patrons and arrival and departure times. Note that there is no need to combine the three station data structures (exit, transfer, and entry) unless measures of platform crowding are needed. Timing is essential to the correct matching of patrons with trains. The patron exit counts are scanned every 2 min, which is less than the minimum scheduled headway. The time required for a patron to leave the train and ride the escalator to the lobby and fare gates must also be considered.

Program 2.3 uses a map showing stations, routes, and transfer requirements. The map tells which route or routes serve any given pair of stations. At BART, three suburban East Bay routes converge into one pair of tracks to serve the central city, San Francisco. The program knows that certain trips (OD pairs) are served by multiple routes. Typically shorter trips are served by multiple routes, whereas longer trips are served by only one route.

Passengers do not need to use their ticket to transfer from one train route to another. In some cases, they have a choice of transfer station. For example, passengers traveling from Fremont to Concord may transfer at any of the three downtown Oakland stations (see Figure 3). The current version of the program simply assumes that MacArthur Station is the preferred transfer station when there is a choice. This does not cause errors in patron travel

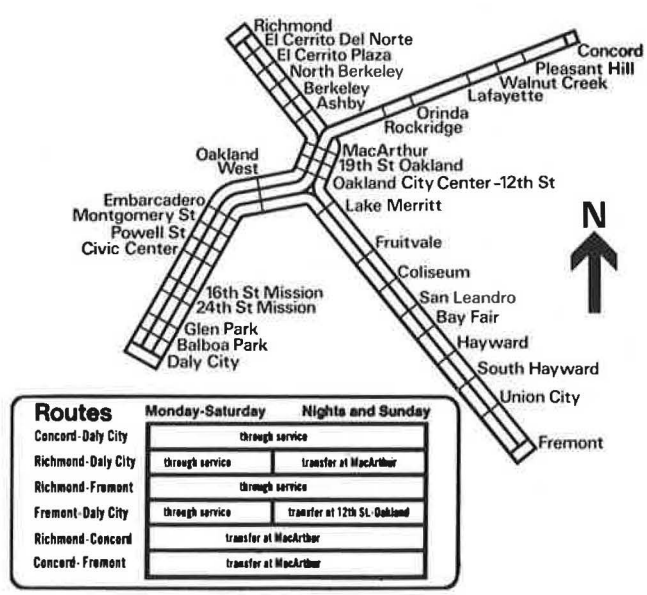

FIGURE 3 BART system map.

time measurement and affects train load measurement accuracy only between MacArthur Station and 12th Street Station. Other possible sources of error in this train load modeling include patrons who transfer even when direct service is available between their origin and destination, patrons who take excursion rides and enter and exit at the same station, and employees traveling on passes.

The train-load output of the program is sorted by location and time and reported for each of several locations. For each location, the report has one line per train, showing time, train identification, train length, number of passengers, and the ratio of passengers to seats. The accuracy of this report has been checked by walking through trains and counting passengers and found to be very high.

Program 2.4: Matching Passenger Trips with Timetable

Program 2.4 finds expected and actual wait time and expected and actual travel times for an entire day of patron trips. All patrons who made the same trip are grouped together into a patron move. Trips are the same if the origin, destination, and train run are the same. A patron move includes entry station, departure time, previous departure time, exit station, arrival time, train identification, and number of patrons involved (frequently just one or a few patrons). To avoid excessive searching, this program proceeds by departure time for all stations and routes in parallel. It maintains pointers into the timetable for each route and station.

The actual departure time and previous departure time define an interval within which the program finds all applicable departures on applicable routes. Three cases are possible: one scheduled departure (normal), no scheduled departure in interval (extra train or early train), or more than one scheduled departure in interval (cancelled or late train or trains). If only one scheduled departure is found between actual departures, there is some chance that a patron expected that departure and some chance that he arrived on the platform too late for that departure and expected the following departure. For example, with a 15-min scheduled headway, a train that started 5 min late appears to some patrons to be 10 min early. Two important assumptions about passenger expectations and passenger behavior are as follows:

1. What is the passenger's expectation of departure time? BART now publishes a detailed timetable, so for passengers who use the timetable the answer is clear. However, some passengers may simply expect a certain headway; they may just expect an average wait time equal to one-half the published headway. This headway expectation is less exacting than a timetable expectation. For transit lines with short route headways, the headway assumption is appropriate. For lines with long route headways such as commuter rail service, the timetable assumption is more appropriate. The patron on-time measures developed so far at BART use the tougher timetable assumption.
2. At what rate do patrons arrive on the platform to wait for their train? This question is relevant for trains that depart a few minutes early and for split-headway operation. Patrons who arrive just before a scheduled departure only to see taillights disappearing into a tunnel must wait a full headway. Although the first assumption implies that all patrons know the timetable, it is also assumed that they do not all arrive exactly at the time the train is scheduled to depart. Nor is it assumed that they all arrive uniformly over time. An estimate of how many patrons arrive promptly rather than uniformly over time must be made to assign delays. This information is not available automatically because only exit data with fine time resolution are available. This question has been investigated at least once before for rail service. In London, a 1970 study (4) suggested that for a 15-min headway, 42 percent of peak-period suburban rail patrons arrive promptly and 23 percent of off-peak patrons do so. The proportion of prompt arrivals tended to decrease with shorter headways. A 1970 study of British bus passengers (5) found similar results. In January 1983, an informal study at BART's suburban stations showed 11 percent prompt arrivals at the 7.5-min headway and 13 percent prompt at the 15-min headway. This is much lower than the British data, perhaps because BART did not then publish a timetable. Because BART now publishes a timetable, an arbitrary promptness criterion of 25 percent prompt arrivals was chosen for the patron on-time measure. This compromise value provides an adequate penalty for early trains.

A numerical example for one origin station follows: Suppose that trains are scheduled to depart at 8:00, 8:15, 8:30, ... Suppose that the 8:00 departure is on time and the next actual departure is at 8:25, and Program 2.3 indicates that 120 patrons rode the 8:25 train. The promptness criterion says that 25 percent of all patrons arriving between 8:00 and 8:15 arrived exactly at 8:15; the remainder arrived uniformly over time. In this example, Program 2.4 will compute the following:

1. Sixty passengers arrived uniformly at 4 per minute between 8:00 and 8:15, expecting to catch the 8:15 schedule;
2. Twenty passengers arrived promptly at 8:15, expecting to catch the 8:15 schedule;
3. Forty passengers arrived uniformly at 4 per minute between 8:15 and 8:25, expecting to catch the 8:30 schedule.

Then of the 120 patrons riding the 8:25 departure, 60 + 20 = 80 expected the 8:15 schedule and 40 expected the 8:30 schedule.

Having determined which schedule each passenger expects, it is easy to determine expected arrival time and compare it with actual arrival time. Figure 4 shows the average weekday distribution of passenger delays for several months at BART. This includes both waiting time and travel time. For transfer passengers, the entire train trip from initial wait through final arrival is measured.

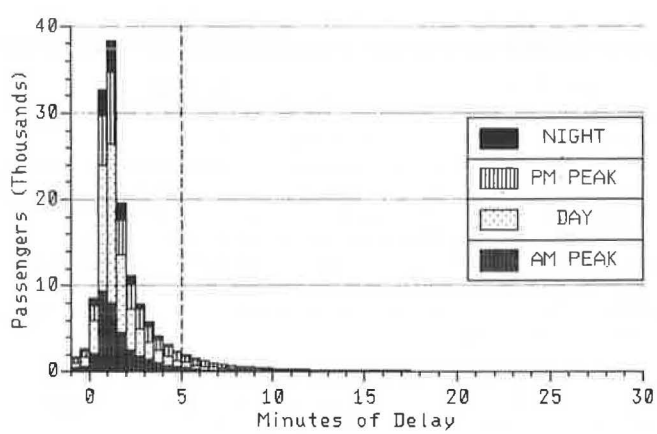

FIGURE 4 Histogram of passenger delays (May-October 1983).

Costs of Automated Performance Measurement

Approximately 3 man-years were required to develop the PFM software. This effort has been divided between independent software contractors and BART's own research engineers. The PFM system runs during the graveyard shift at the end of revenue train operations. It requires about half an hour to execute each night on BART's IBM 4341 and uses up to 1 megabyte of virtual memory. During the graveyard shift, however, there is little contention for the computer and a full megabyte of real memory is generally available. Operation and use of the PFM system requires less than 1 hr a day from data processing operations staff and an hour per day of a schedule analyst's time.

Automated performance measurement is considered a by-product, not the primary purpose, of the automatic fare collection (AFC) system, so it is inappropriate to include the costs of AFC as part of the costs of this measurement. The station computers and central polling computer cost about $750,000 in 1976 dollars and require an average of 0.4 full-time electronic technician to maintain. However, these computers provide other benefits by collecting data from other station equipment, such as ticket vending machines, for use by the treasury and police departments.

USE OF AUTOMATED PERFORMANCE MEASURES

The train-based delay event list generated by Program 2.2 is used during next-day analysis of train delays. It helps to bridge the gap between aggregate service quality measures, such as total train delays or total patron delays, and individual component failure measures. It helps distinguish between primary delays and secondary delays. Primary delays are those directly caused by a failure, whereas secondary delays are due to the train traffic congestion that follows a primary delay.

The passenger loads are averaged by train over 10 weeks, with separate moving averages for each day of the week. From these average passenger loads, measured at the five critical locations in the BART system, an optimal allocation of revenue vehicles is made. This allocation is to minimize crowding subject to several constraints such as minimum and maximum train lengths and the timetable (minimum headways).

Ongoing passenger load monitoring is necessary as demand patterns shift. Some changes are predictable, such as seasonal changes and changes during the week. Some demand changes are not so predictable, such as the ongoing overall increase in BART patronage or public response to changes in feeder-bus routes or station parking availability.

Approximately 12.5 full-time clerical positions would be required for passenger load monitoring if the automated system were not available. Whether loads are monitored manually or automatically, the monitoring is worthwhile. Operation of just 2 percent more vehicle miles than the load requires would cost about $584,000 per year just for energy and maintenance.

Passenger loads are also reported quarterly and used for long-range planning. As patronage grows and trains become more crowded, it is necessary to acquire more vehicles and reduce headways. Examination of passenger loads by time and by route can show the shape of the peak-period demand. A flatter, wider peak suggests that transportation systems management measures such as flextime are effective.

A project is under consideration to communicate detailed passenger load information to the passengers. This would inform passengers which trains are usually more crowded and which trains usually have remaining capacity. Surveys show that about one-third of BART's passengers could vary their working hours. These passengers might switch to less crowded trains if they were provided with suitable advisories.

Although automated measurement of passenger loads has been in use at BART for several years, the passenger-based delay measures are newly developed. The exact form of a measurement of passenger delays has not yet been decided. Two basic forms are possible:

1. Trip dependability, defined as follows:

D = (number of passenger trips on time times 100) divided by total trips,

where an on-time trip is a trip with delay less than some tolerable amount, such as 5 min. This measure would be expressed as a percentage; 99 percent would be very dependable service, whereas 50 percent would be very poor service.

2. Expected delay, defined as follows:

D = total passenger minutes of delay divided by total trips,

where a passenger minute of delay is one passenger delayed 1 min. This measure would be expressed in minutes; 0.1 min would be very dependable service, whereas 5 min would be very poor service. Unlike the trip-dependability measure, this is sensitive to the duration of delays.

Little is known about passenger disutilities with respect to length of delay for transit systems. For example, is a 10-min delay once a month better or worse than a 5-min delay twice a month? Leis (1) raises this type of question in greater detail.

Summary measures of on-time performance or delay, or even the distribution of delay such as that shown in Figure 4, are most useful if a link can be established between the aggregate measure and specific problems such as component failure. The causative relationship between component failure and total passenger delays can be broken down into four steps:

1. Equipment failure causes a slow or stopped train;
2. Time is required to diagnose the problem and restart or remove the train (primary train delay);
3. Secondary train delays occur because the primary train blocked the track and because the resultant high passenger loads cause station dwell delays; combined primary and secondary delays are a delay event;
4. Passenger delays occur both on board the delayed trains and while waiting downstream of the delay.

As Heimann (2) points out, much is known about step 1 and the relationship between steps 1 and 2. An automated vehicle and component repair tracking system is in use at BART. The link between steps 2 and 3 is currently established manually for each major delay event with computer assistance. This could be largely automated except that multiple primary delays occur and interfere with each other, making it difficult to distinguish delay events. The link between steps 3 and 4 has not yet been well established. Because passengers may experience the effects of more than one delay event in the course of a trip, this linkage should be further analyzed. Currently a computer report of passenger delays by

train run is used to support the manual analysis of train delay events and to identify the likely cause of major passenger delays. The report breaks down delay time into an en-route component and a waiting-time (at platform) component. Results so far show that the waiting-time component of delay is generally larger than the en-route component.

Although this discussion has been largely oriented to measurement of the current state of an existing system, the passenger-based aggregate measure can be applied to future systems or to future configurations of existing systems. The real train actions can be replaced by simulated disturbed train actions produced by a train system simulator. The real passenger counts can be replaced by forecast passenger counts. Program 2.3 would be replaced by time-forward matching of trains and passengers.

CONCLUSIONS

With increasing automation in transit operations, the amount of operating data generated will increase. Automated OD ticketing and automated recording of vehicle movement make possible accurate passenger-based performance measurement. The PFM software could be configured to serve any guideway transit system that has data on all train movements and all patron exits by origin. Even automated zone-based or flat-fare ticketing systems, which record entries but not exits, can provide data to supplement manual counts and sampling.

Further research is needed to better understand the relationship between primary vehicle delays, secondary vehicle delays, and passenger delays. When these links are understood, it will be possible to allocate maintenance resources where they will best benefit the passenger. More research is also needed to investigate passenger expectations regarding wait time and timetable adherence and passenger annoyance due to delays of differing lengths.

ACKNOWLEDGMENT

An earlier version of the PFM was developed in conjunction with the BART 1980-1981 Short Range Transit Plan update funded by an UMTA Section 8 Technical Studies Grant.

REFERENCES

1. R.D. Leis. The Development of Measures of Service Availability. U.S. Department of Transportation, 1978.
2. D. Heimann. The Determination of Transit System Dependability. Proc., 1979 Annual Reliability and Maintainability Symposium, IEEE, New York, 1979.
3. T. DeMarco. Structured Analysis and System Specification. Yourdon, New York, 1979.
4. D.A. Lynam and P.F. Everall. Public Transport Journey Times in London, 1970. TRRL Report LR 413. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1971.
5. C. O'Flaherty and D.O. Mangan. Bus Passenger Waiting Times in Central Areas. Traffic Engineering and Control, Vol. 2, No. 9, 1970.

Publication of this paper sponsored by Committee on Rail Transit Systems.

Monitoring the Quality of Service from the Passengers' Perspective

MATTHEW K. du PLESSIS

ABSTRACT

Management's concern with customer satisfaction and the common methods of gauging patrons' assessment of service are discussed. A method of performing surveys of trains and stations based on sampling techniques is then described. Performed on a periodic basis, the studies have an audit-type quality that helps alert management to potential problems and areas needing further investigation. The results of the studies are reviewed, and sample tables and graphs are presented. As a result of the data generated by the surveys, changes in train schedules were developed and further studies of the vehicle-cleaning process initiated. The increased reliability of the system is shown dramatically in a graph of published travel time variance.

Customer satisfaction is an important concern to managers in any organization but especially to those in a service industry such as public transportation. Being publicly owned, such transportation agencies find themselves subject to even closer scrutiny than private companies. For these and other reasons, senior managers of rapid transit agencies are