Paper Submitted for TRB Conference – January 2010
Submission Date: July 29/09
Word Count: 5,815 + 8 Figures & Tables

Service Reliability Measurement using Automated Fare Card Data: Application to the London Underground

David L. Uniman* <uniman@alum.mit.edu>
Massachusetts Institute of Technology &
EMBARQ Center for Sustainable Transport – Mexico
Felipe Carrillo Puerto 54
Col. Villa Coyoacan, C.P. 04000
Mexico City, D.F.
(52) 04455-357-16432

John Attanucci <jattan@mit.edu>
Massachusetts Institute of Technology
77 Massachusetts Ave. Rm 1-274
Cambridge, MA 02139
(617) 253-7022

Rabi G. Mishalani <mishalani@osu.edu>
The Ohio State University
2070 Neil Ave., Rm 470
Columbus, OH 43210
(614) 292-5949

Nigel H. M. Wilson <nhmw@mit.edu>
Massachusetts Institute of Technology
77 Massachusetts Ave. Rm 1-238
Cambridge, MA 02139
(617) 253-5046

**Corresponding Author*

Service Reliability Measurement using Automated Fare Card Data: Application to the London Underground

David L. Uniman, John Attanucci, Rabi G. Mishalani, Nigel H. M. Wilson

Abstract

This paper explores the potential of using Automated Fare Card data to quantify the reliability of service as experienced by passengers of rail transit systems. For those systems requiring both entry and exit fare card validation, the distribution of individual passenger journey times can be accurately estimated. Using this information, a set of service reliability measures are developed that can be used to routinely monitor performance, gain insights into the causes of unreliability, and serve as an input into the evaluation of transit service. An estimation methodology is proposed that classifies performance into typical and non-recurring conditions, which allows analysts to estimate the level of unreliability attributable to incidents or disruptions. The proposed measures are used to characterize the reliability of one line within the London Underground under typical and incident-affected conditions using data from the Oyster Smart Card system for the morning peak period. A validation of the methodology using incident-log data confirms that a large proportion of the unreliability experienced by passengers can be attributed to incident-related disruptions. In addition, the study revealed that the perceived reliability component of the typical Underground trip exceeds its platform wait time component and equals about half of its on-train travel time as well as its station access and egress time components, suggesting that sizeable improvements in overall service quality can be attained through reliability improvements.

Introduction

The increasing adoption of Automated Fare Collection (AFC) systems by transit agencies presents an opportunity to study and improve service quality by allowing analysts to measure the performance of the system from the perspective of its passengers' actual experiences. Given that there is not always a one-to-one relationship between the delivery of the service and passengers' experiences, analysts have searched for more accurate and cost-effective ways to capture service quality, which go beyond a traditional focus on adherence to the operating plan. The emerging availability of AFC data is allowing for further developments in this regard where passengers' travel experiences can be observed directly in a comprehensive manner and on an on-going basis.

For transit operators, the automated measurement of service quality improves several different functions central to the provision of service, including:

- routine monitoring and detection of changes in service quality,
- evaluation and management of operator performance,
- identification of the causes of service quality problems and appropriate strategies to address them, and
- prediction of travel behaviour responses to changes in transit level of service and ridership/revenue forecasting.

To a large extent, the development of these and other applications depended on the data that were available for measuring service quality. Initially, manually collected data from surveys allowed transit agencies to directly observe a limited snapshot of the passenger experience due to its high-cost of collection and processing, restricting sampling frequency and system coverage. The emergence of Automated Vehicle Location (AVL) and Automated Passenger Counting (APC) technologies led to improvements in this area, producing a wealth of data on individual vehicle movements and passenger demand that could be used to infer the passenger experience more cost-effectively. In recent years, the proliferation of individual passenger Smart Cards as part of AFC systems has produced another important set of data, especially in transit systems requiring entry- and exit-validation where individual passenger journey times can be measured. This latter source of data allows transit analysts to directly and cost-effectively *measure*, as opposed to indirectly *estimate*, the passenger experience in terms of individual origin to destination (OD) travel times on the system, thus providing a new, easy-to-access resource for the monitoring, evaluation, and analysis of service quality.

One aspect of service quality that may be more readily available from AFC data is the assessment of reliability, or the degree of variability of certain attributes of service. Due to its focus on variability, the study of reliability in the context of service quality requires a large number of disaggregate observations, which were previously difficult to obtain. Taking advantage of the availability of AFC data for a system requiring both entry and exit validation, this paper proposes a set of service reliability measures that can be used to both gain insight into the importance and nature of this attribute of service, and develop practical applications for use by transit agencies. Specifically, the paper seeks to achieve the following three objectives:

- explore the feasibility of AFC data as part of transit performance measurement efforts,
- use the set of measures to gain insight into the contribution of reliability to overall service quality, and some of the causes of reliability problems, and
- operationalize passenger views on reliability into a set of practical measures and use them to illustrate potential applications.

The paper achieves these three objectives by first presenting in the next section an overview of reliability as defined from the point of view of passengers and some previous passenger-oriented measures. This is followed by a definition of the proposed set of reliability measures and the methodology for their estimation. The subsequent section provides insight into the contribution of non-recurring incidents to unreliability and overall service quality and illustrates possible applications of the measures by characterizing the performance of the London Underground using data from the Oyster Smart Card system. The final section provides some general conclusions and future research directions that stem from this study.

Overview of Service Reliability

A study by Abkowitz et al. defined reliability as *"the invariability of service attributes which influence the decision of travelers and transportation providers."* (1) The distinction made by this definition between the perspective of passengers and of operators is useful when developing measures of reliability since each is affected differently by uncertainty in the service. The effects of service variability on the effectiveness and efficiency of the service are well studied, including the underutilization of vehicle capacity due to arrival time irregularity, and the need for additional resources to compensate for longer recovery times at terminals (2). Transit patrons, however, are affected differently, which is discussed further in the next section.

Passenger Perspectives on Reliability

Attitudinal surveys of reliability indicate that besides the consistency of service attributes related to comfort and safety, passengers are most concerned about the predictability of total journey time and its individual components (3). Focusing on the variability of travel times, there are two differing perspectives on the precise way in which passengers are impacted by unreliable service (4). The first contends that the uncertainty of total journey time, or its components (e.g. wait time), has an inherent disutility for passengers, and can be quantified through a measure of variability such as the standard deviation of the travel time distribution. The second perspective focuses on the way unreliable service hinders passengers' ability to make optimal travel decisions that minimize their disutility. In the case of low-frequency service, passengers would be interested in arriving as close to their desired vehicle's departure without missing the service, while generally expecting that the departure takes place on time (5).

In the case of high-frequency service,, passengers would be more concerned with choosing an appropriate departure time that would minimize the chance of arriving late at their destination, assuming that they have an expected arrival time for the trip. This scenario can be understood by taking the case of a passenger with a desired arrival time T_{DES}, traveling in a system with perfectly deterministic journey times. In this case, the optimal departure time would be the time calculated by subtracting exactly the expected duration of the trip from T_{DES}. In a stochastic environment, however, for every departure time, passengers experience an arrival time distribution with a non-zero probability of arriving later than desired. This is represented in Figure 1, showing how for a departure time T_{DEP}, an average travel time T_{AVG}, and an expected arrival time T_{ARR}, there is a non-zero probability of arriving after the desired arrival time T_{DES}.

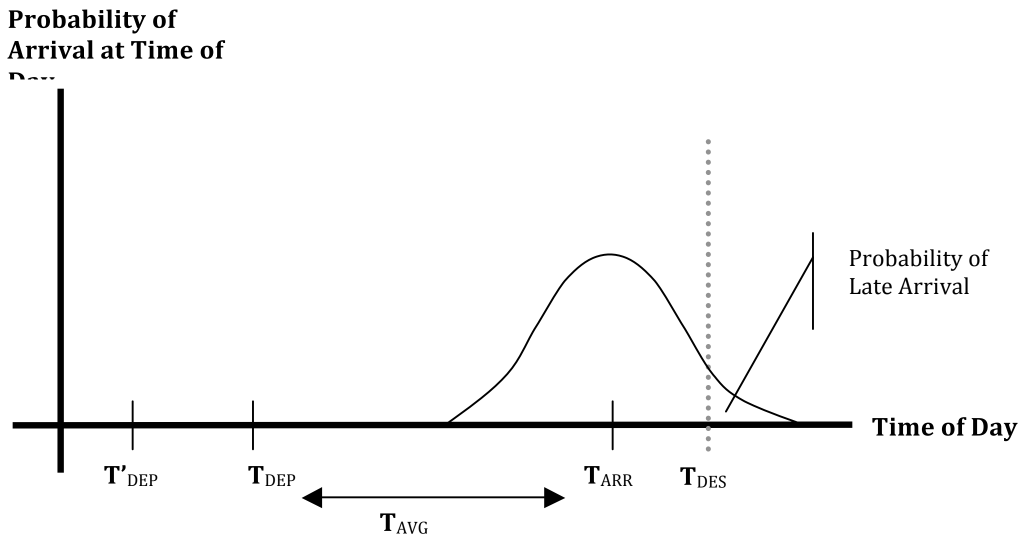

Figure 1: Journey departure time decision and probability of late arrival

If the passenger valued an on-time arrival (where any arrival at or before T_{DES} is considered to be "on-time") more than the time spent traveling, the departure time would be shifted earlier to reduce the probability of being late. This additional time budgeted by a traveler, for example the difference in time between T_{DEP} and T'_{DEP}, can be thought of as a "slack" or "buffer" time. Naturally, as the variability of travel time increases, passengers must leave a greater buffer time for a similar probability of on-time arrival, assuming the problem is not severe enough to warrant a mode change. There is evidence that some passengers value travel time reliability higher than the average speed of the service (6), in addition to anecdotal evidence such as that found by a London Bus study where a passenger described reliability as: *"not having to leave ¾ of an hour early to get to work on time"* (7). This understanding of the way passengers perceive and react to reliability serves as the basis for the measures proposed in the following section.

Passenger-Oriented Reliability Measurement

There are a number of additional studies on the development of reliability measures, with a large proportion of them focused on users of private autos. These studies reflect important parallels with measures of reliability focused on transit passengers because both consider the variability of total travel time at the individual user level.

Lomax et al. (8) identified three types of reliability measures: measures of statistical range or variability, measures of additional budgeted travel time, and measures of extremely long delays or "tardy trips". The first type focuses on statistically measuring the compactness of the travel time distribution about some "central" value such as the mean or median. De Jong et al. (9) identify variance and percentile-based measures of compactness, with the study by Lam and Small (6) finding evidence that the latter more effectively represents passenger perceptions of unreliability. The underlying concept behind these measures of compactness is that variability has an inherent disutility for passengers and excludes any effects of unreliability through schedule delays. The second type of measure is related to the "slack" or "buffer" time that is added by passengers through a shift in their departure time. This additional time can be expressed as a percentage of the average trip duration or as an absolute value additional to some expected travel time. The third type of measure relates to the concept of schedule delay, and consists of the likelihood that a passenger will arrive at their destination *unacceptably* late. This is estimated by determining a threshold for what is considered to be an "unacceptable" travel time for passengers, either as a percentage of the typical travel time, or a certain fixed time in minutes.

Two studies with a particular focus on developing reliability measures for public transport users are those by Furth et al. (10) and Chan (11). The first study argues that the effects of unreliable (bus) service are not accurately represented by traditional measures of reliability, such as mean passenger wait time, since they underestimate the effects of unreliability on passengers. This is because unreliability is expected to force passengers to allocate an additional amount of time to complete a journey through shifts in their departure time decision. The authors argue that in the case of high-frequency bus service, this also affects the amount of wait time passengers include in their travel schedules, because if passengers only allowed for the mean wait time when making their trip, they would arrive late around half the time. Therefore, it is proposed that a better measure of the effects of reliability on passengers would be the 95th percentile wait time, or the amount of wait time budgeted in order to complete a journey by the desired arrival time. In addition, the difference between the budgeted wait time and the mean wait time can be considered as the *potential wait time*, or the time that would have potentially been used for waiting, except that in most cases it would actually be spent at the destination after an early arrival. While the Furth study focuses only the effects of reliability on wait times (thus ignoring the variability of in-vehicle travel time), its use of actual vehicle headway data to represent the effects of unreliability on passengers more accurately provides a more passenger-centric measure.

The study by Chan (11) also proposed a passenger-centric measure of reliability that focused on capturing the compactness of the travel time distribution instead. The proposed measure was defined as the difference between the 95th percentile travel time and the median travel time of the travel time distribution for a particular OD pair, and was measured using AFC data from London, which record time and location of entry and exit for all Oyster Smart Cards. This research was one of the first to begin to explore the potential for using this source of data for measuring service reliability, serving as a useful starting point for the measures proposed here.

The Reliability Buffer Time Metric

Building on the prior work to define and quantify reliability from the passenger perspective, and the characteristics of AFC Smart Card data, a set of measures is proposed that capture total passenger travel time variability (in the context of their departure time decisions), on both a "typical" or recurrent basis and a more exceptional, incident-related basis. The Reliability Buffer Time (RBT) is defined as the amount of “slack” or “buffer” time that passengers must allow for above their typical travel time in order to arrive with certainty at their destination with a specified level of probability. This measure is defined mathematically as the difference between an upper percentile value N and an indicator for the typical travel time, or M percentile. Setting N to be the 95th percentile and M to be the 50th percentile travel time for an O-D pair’s total travel time distribution, the RBT is given by:

$$\text{RBT} = (95^{\text{th}} \text{ percentile travel time} - \text{median travel time})_{\textit{O-D, Within-Day Time Interval, n-Days}} \quad [1]$$

The median travel time represents the typical duration of a journey, and is preferred to the mean due to its insensitivity to outliers. The upper percentile, in this case the 95th, captures the level of certainty to which passengers would like to ensure their on-time arrival, approximately representing a once-a-month chance of late arrival for commuters. The proposed 95th percentile value strikes a balance between passenger relevance and realistic expectations of the service, and is not susceptible to any biases due to unusual individual passenger behaviour (e.g., waiting for a friend inside a station), which Chan (11) found generally to occur only beyond the 99.5th percentile for the case of the London Underground. This parameter can be modified to represent the preferences of transit passengers in a specific market. The subscripts in Equation 1 indicate various dimensions of aggregation including OD pair, within-day time interval (e.g. 15-minute, 3-hour, full-day), and a certain span of time over multiple days (e.g., a 20-weekday sample). Modifying the latter two aggregation dimensions allows analysts to estimate the RBT across varying levels of temporal resolution (e.g., within the AM Peak). The RBT can also be aggregated spatially beyond the individual OD pair. Uniman (12) presents one approach for estimating a line level measure of reliability, where the RBT of each individual OD pair within a line is weighted by the volume of journeys it carries during the time interval being studied, and a weighted average performance is found. This is given by:

$$RBT_{Line} = \frac{\sum_{OD \in Line} Vol_{OD} \cdot RBT_{OD}}{\sum_{OD \in Line} Vol_{OD}} \quad [2]$$

Where Vol_{OD} = total passenger journeys for origin-destination pair “OD” within “Line”, and
RBT_{OD} = reliability buffer time for origin-destination pair “OD” within “Line”.

This definition of a line-level measure of RBT does not take into account journeys transferring from other lines. However, these transfer trips could be taken into account in final performance estimates through a network assignment, using the assumption that single-line segments of a transfer trip would have a similar travel time distribution as the "same line" OD trips on the selected line.

The Excess Reliability Buffer Time Metric

An extension to the RBT metric is developed by specifying a baseline measure of reliability from which to compare the observed performance of any given sample period. This baseline allows the analyst to define quantitatively a desired or expected performance, and use it to separate the causes and levels of unreliability that are tolerated and penalized. The proposed baseline is developed based on two types of factors that were observed to affect performance. The first is the effect of recurring service characteristics – such as journey length, scheduled headway, and whether the trip involved interchanges – on passenger travel time (see 11 and 12). The second type captures the effects of incident-related service disruptions on the reliability experienced by passengers. These non-recurring events were found by Uniman (12) to have a substantial impact on performance relative to the contribution of service characteristics, raising the concern that the quality of service for a particular period could be strongly affected by a handful of days with severe delays largely caused by factors external to the operating plan.

Based on these insights, a reliability baseline is developed by separating performance into two categories: "typical" (or "recurrent") and "incident-affected." The former represents the performance of the system under typical conditions and captures the effects of the first type of factors including the irreducible level of travel time variability caused by the discrete nature of transit service (e.g., even under perfectly regular headways, wait time variability will be non-zero), and can be expected to be stable over time for a given system and operating plan. The latter category represents the performance of the system under disruptions due to serious incidents. Together, these two performance categories represent the actual experience of passengers. Thus, the Excess Reliability Buffer Time (ERBT) is defined as the amount of buffer time required by passengers to arrive on-time with 95% certainty *in addition* to the amount of buffer time that would have been required under typical conditions, and is given by:

$$\mathrm{ERBT} = (\mathrm{RBT}_{\mathrm{Overall}} - \mathrm{RBT}_{\mathrm{Typical}})_{\textit{O-D, Within-Day Time Interval, n-Days}} \qquad [3]$$

Where the $\mathrm{RBT}_{\mathrm{Overall}}$ represents the overall actual buffer time experienced by passengers and the $\mathrm{RBT}_{\mathrm{Typical}}$ represents the performance under typical conditions. In addition, the subscript *n-Days* would pertain to the sample period for the former component of the ERBT measure. The use of the typical performance category as the basis for the measure's baseline measure of reliability has the advantage that it establishes a realistic performance reference that takes into account the characteristics of the service. This baseline has the potential benefit of allowing comparative analyses across parts of the network with different characteristics. This approach also provides an opportunity to begin to distill the different factors that contribute to unreliability by separating out the effects of disruptions on service quality, thus aiding in the selection of strategies aimed at improving the level of service of the system.

A line level measure, the $\mathrm{ERBT}_{\mathrm{Line}}$ can be estimated similarly to the approach described for the line-level RBT, weighing the performance of each OD pair by the volume of passengers experiencing that level of reliability. In addition, based on these two aggregations, it becomes straightforward to derive the relationship:

$$RBT_{Line} = RBT_{Line,Baseline} + ERBT_{Line} \qquad [4]$$

Where $RBT_{Line,\ Baseline}$ = the line level measure of the baseline RBT.

Equation 4 simply states that the measure of the overall RBT at the line level is the sum of the baseline RBT at the line level given by Equation 2 and the ERBT at the line level as defined above. This simple relationship makes it possible to also decompose the overall reliability at the line level into a baseline and excess level of performance.

The classification approach used to separate the typical and incident-related performance of the system was based on stepwise regression (see 13, 14, and 12). Each day-specific peak period was compared with the remaining day-specific peak periods in terms of the magnitude and degree of recurrence of delays, as measured by the 95th percentile travel time, to determine whether a particular day should be classified as incident-affected or typical. Those days that exhibited non-recurring and comparatively large delays were separated from those remaining days, which were empirically observed to have a typical travel time distribution. The journeys for all peak periods under each category were then pooled to estimate the travel time distribution representing performance under each of the two types of conditions on which the RBT measure was applied. This procedure could be applied to estimate a baseline performance using a sample period of a fairly long duration (e.g., 20 weekdays or a 3-month sample).

Finally, the typical RBT metric also allows analysts to estimate the proportion of journeys that experienced unacceptably high levels of unreliability. Taking the baseline performance as the threshold for acceptable service, the proportion of journeys with a travel time higher than the 95th percentile travel time of the typical travel time distribution could be considered unreliable. Since this value would include 5% of all journeys under typical conditions by construction, the percentage of journeys greater than this value would represent those unreliable journeys caused by incident-related disruptions. Figure 2 summarizes the RBT, ERBT, and Percentage of Unreliable Journeys (PUJ) for an OD pair's travel time distribution.

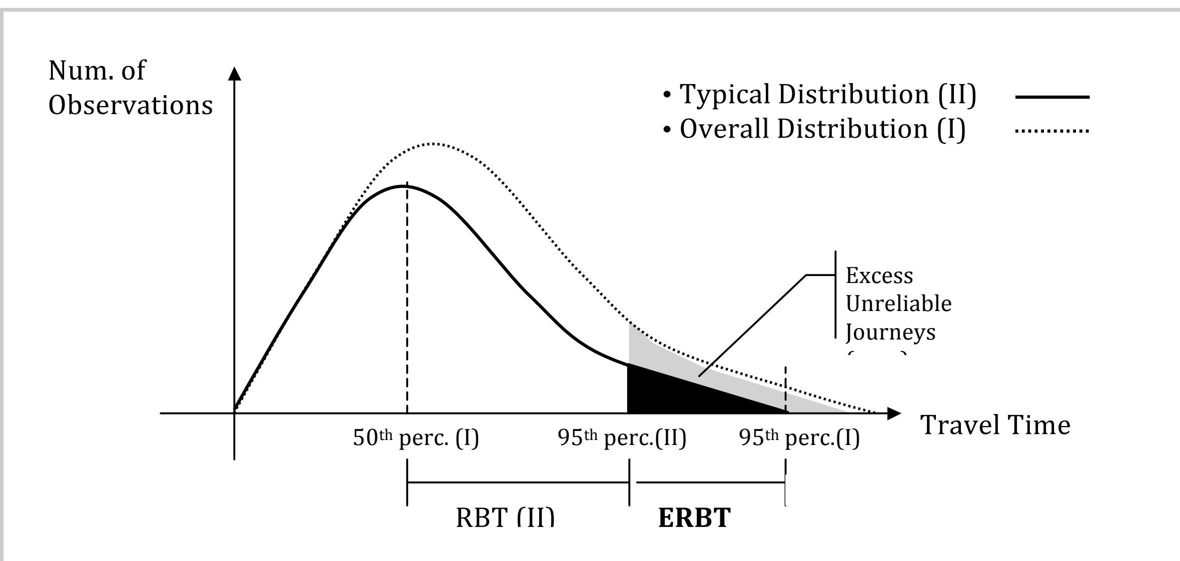

Figure 2: Illustration of the RBT, ERBT, and PUJ

The area shaded in black represents those journeys that by construction experienced unreliable service and were made on days classified to be part of the typical performance of the system (i.e., 5% of all journeys on typical days). The area shaded in gray is the number of unreliable journeys that occurred in excess of those journeys considered unreliable by construction, and it represents the impact of incidents on the proportion of all trips receiving unreliable service.

Application to the London Underground

Using data from the Oyster Smart Card, the proposed reliability measures are used to develop applications for the London Underground. For the first application, the RBT metric is used to characterize the reliability of the service and to validate the classification of performance against incident-log data. Two additional applications illustrate how reliability can be measured as part of routine service quality monitoring efforts, and how the contribution of this attribute relative to average travel times can be estimated. Lastly, a brief discussion on the potential use of the proposed measures for supplementing passenger information systems is included.

Description of the London Underground and Oyster Smart Card Data

The London Underground is a primary component of the city's transport system, with 12 heavy rail lines providing high-frequency service(every 2-5 minutes) over the trunk portions of the network throughout the day (15). Within the public transport network, the Underground carries 27% of all journeys and 53% of all journeys to and from Central London during the peaks, reflecting its importance to the commuter market.

Key performance objectives for the Underground are improvements in the duration and consistency of door-to-door travel times (16). Reflecting this concern for the passenger experience, TfL uses two performance measurement systems relevant to this study: the Journey Time Metric (JTM) service quality monitoring system, and the Nominally Accumulated Customer Hours (NACHs) system for quantifying average delays attributable to service disruptions.

The JTM is a passenger-focused performance measurement system, where through the use of multiple data sources, surveys, and models, a passenger's *average* total travel time is estimated. This figure is assembled from individual estimates at the trip component level, including access and egress times, platform wait times, in-vehicle times, ticket purchase time, and time lost due to closures and incidents. The estimated passenger experience is then compared against a benchmark reflecting ideal conditions (i.e., service operating according to plan), and the excess journey time is derived at the line segment, line, and network levels. Finally, to fully represent the impact of the service on passengers, weights are applied to reflect overall demand and the value of time for each trip component, resulting in a final weighted excess journey time at various levels of spatial and temporal aggregation (17).

The NACHs system measures the contribution of each incident and closure to *average* passenger delays through the use of models and look-up tables capturing a range of events (18). It is an important part of the evaluation of line upgrades being carried out under the Underground's

Public Private Partnership, and reflects the important contribution service disruptions play in the overall level of service.

The Oyster Smart Card, introduced by TfL in 2003, is a contactless fare media that supports London's complex zone- and time period-based fare structure. With its acceptance on all modes of the TfL network, it accounts for over 70% of all journeys in the system (11). The fact that the London Underground fare structure requires validation both upon entry and exit, as well as the high level of Oyster Card penetration, makes data from the Oyster system an ideal source from which to develop applications based on the proposed set of reliability measures. Measurements of total passenger travel times, however, experience a plus or minus 1 minute margin of error due to the truncation of seconds on the Oyster data timestamps in the current Oyster database (which will be updated in a new version of the Oyster system software).

Two four-week samples of morning peak period Underground journeys made on Oyster in 2007 were used for the following applications. Journeys with incomplete transactions (i.e. missing entry or exit validation), which amounted to 9% of all recorded trips, were excluded. In addition, the estimation methodology for the ERBT was validated using four weeks of incident-log data from the NACHs system, which included both the time and place of the event, as well as an estimated level of delays in units of hundreds of passenger hours, referred to as NAX units.

Characterization and Validation

The framework's potential for characterizing service reliability is illustrated through an application to one of the highest-volume AM Peak O-D pairs in the system: Waterloo-Canary Wharf, eastbound on the Jubilee Line.

The first step of the application classifies the peak period performance of each day in order to pool the journeys made under typical and incident-affected conditions and hence estimate a baseline performance. For journeys on this particular OD pair during the four weeks in February, a total of six weekdays were classified as being affected by severe disruptions. This is higher than the rate typically found for the largest 800 OD pairs in the system, which was 3-4 days out of the 20 days analyzed. Figure 3 shows the classification results for this OD pair. Those days classified as incident-affected are shown in green (or the lighter color).

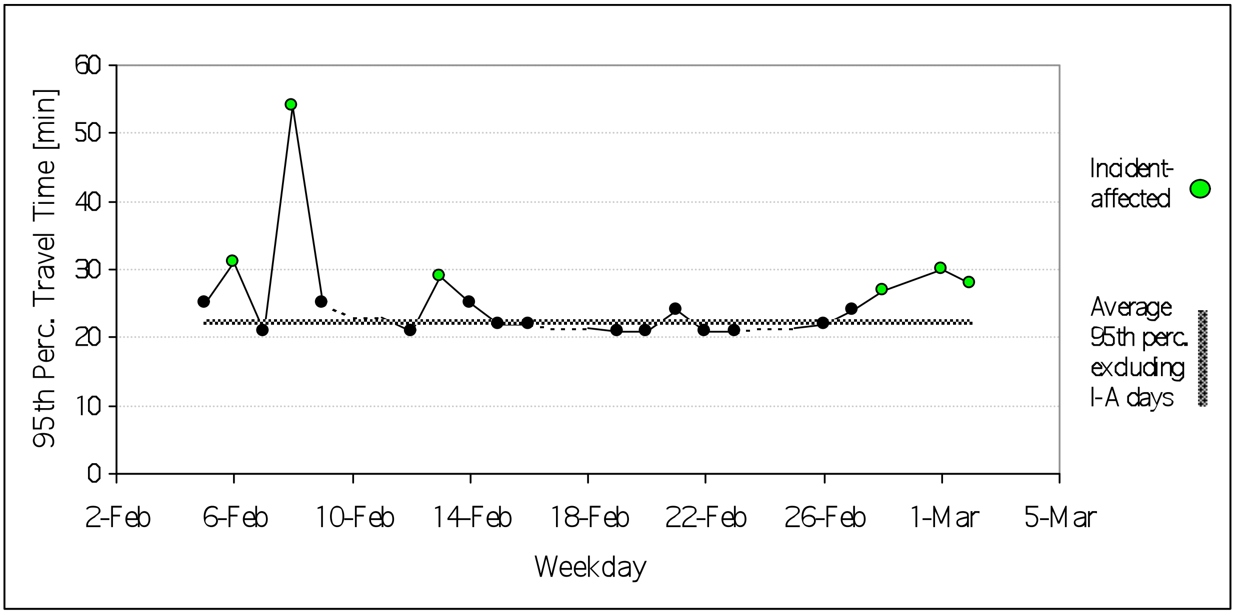

Figure 3: Classification of Daily Performance:
Waterloo to Canary Wharf – AM Peak, February 2007

One incident-affected period is shown in detail in Figure 4, which includes the travel time for every complete Oyster journey during the morning peak of February 13, 2007. Figure 4 indicates how passengers entering Waterloo station after 8:20 am experienced sudden large increases in travel time, which continued until the end of the morning peak. The incident-log for that day identified four incidents affecting the Jubilee Line AM Peak service as shown in Table 1, which includes the estimated passenger impact (in NAX units) of each event. There was indeed a major incident at approximately 8:26 am explaining the poor service quality after that time. Similar validations were achieved for all other typical and incident-affected classifications for all OD pairs analyzed.

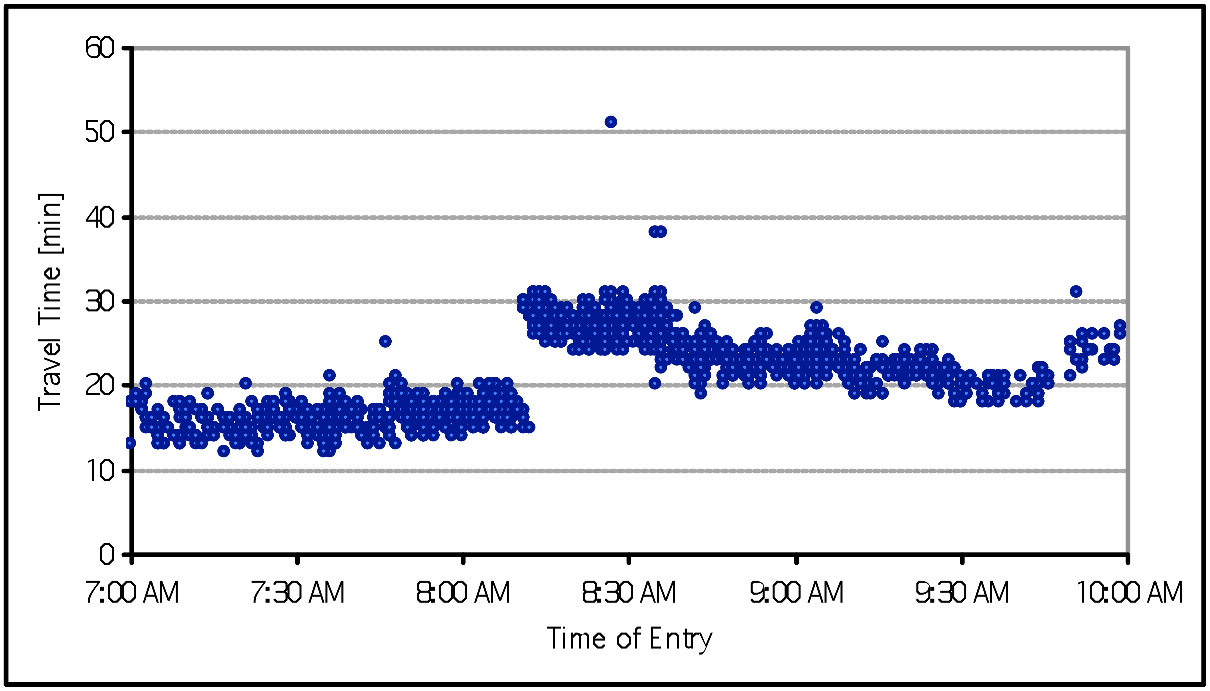

Figure 4: Individual Oyster journey times:
Waterloo to Canary Wharf – AM Peak, 13 Feb., 2007

Table 1: Incident-log for the Jubilee Line: AM Peak, 13 February, 2007

Date	Start Time	Location	Cause	Result	Indicative NAX
13/2/07	7:01 AM	Wembley Park	Fleet – Defective in Service	Train Delays	2.4189
13/2/07	8:06 AM	Canary Wharf	Customers – Crowding	Train Delays	3.5425
13/2/07	8:26 AM	North Greenwich	Track Power Failure	Train Delays	25.1754
13/2/07	9:32 AM	London Bridge	Customers – Disruption	Train Delays	2.1062

Having classified performance, the ERBT can be calculated using Equation 3 by comparing the typical and overall travel time distributions as shown in Figure 5. The buffer time under typical conditions is calculated at 5 minutes based on Equation 1, or 4 minutes less than the overall buffer time. This implies that passengers would have needed to budget into their schedules an additional (or *excess)* 4 minutes of buffer time to arrive on-time with 95% certainty at Canary Wharf station during February 2007. This excess buffer time can be largely attributed to the effects of incidents on service quality.

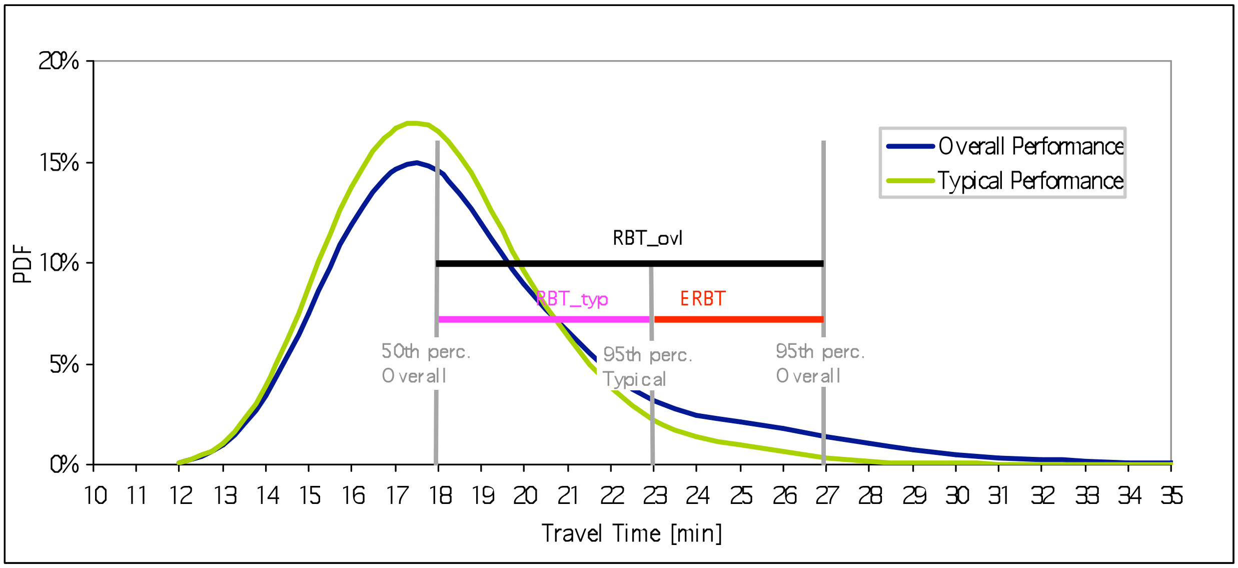

Figure 5: Overall and Typical travel time distributions: Waterloo to Canary Wharf – AM Peak, February 2007

Performance Monitoring and Passenger Information

One of the primary applications of the proposed framework is to monitor and compare service quality across different parts of the network and over time. This depends on the sample sizes available for various OD pairs in the system, themselves a function of the demand and the Oyster penetration rate. Based on an analysis of Victoria Line data, it was determined that all OD pairs satisfied the selected minimum sample size of 20 journeys over a four-week period during the morning peak to accurately measure the RBT. In order to estimate service reliability for the entire Victoria Line, the framework was first applied at the OD level, and the results aggregated as defined by Equations 2 and 4.

Table 2 shows the reliability buffer time results for the Victoria Line for two months in 2007. The baseline RBT is the same for February and November because the stability of typical conditions during the morning peak made it feasible to estimate the typical RBT using data pooled for both months. However, incident-related disruptions had a larger effect on reliability during the month of February, adding 3.62 minutes to the typical (baseline) buffer time. This represents a 73% increase in the amount of time passengers would need to budget to be sure of on-time arrival above that under typical conditions. For November, the corresponding figure was a 42% increase in the buffer time. In both cases though, the contribution of service disruptions to the total unreliability of the line is clearly appreciable and comparable to the contribution of all the remaining stochastic elements in service delivery.

Table 2: Reliability Buffer Time: Victoria Line, AM Peak, Feb and Nov 2007

	Reliability Buffer Time [min]		
Four-week Period	Typical (Baseline)	Overall	Excess
February 2007	4.93	8.55	3.62
November 2007	4.93	7.01	2.08

Another application of the proposed reliability measures enhances travel information provided to passengers through tools like TfL's web-based "Journey Planner" by complementing average performance figures with reliability estimates of the service. Specifically, the high level of resolution and coverage of Oyster data and the RBT metric can be used to supply information on the amount of buffer time that passengers need to budget in order to arrive on-time with 95% certainty alongside estimates of average or typical journey times.

The current Journey Planner system uses scheduled train times from station to station, and often fails to include the additional station access, egress, and platform wait time for an Underground trip when providing average journey time information. More importantly, however, current trip planning software like the Journey Planner fail to provide any information on the range of system performance that can be expected. By using distributions of travel times obtained from Oyster data, more accurate and more complete information can be provided to passengers in the form of the RBT. This would reduce the uncertainty for passengers making new or unfamiliar journeys (expected to be the majority of Journey Planner users) and help frequent passengers counteract the effects of incidents by increasing their chances of on-time arrival at their destination.

Service Quality Impact

While the RBT (and EBRT) metric complements the average performance captured by JTM, it is useful to compare reliability measurements with estimates of the average service quality of the system. For this analysis, the median Oyster travel time is split into platform wait time (PWT), access, egress and interchange time (AEI), on-train time (OTT), and the contribution of closures (CLRS) using the proportions of each as estimated based on JTM results for the Victoria line during February 2007. Each component of a trip is also weighted by the value of time attributed to it by passengers (also obtained from JTM) to estimate their individual contribution to the perceived travel time. These values are compared to the contribution of the RBT to overall perceived travel time using a value of 0.6 (relative to on-train travel time), which is at the low end of the range of estimates found in the literature where values as high as 1.4 are reported (6).

Figure 6 shows this comparison where, in addition to the reliability buffer time at the line level, the journey components used in JTM are shown. The left pie-chart shows the line-level reliability buffer time of 8.55 minutes and the median total actual journey time of 16.71 minutes. The center pie-chart applies the JTM proportions of the components of the total journey time to the median journey time. Finally, the right pie-chart shows that at the assumed value of reliability buffer time of 0.6, unreliability contributed around 16% of total perceived journey time. This is comparable in magnitude with other components of the trip, such as the platform wait time. About 7% of the 16% can be attributed to the effects of disruptions. From the perspective of the operator, this means that reliability can be improved by up to 40% and the total perceived travel time for this particular line by up to 7%, if incidents could be completely eliminated or substantially reduced.

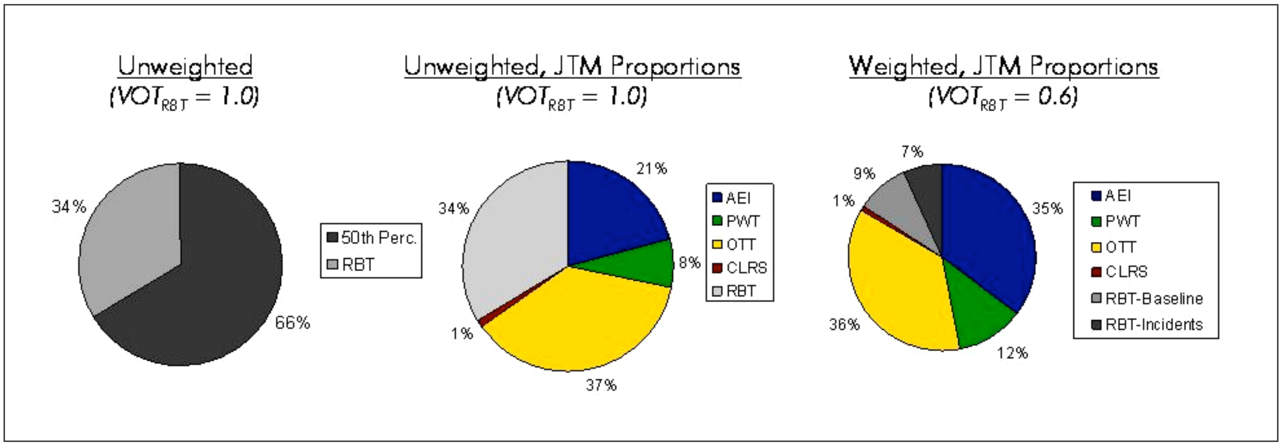

Figure 6: Contribution of JTM journey components and RBT to total journey time, unweighted and weighted by value of time: Victoria Line – AM Peak, Feb. 2007

Conclusions

Specific to the three objectives stated at the outset, this paper illustrates how data from transit Smart Cards can be used to quantify an important yet complex aspect of service quality such as reliability. Moreover, reliability measurements are extended to show how they can be used as part of routine performance monitoring efforts, as an input into the evaluation of the service, and to gain insights into the causes of unreliability and their contribution to overall service quality.

At a broader level, however, three important implications can be derived from this research. First, this research shows how, using the particular case of the London Underground as an example, the current focus on average performance by transit providers does not take into account the important impact of reliability on passenger perceptions of service quality, leading to an undervaluing of improvements or problems in this area. Second, from this study it is possible to gain insight into the important role that non-recurring disruptions play in determining the level of reliability passengers experience over time, compared to the contribution of recurring variability of the operation under typical conditions, which might alter ongoing efforts by operators to improve reliability. Lastly, an initial attempt at identifying and measuring the impact of these non-recurring disruptions was included here, illustrating the potential for using AFC data not only for service quality monitoring exercises, but also for detailed analyses of specific service disruptions.

These implications also suggest areas of research where future studies can focus to increase our understanding of transit service reliability. An immediate extension of this research involves extending the two performance categories presented here to study additional causes of unreliability from the passenger perspective. This could include the effects of operations control interventions on service quality, or more simply the contribution of various service characteristics such as length of line and service frequency on travel time variability. Longer term studies based on this work could focus on using the measurement methodology presented here to calibrate existing reliability measures that use supply-side automated data (such as headway and vehicle travel time estimates derived from signal or AVL systems), benefiting a wider array of transit agencies that do not collect AFC data with entry- and exit-validation. Finally, the ability to cost-effectively and accurately measure reliability as experienced by passengers could be used as an input into passenger choice models, which are important tools for transit service planning.

Acknowledgements

This research was made possible due to the generous support of Transport for London, who provided the authors not only with the data and resources necessary to complete the study but also their time and expertise.

Bibliography

(1) Abkowitz, M., Slavin, H., Waksman, Englisher, L., and Wilson, N. (1978). *Transit Service* Reliability, Report UMTA-MA-06-0049-78-1. USDOT Transportation Systems Center. Cambridge, MA.

(2) Strathman, J., Kimpel, T., Dueker, K. (2000) *Time Point-Level Analysis of Passenger Demand and Transit Service Reliability*. Final Technical Report TNW2000-03, Portland State University.

(3) Paine, F. T., Nash A. N., Hille S. T., Brunner, G. A. (1976) *Consumer Attitudes Toward Auto vs. Public Transport Alternatives*. Journal of Applied Psychology, Vol. 53 (6), pp. 472-480

(4) Bates, J., Polak, J., Jones, P., Cook, A. (2001) *The Valuation of Reliability for Personal Travel*. Transportation Research (Part E), Vol. 37, No. 2, pp. 191-229.

(5) Frumin, M., Uniman, D., Wilson, N.H.M., Mishalani, R.G., Attanucci, J.P. (2009) *Service Quality Measurement in Urban Rail Networks with Data from Automated Fare Collection Systems*. Presented at the Conference on Advanced Studies for Public Transport, 2009. Hong Kong.

(6) Lam, T., Small, K. (2001) *The Value of Time and Reliability: Measurement from a Value Pricing Experiment*. Transportation Research (Part E) Vol. 37, pp. 231-251.

(7) Love, A., Jackson, P. (2000) *The Meaning of Reliability*. Qualitative Research Commentary Report S.00756, London Buses, UK.

(8) Lomax, T., Schrank, D., Turner, S., Margiotta, R. (2003) *Selecting travel time reliability measures*. Texas Transportation Institute and Cambridge Systematics, Inc.

(9) De Jong, G., Kroes E., Plasmeijer, R., Sanders, P., Warffemius, P. (2004) *The value of reliability*. Proceedings of the European Transport Conference – Strasbourg, 4-6 October, 2004.

(10) Furth, P., Muller, T. (2006) *Service Reliability and Hidden Waiting Time*. Transportation Research Record 1955, pp. 79-87.

(11) Chan, J. (2007). *Rail OD Estimation and Journey Time Reliability Metrics Using Automated Fare Data*. Thesis, Master of Science in Transportation, MIT.

(12) Uniman, D. (2009) *Service Reliability Measurement using Smart Card Data: Application to the London Underground*, Master's thesis, Massachusetts Institute of Technology.

(13) Draper, N., Smith, H. (1998) *Applied Regression Analysis*. 3rd Edition, John Wiley & Sons.

(14) Wolberg, J. (2006), *Data Analysis using the method of least squares: extracting the most information from experiments*, Springer, New York.

(15) TfL – Group Business Planning and Performance (2007) *TfL Business Plan 2005/6 to 2009/10*. Transport for London, UK.

(16) TfL – Transport Planning Business Operations (2007b) *London Travel Report 2007*. Transport for London, UK.

(17) TfL – Group Transport Planning and Policy (2006) *Transport 2025: Transport vision for a growing world city*. Transport for London, U.K.

(18) London Underground Limited (2004), Modelling and performance tools. retrieved from LU Intranet on 10 August, 2006.