A Framework for Measuring Passenger-Experienced Transit Reliability Using Automated Data

by

Daniel Alden Wood

B.S., Mechanical Engineering,
University of Pennsylvania, 2013

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

© 2015 Massachusetts Institute of Technology. All rights reserved.
A Framework for Measuring Passenger-Experienced Transit Reliability
Using Automated Data
by
Daniel Alden Wood

Submitted to the Department of Civil and Environmental Engineering
on May 21, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

Abstract

A public transport operator’s ability to understand and improve service reliability experienced by passengers depends upon their ability to measure it. Traditionally, largely due to data collection limitations, passengers’ experience of reliability has not been measured directly. As a result, operators often fail to effectively measure, and thus manage, key issue affecting passengers’ perceived reliability. However, with the relatively recent availability of automatic data collection (ADC) systems, it has become technically feasible to measure passengers’ reliability experience at a detailed level. If used in practice, passenger-experienced reliability measurement has the potential to improve public transport systems’ responsiveness to passengers needs across many functional areas.

This thesis develops a framework for the understanding and practical use of passenger-experienced reliability measurement on high-frequency transit systems. A model of passenger-experienced reliability based on total travel time variability is developed, and the key differences from “operational” reliability identified. This model is applied to identify public transport management functions which should be targeted as a result of passenger-experienced reliability measurement. The model and potential applications are then synthesized to develop a set of design criteria for passenger-experienced reliability metrics.

Two new measures of passenger-experienced reliability are developed, both aiming to quantify the “buffer time” passengers must add to compensate for travel time variability. The first measure, derived from passengers’ actual travel times from automatic fare collection (AFC) data, is essentially the median travel time variability experienced by frequent travelers over each origin-destination (OD) pair of interest. The second measure, derived from vehicle location data, OD matrices, and train load estimates, is based on a simulation of passengers’ waiting, boarding, transfer, and in-vehicle travel process. This second metric is aimed at “non-gated” systems without exit AFC data, for which passengers’ travel times cannot be measured directly.

These two metrics are tested and evaluated using data from the Hong Kong MTR system. These metrics’ response to incidents, scheduled headways, and passenger demand are tested at the OD pair and line levels. The results are used to evaluate these metrics according to the previously-developed design criteria for passenger-experienced reliability metrics. The first metric is found to be suitable for implementation (where adequate data is available), while the second is found to inadequately measure demand-related delays.

An implementation guide for the AFC-based metric is developed. This guide is structured around four main implementation decisions: (1) coordination with an operator’s
existing metrics, (2) defining the service scope, (3) determining an appropriate frequency of calculation, and (4) defining appropriate time of day intervals and date periods. This guide is then demonstrated using a case study application from MTR: an investigation of the 2014 Hong Kong political demonstrations’ impact on MTR reliability.

Thesis Supervisor: Nigel H. M. Wilson
Title: Professor of Civil and Environmental Engineering

Thesis Supervisor: Haris N. Koutsopoulos
Title: Professor of Civil and Environmental Engineering
Northeastern University
Acknowledgments

This thesis, and the two years of work it represents, were only possible with the help of many people whom I would like to acknowledge:

I would like to extend my deepest thanks to my advisors, Nigel Wilson and Haris Koutsopoulos, for their valuable guidance, support, and encouragement. Their drive towards excellence has not only aided my research, but made me a stronger professional. Furthermore, our challenging and rewarding discussions have broadened my perspective on transportation, and have confirmed my decision to pursue a career in it.

MTR has generously made this research possible. I want to especially thank Pandora Leung, Felix Ng, and Henry Lo for their valuable assistance and feedback, as well as their efforts in making our visits to Hong Kong possible. Thank you as well to Jackie Cheung, whose advice has been valuable in shaping this research to MTR’s needs.

Thank you to Kiley, Ginny, Shannon, and all of the CEE staff for your dedication to students, and for keeping things running in the MST program and department.

To my friends at MIT—you have been the best “comrades in arms” one could ask for. You have made the past two years, despite their challenges, some of the best so far. Thank you also to Mike and Jay for your friendship, and providing a welcome escape from the “MIT bubble”.

Finally, I would like to thank my family for their tireless love, support, and encouragement along the way. Without you, none of this would have been possible.
# Contents

1 Introduction ........................................... 15
   1.1 Motivation ........................................... 16
   1.2 Research Objectives ................................ 18
   1.3 Research Approach .................................. 19
   1.4 Introduction to the MTR ............................. 20
   1.5 Thesis Structure ..................................... 21

2 Literature Review on Reliability ......................... 23
   2.1 Overview of Operational Reliability ............... 24
      2.1.1 Timetable Adherence Measures ................. 24
      2.1.2 Headway-Based Measures ....................... 25
   2.2 The Passenger Experience of Reliability .......... 25
      2.2.1 Direct Surveys on Reliability ................. 26
      2.2.2 Empirical Models ............................... 26
      2.2.3 Theoretical Model of Passenger Reliability ...... 27
   2.3 Comparison of Passenger and Operational Reliability ... 30

3 Uses and Benefits of Reliability Measures .............. 33
   3.1 Reliability Management ................................ 33
      3.1.1 Overview of Reliability Management .......... 33
      3.1.2 Benefits of Passenger-Experienced Reliability Measures ...... 35
   3.2 Passenger Information ................................ 37
   3.3 Reporting to Stakeholders ........................... 38
   3.4 Cost-Benefit Analysis ............................... 40
   3.5 Long-Range and Regional Planning ................ 40
7.4 Time Period Setting

7.4.1 Timeframe of Reliability Factors of Interest

7.4.2 Rate of Variation in Reliability Factors

7.4.3 Frequency of Information Update

7.4.4 Target Audience

7.4.5 Frequent Traveler Group Size Constraints

7.5 Case Study: Evaluating Effects of Political Protests

7.5.1 Background on Effects on MTR Service

7.5.2 IRBT Implementation

7.5.3 IRBT Results

7.5.4 Incident Results and Causal Analysis

8 Conclusions

8.1 Research Summary and Findings

8.2 Recommendations

8.3 Future Research
List of Figures

1-1 Map of MTR System .................................................. 22

3-1 Proposed presentation of reliability information in Journey Planner\textsuperscript{1} ........ 38

4-1 The likelihood of delays as a function of RBT upper percentile ............... 53
4-2 Journey time PDFs for 3 bus trips\textsuperscript{2} .................................. 56

5-1 Representative IBT Distribution ........................................... 68
5-2 Wan Chai to Quarry Bay RBT and IRBT (Island Line) ......................... 73
5-3 Kowloon Bay to Lam Tin RBT and IRBT (Kwun Tong Line) ................. 73
5-4 Wan Chai to Jordan RBT and IRBT (Transfer) ................................ 74
5-5 Island Line RBT and IRBT .................................................. 75
5-6 Kwun Tong Line RBT and IRBT ............................................. 76
5-7 Island Line to Tsuen Wan Line Transfer OD pairs, RBT and IRBT ........ 76
5-8 IRBT and Major Incidents, Island Line WB OD Pairs ......................... 78
5-9 IRBT and Major Incidents, Island Line WB and EB ........................... 78
5-10 IRBT and Major Incidents, Kwun Tong Line WB and EB .................... 78
5-11 Island Line WB, Half-Hour Periods ........................................ 80
5-12 Tsuen Wan Line NB, Half-Hour Periods .................................... 80
5-13 ISL WB Excess IRBT, Average Passenger Flow (at Peak Load Point), and Passenger Capacity ................................................... 83
5-14 TWL NB Excess IRBT, Average Passenger Flow (at Peak Load Point), and Passenger Capacity ................................................... 83
5-15 ISL WB Excess IRBT and Capacity Utilization (at Peak Load Point) ..... 84
5-16 TWL NB Excess IRBT and Capacity Utilization (at Peak Load Point) ..... 84
5-17 ISL 8-9 AM IRBT and intra-line demand, by direction ...................... 86
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-18</td>
<td>OD pair IRBT, Normal and for Limited Passenger Sample Size</td>
<td>89</td>
</tr>
<tr>
<td>5-19</td>
<td>Line-level IRBT, Normal and for Limited Passenger Sample Size</td>
<td>90</td>
</tr>
<tr>
<td>6-1</td>
<td>Admiralty to Fortress Hill</td>
<td>104</td>
</tr>
<tr>
<td>6-2</td>
<td>Transfer Arrival Rate $\lambda_k(t)$</td>
<td>108</td>
</tr>
<tr>
<td>6-3</td>
<td>Line-level PPRBT and IRBT, and Avg. Line Flow, by time-of-day</td>
<td>121</td>
</tr>
<tr>
<td>6-4</td>
<td>Line-level PPRBT and IRBT, for rolling 6-week periods</td>
<td>121</td>
</tr>
<tr>
<td>6-5</td>
<td>OD PPRBT and IRBT, and Avg. Line Flow, by time-of-day</td>
<td>123</td>
</tr>
<tr>
<td>6-6</td>
<td>Avg. Line Flow for each OD leg, by time-of-day</td>
<td>123</td>
</tr>
<tr>
<td>6-7</td>
<td>OD PPRBT and IRBT, for 6-week rolling date periods</td>
<td>124</td>
</tr>
<tr>
<td>7-1</td>
<td>KTL EB IRBT for 15-min, 30-min, and 60-min intervals and Pax. Flow</td>
<td>144</td>
</tr>
<tr>
<td>7-2</td>
<td>Map of areas affected by protests (in red)</td>
<td>148</td>
</tr>
<tr>
<td>7-3</td>
<td>Change in Average Weekday Journeys, by Half Hour</td>
<td>149</td>
</tr>
<tr>
<td>7-4</td>
<td>Peak Half-Hour Line Flow (at Peak Load Point), by Line</td>
<td>150</td>
</tr>
<tr>
<td>7-5</td>
<td>Peak Half-Hour Capacity Utilization (at Peak Load Point), by Line</td>
<td>150</td>
</tr>
<tr>
<td>7-6</td>
<td>Line-direction IRBT Results, Before and During Protests</td>
<td>154</td>
</tr>
<tr>
<td>7-7</td>
<td>Transfer Pattern IRBT Results, Before and During Protests</td>
<td>155</td>
</tr>
<tr>
<td>7-8</td>
<td>Change in number of incidents, Base Period to Protest Period</td>
<td>158</td>
</tr>
</tbody>
</table>
# List of Tables

5.1 RBT–IRBT Difference Statistics 75
5.2 Root-Mean-Square Deviation and Correlation between OD-pair IRBT and Sampled OD-pair IRBT 90
5.3 Root-Mean-Square Deviation and Correlation between Line-direction IRBT and Sampled Line-direction IRBT 91
5.4 Percentage of Island Line WB OD pair IRBTs meeting min. group size 93
5.5 Line-direction group size, green meeting min. line group size 93
5.6 Percentage of Passenger Trips Represented by Island Line WB OD pair IRBTs meeting min. group size 94
7.1 Time Period Decision Factors Summary 146
7.2 Ridership Change, from Base Period (8/18-9/26) to Protest Period (9/29-11/7) 149
7.3 Time-of-Day Intervals for IRBT Calculation 155
7.4 Peak Half-Hour Average Passenger Flow (PPHPD) 157
7.5 Peak Half-Hour Capacity Utilization 157
8.1 Metric Application Strengths and Limitations 168
8.2 Metric Application Recommendations 169
Chapter 1

Introduction

“What gets measured, gets managed.” —Peter Drucker

This well-known maxim captures the essence of what reliability measurement means to the transit agency. Transit agencies, like all large organizations, manage and improve their services through the monitoring and collection of a set of performance metrics. Thus, a transit agency’s set of reliability metrics should effectively capture all important aspects of transit service reliability. While conceptually simple, in practice designing and implementing such measures is a challenge, given the complexity of transit operations, and the limits of information available to the transit agency.

The introduction of Automatic Data Collection (ADC) systems in the transit industry, over the past several decades, has greatly enhanced the depth and quality of information operators can gather about their service reliability. Information on vehicle running times and passenger journey times, once obtainable only through imprecise and costly manual sampling, is now available at a detailed, disaggregate level with data from Automatic Vehicle Location (AVL) and Automatic Fare Collection (AFC) systems. This data, unlike traditional manually-sampled data, is able to directly capture system reliability as experienced by customers. However, new analytical methods are required to translate this “big data” into reliability metrics that are useful for transit agencies, passengers, and other stakeholders.

This thesis has two primary aims. First, to derive from this automatic data new reliability metrics that effectively communicate the passenger’s experience of service reliability. These metrics are meant to address weaknesses of previously-developed ADC-based reliabil-
ity metrics, by Frumin [1], Uniman [2], and others. Second, to develop a practical guide for public transport operators on implementing these and similar “customer-centric” reliability metrics, in a manner effecting tangible improvements to operators’ reliability management applications. These objectives are largely pursued in the context of the Hong Kong Mass Transit Railway subway system, known as MTR, whose data is used to test and evaluate the new metrics, and whose reliability measurement needs are used to demonstrate the metric implementation guide.

1.1 Motivation

This study’s main motivation is the need for public transport operators to transition from a strictly operational view of reliability, to one incorporating passengers’ direct experiences of reliability. Reliability from the operator’s perspective is generally expressed in terms of train (or bus) adherence to an operating plan, i.e. a timetable or operating frequency. This “operational” reliability is assessed using performance measures such as on-time performance, or headway regularity. Most large transit agencies have decades of experience measuring reliability using and interpreting these types of measures. Furthermore, such measures can be calculated without large sample sizes of operational data, thus suiting them to calculation via manual data collection—the only method available before the recent introduction of ADC systems.

Reliability as experienced by passengers is very different. Studies [3] have shown that, for most passengers, reliability is seen as the likelihood of on-time arrival, for a particular journey between a specific origin-destination (OD) pair, at a given time of day. Included in this reliability experience are delays from sources such as denied boardings (a.k.a. left-behinds) and congestion in stations, factors not captured by operational reliability measures, which capture only bus (train) delays. Reliability metrics, such as the reliability buffer time [4, 2], have been developed by researchers and transit agencies to measure this passenger experience of reliability. These proposed metrics require very large sample sizes of operational and passenger travel time data, now available from AFC and AVL systems.

However, the adoption of such measures by the transit industry has been limited. This may be due to the greater difficulty of implementing passenger-oriented reliability measures, relative to operational measures. Besides their larger data requirements, which have
largely been addressed with the introduction of ADC systems and more computing power, passenger-focused metrics are also more challenging to analyze than their operational counterparts, often requiring more interpretation to yield actionable information about system performance.

This thesis addresses this need. If new measures of passenger reliability are developed that provide more actionable information, and are easier to implement, public transport operators may be more willing to adopt passenger reliability measures as part of their performance measurement programs. Achieving this would be a great benefit to the public transport industry, as passenger-focused reliability measures can improve a transit agency’s capabilities in a variety of ways, including:

- **Better measure (and manage) effects of overcrowding:** Overcrowding in public transport systems can produce significant delays for the individual passenger; e.g., denied boardings. These delays are difficult to measure with operational reliability measures, which focus on vehicle performance. Passenger-focused metrics can measure these delays, most of which occur during the “out of vehicle” portions of the passenger journey. With such information, public transport operators can better identify overcrowding problems when they arise, and plan strategies to address them.

- **Better evaluate reliability improvements:** With passenger-focused reliability metrics, an operator can evaluate the effects of system improvements (e.g., new signaling system) directly in terms of improvement to passengers’ experience of reliability, rather than indirectly through improvements to train schedule adherence. This better aligns the evaluation with an operator’s overall objective, to provide as good a passenger experience as possible within budget constraints.

- **Improved stakeholder communication:** Reporting passenger-focused reliability metrics could help a transit agency’s stakeholders from non-technical backgrounds (e.g., elected officials) better understand the reliability performance of the agency’s service. This, of course, requires metrics that not only capture the passenger view of reliability, but present it in an intuitive manner.

- **Improved passenger information:** By calculating an appropriate passenger-oriented reliability measure for each OD pair (a level of detail difficult with operational reliability measures) in a transit network, journey planners can be enhanced to include
information about the reliability of a particular journey, in a manner allowing passengers to more effectively plan their journeys.

- **Long-term planning:** Passenger-experienced reliability measures can be used to support long-term planning, such as identifying critical infrastructure needs, or analyzing the geographic and/or demographic equity of service reliability.

## 1.2 Research Objectives

The overarching objective of this thesis is to present a comprehensive framework for effectively incorporating passenger-experienced reliability measurement into public transport system management, from a conceptual understanding of passengers’ reliability experience to the practical considerations of using new performance measures for specific applications. This framework is intended to convey not only how passenger-experienced reliability measures can be implemented, but, just as importantly, why an operator should expend the cost and effort to do so. To achieve this, this thesis sets out to achieve the following objectives:

1. **Develop model of the passenger’s experience of reliability:** Through a review and synthesis of the literature on passengers’ opinions on and response to transit service reliability, develop a workable model of how transit passengers experience reliability, and identify the key differences between passenger-experienced and operational reliability.

2. **Identify key applications for passenger-experienced reliability measurement:** Identify how measuring passengers’ experience of reliability can improve a transit operator’s key reliability measurement applications.

3. **Define design criteria for passenger-experienced reliability metrics:** Develop a set of design criteria for potential passenger-experienced reliability metrics. Such criteria, if met, should produce metrics that accurately capture the passenger’s experience of reliability (as described by the derived model), and are suitable for the identified practical applications of passenger-experienced reliability measurement.

4. **Develop two new reliability metrics:** Develop two new passenger-experienced reliability metrics that overcome the main shortcomings of previously-developed passenger-experienced reliability metrics. The first metric is based on passengers’ actual travel
times, from “closed” AFC data of the type available on MTR. The second metric does not require passenger’s actual travel times, thus making it suitable for “open” systems (i.e., only an entry transaction).

5. **Test and evaluate new metrics:** Using data from the MTR system, demonstrate the calculation of the new metrics, and test their behavior with respect to external factors (e.g., passenger demand, incidents, etc.). Based on the results, evaluate the metrics with respect to the previously-developed design criteria. Identify the metric most promising for implementation on the MTR system.

6. **Develop implementation framework:** Develop a framework for practically implementing the recommended new metric, for a given reliability measurement application.

### 1.3 Research Approach

The specific research approach for the development, testing, and evaluation of new reliability measures and the development of the implementation framework is described below.

**Development of new reliability measures:** The two new reliability metrics proposed in this thesis are based on the Reliability Buffer Time (RBT), a metric developed by Chan [4] and Uniman [2]. Like the RBT, these measures estimate the typical ext “buffer time” a passenger should budget for their journey to ensure an acceptable likelihood of on-time arrival, due to service unreliability—as unreliability increases, so should the buffer time required.

The first metric, termed the Individual RBT, or IRBT, is derived, like Chan and Uniman’s RBT, from passengers’ travel times inferred from AFC entry and exit times. The IRBT is, essentially, the average of the buffer times experienced by frequent travelers on a particular OD pair. This method aims to significantly reduce the “noise” in the RBT due to behavior variation across passengers. The second metric, the Platform to Platform RBT, or PPRBT, is derived by simulating each stage of every passengers’ journey, based on AVL data, OD matrices, and route choice data. From the simulation results, the typical passenger’s buffer time is estimated. This approach extends the AVL-based RBT research of Ehrlich [5] and Schil [6].
Testing and evaluation of reliability measures: The two metrics are tested and evaluated differently. The IRBT is tested by examining its sensitivity to several key external factors: the presence of incidents, the scheduled headway, and passenger demand; limitations to the calculation of the IRBT are also described. The PPRBT, on the other hand, is mainly tested by comparing PPRBT results with corresponding IRBT results. Because the two metrics are theoretically estimating the same buffer time, their results should be similar. The PPRBT, then, is evaluated largely with respect to how well its model of passenger experience can replicate the actual experience seen in the IRBT.

Framework for Implementation: The metric implementation framework is comprised of three main parts: (1) integration with an operator’s existing metrics, (2) setting an appropriate spatial scope, and (3) setting appropriate time periods for calculation. This framework is then applied to investigate the effects of the Fall 2014 political demonstrations in Hong Kong on MTR’s reliability.

1.4 Introduction to the MTR

The Mass Transit Railway, referred to as MTR, is the rapid transit rail system in Hong Kong. MTR operates 9 metro lines (serving 87 stations), a 36 km light rail system, and an airport express train in Hong Kong, as shown in Figure 1-1. Together, these systems have an average weekday ridership of over 5 million, with over 4.6 million served by the MTR heavy rail metro system alone [7]. The MTR metro ridership occurs over a small (compared to peer agencies like London, New York, etc.) 175 km network, resulting in MTR’s ridership per route-km being among the highest in the world\(^1\). MTR’s services comprise only the rail portion of Hong Kong’s public transport network. Public bus service in the city is operated by five private bus companies, which together serve about 3.9 million passenger trips per day (a handful of feeder bus routes are operated by MTR) [9].

To serve this high level of demand, MTR operates high service frequencies on all of its lines, with headways as short as 1.9 minutes during the peak hours, and 4-6 minutes during off-peak hours [10]. Even with such high service frequencies, train overcrowding and denied boardings have become a serious public concern in the face of increasing passenger demand, with demand regularly exceeding 90% of capacity on most of MTR’s busiest lines [11].

\(^1\)9.16 million/route-km/yr. For reference, this is about twice that of the New York City Subway [8]

20
is making serious efforts to address this issue, such as a 25% fare discount from 7:15-8:15 AM for passengers exiting at 29 designated core urban stations, to encourage passengers to shift their journeys away from the peak hour (8:15-9:15 AM)[12]. However, MTR’s main reliability metrics in use—Passengers Affected Ratio, running time adherence, and number of incidents ≥X minutes—do not completely capture the effects of demand-related delays (e.g., denied boardings) on passengers.

Fares on the MTR system are distance-based, and are collected almost entirely through MTR’s Octopus smart card system, one of the first in the world to be implemented on a metro system. Passengers pay their fares by “tapping-in” at entry faregates at their origins, and “tapping-out” at exit faregates at their destination. These transactions are recorded, and constitute the bulk AFC data that is used in this thesis to derive information about passengers’ travel times. MTR also produces high-quality AVL data, used extensively in the research presented in this thesis.

1.5 Thesis Structure

Chapter 2 reviews the literature on operational and passenger-experienced reliability, to identify the key characteristics of passenger-experienced reliability, and the main differences between it and operational reliability. Chapter 3 reviews the potential uses and benefits of passenger-experienced reliability measures. Chapter 4 develops a set of design criteria for passenger-experienced reliability metrics. This chapter also reviews previously-developed passenger-focused reliability metrics, and assesses them according to the developed framework. Chapters 5 and 6 develop, test, and evaluate the two new passenger-experienced reliability measures, the IRBT and PPRBT. The specifics of how each measure is calculated are defined, and application results from the MTR network presented. Chapter 7 develops a framework for implementing the IRBT (the recommended metric for MTR), and demonstrates this framework using an analysis of the 2014 Hong Kong political demonstrations as a case study. Chapter 8 summarizes the conclusions of the thesis research, and identifies opportunities for future research.
Figure 1-1: Map of MTR System
Chapter 2

Literature Review on Reliability

The object of this thesis is the measurement of transit service reliability, a concept best defined in the seminal Transit Reliability Study [3] as “the invariability of service attributes which influence the decisions of travelers and transportation providers”. As Uniman [2] notes, this definition makes two important distinctions about transit service reliability.

First, it defines reliability in terms of a provided service; for transit, the service is mobility for passengers. Thus, service reliability is the reliability of passengers’ transit journeys. This is different than the reliability of equipment and infrastructure, which does not relate directly to movement of passengers, and is only one of many inputs to transit service. This distinction is generally observed in public transport operators’ reliability metrics; separate measures are used for service reliability (e.g., on-time performance) and equipment reliability (e.g., mean distance between failures for rolling stock).

Second, and more importantly for this thesis, this definition distinguishes between the passenger’s perspective, and the operator’s. The main focus of this chapter is to show, through a review of literature on transit service reliability, how these two viewpoints lead to two distinct aspects of reliability, with very different characteristics: “operational reliability”, and “passenger-experienced reliability”. Section 2.1 reviews how service reliability has been defined and measured from the operator’s perspective. Section 2.2 reviews studies of how reliability is experienced by passengers. Section 2.3 summarizes the key differences between these two views of service reliability.
2.1 Overview of Operational Reliability

Transit operators generally view service reliability as vehicle trips’ performance relative to the schedule [3, 13], which can be either a timetable, or a scheduled headway (e.g., headway at Stop A will be 5 minutes). Two explanations can be given for this practice. First, almost all aspects of transit operations—dispatching of buses and trains, availability of vehicles, assignment of crews to vehicles, recovery from service disruptions—are planned and executed in accordance with a defined schedule, at the route or line-level. Thus, reliability in terms of vehicle runs and timetables can be easily interpreted and acted upon by transit operators’ staff.

Second, until recently, transit operators’ only detailed source of service performance data has been recorded vehicle arrival and/or departure times at various points along the route. Headways, departure times, and running times for vehicle trips can be readily derived from this data, and compared with the schedule. Inferring information about passengers’ journey times, however, is not straightforward with this data. Thus, operational measures have traditionally been much easier for operators’ staff to calculate.

Operational measures of service reliability can be categorized into two groups: measures of timetable adherence, and measures of headway adherence [14].

2.1.1 Timetable Adherence Measures

Timetable adherence reliability measures are based on the difference between actual and scheduled vehicle departure, arrival, or running times. The most common is on-time performance (OTP), the percentage of departures or arrivals at a given location (or “timepoint”) within a specified time window, relative to the scheduled time. This window varies across operators; MBTA uses 0 to 5 minutes late, while London Buses uses 2.5 minutes early to 5 minutes late [1]. OTP is often measured at a single timepoint for a given route or line, such as the terminal or the most heavily-used stop or station. In other systems, OTP is averaged across multiple timepoints (sometimes referred to as en-route adherence, or ESA), providing a more complete picture of route performance.

Another timetable-based measure is running time adherence, the percentage of terminal-to-terminal running times within a time window around the scheduled time. This is one of MTR’s main reliability measures; MTR considers train trips “late” if their running time
exceeds the schedule by 2 minutes or more.

2.1.2 Headway-Based Measures

For high frequency services, often defined as headways of 10 minutes or less [15], passengers are assumed to arrive randomly at stops, rather than coordinating their arrival with published departure times. These passengers’ wait times, and thus their wait time reliability, will be a function of the operated headways, rather than timetable adherence. Recognizing this, many transit agencies measure their high-frequency services’ reliability in terms of headway regularity.

The simplest, and most commonly used headway regularity measure is the percentage of headways within some absolute or relative deviation around the scheduled headway. For example, the MBTA uses the percentage of headways less than 150% of scheduled headway. Eight of the nine members of the International Bus Benchmarking Group (IBBG) using headway regularity indicators use similar definitions [15].

The other noteworthy headway regularity metric used in practice is the Excess Wait Time (EWT), used by London Buses. The EWT estimates the difference, for a given service location, between the average passenger wait time, and the average passenger wait time if all buses had arrived on schedule. Both the average actual and scheduled wait time are calculated using the following formula for average waiting time \( E[w] \) [16]:

\[
E[w] = \frac{1}{2} E[h] \left( 1 + cov^2(h) \right)
\] (2.1)

where \( E[h] \) is the average headway, and \( cov(h) \) is the coefficient of variation of headways. Thus, the EWT is defined as:

\[
EWT = \frac{1}{2} E[h_{actual}] \left( 1 + cov^2(h_{actual}) \right) - \frac{1}{2} E[h_{schedule}] \left( 1 + cov^2(h_{schedule}) \right)
\] (2.2)

2.2 The Passenger Experience of Reliability

This section reviews the literature on how transit service reliability is experienced by passengers. Section 2.2.1 describes findings from direct passenger surveys about reliability (i.e., asking passengers directly what they think about service reliability). Section 2.2.2 describes empirical studies on service reliability. Section 2.2.3 proposes a detailed model of
passenger-experienced reliability, based on a synthesis of the theoretical work this topic.

2.2.1 Direct Surveys on Reliability

Probably the most in-depth passenger surveys about reliability are two surveys conducted by London Transport (predecessor to Transport for London); one by Love and Jackson for London Buses [17] (as cited in Uniman [2], and the other done by London Underground [18]. In both surveys, passengers were asked their general feelings on reliability.

An important survey finding was that passengers’ perception of transit reliability is very subjective. Unreliability was something passengers “felt”, based on their remembered reliability experiences, a “data set...subconsciously re-calibrated during each moment of travel”, and influenced by their feelings about other aspects of service quality, such as comfort. Passengers’ assessment of transit reliability also depended on their trip purpose and sensitivity to late arrivals. For example, commuters were usually much more concerned with reliability than “leisure” travelers. The surveys also found a variety of passenger behavior in response to unreliability. Some passengers adjusted their departure times, some chose different routes or modes, and some did not change their behavior at all, depending on their perception of reliability, value of on-time arrival, and available route choices.

Given such subjectivity and behavior variation in passengers’ responses to reliability, it is clear that developing a workable, quantitative model of passenger-experienced transit reliability based on direct survey results is infeasible. However, survey findings can be used to validate passenger reliability models developed by other, more analytical methods.

2.2.2 Empirical Models

Several studies [19, 20] have used stated preference surveys to compare two potential mathematical models of the disutility of “unreliability”—“mean-variance” and “schedule delay”. In the “mean-variance” model, service reliability is represented explicitly by a travel time variance factor, as in the following equation:

\[ U = \alpha T_{avg} + \beta \sigma_T \]  

(2.3)

where \( U \) is the passenger’s (dis)utility, \( T_{avg} \) is the mean travel time, \( \sigma_T \) is the travel time standard deviation, and \( \alpha, \beta \) are the individual’s weights given to each factor. In the
“schedule delay” approach, reliability is represented less directly, with terms representing changes in passenger scheduling decisions. The results of these studies tend to favor the “schedule delay” approach, but there is no consensus on this point.

These two models could be more realistically tested using a revealed preference experiment, based on passengers’ “real-world” travel decisions. Unfortunately, no significant revealed preference studies of transit passenger reliability behavior have been published to date (presumably because of the expense involved). However, a similar study was undertaken for car travel reliability, using a California highway with free and (more reliable) toll lanes as a case study [19]. It tested whether unreliability cost was better modeled by the standard deviation of travel times, or the difference between the 90th percentile and median travel time. The results found the latter to have greater explanatory power for passengers’ route choice decisions, agreeing with the stated preference studies that explicit travel time delay better explains passengers’ unreliability cost than the variance.

### 2.2.3 Theoretical Model of Passenger Reliability

The limited surveys and experimental studies described in this section cannot, by themselves, provide a sufficient basis for a useful model of passenger-experienced reliability. To develop such a model, this section turns to the prior theoretical work on passenger reliability; in particular, that by Abkowitz et al [3] and Bates et al [19]. The rest of this section lays out what is intended to be a comprehensive model of passenger-experienced reliability, based on this work.

According to the model, the passenger experience of reliability consists of two processes: first the perception of reliability, then the reaction to it. The former can be defined as how the passenger observes service reliability, and the latter as how the passenger responds to minimize the cost of this unreliability.

**Perception of Reliability**

Service reliability from the passenger’s perspective, can be considered, at the most basic level, as the deviation of transit service outcomes from what passengers think the outcomes should be. What the outcome “should” be, however, depends greatly on the type of traveler, type of service, and time when the travel takes place. First-time travelers’ service expectations will likely be based on the published schedule, simply because no other information is
available. Thus, for such schedule-driven travelers, reliability will be perceived as adherence to the schedule.

Frequent travelers’ expectations, in contrast, will come from past experience with the service as operated, which may differ significantly from schedule. For such travelers, reliability will be perceived as the frequency and magnitude of deviation from the expected “typical” service, not simply schedule adherence. For the most part, this deviation will be perceived in terms of travel time variability. This variability is not equivalent to statistical variance, but rather a “sense” of travel time variability, based on an individual’s experience, relative to their expected “typical” travel time.

Abkowitz et al [3] give an example to illustrate the difference between reliability for first-time and frequent travelers: Suppose buses arrive at a stop exactly 5 minutes late, every day. For frequent travelers, this service would be considered very reliable, as it always arrives at the expected “typical” time. They can effectively plan future journeys based on this pattern; if they always arrive at the stop exactly 5 minutes after the scheduled departure time, they will always have zero waiting time. For the traveler relying on a schedule, however, this service would not be considered reliable, as it is always five minutes late relative to their expectation, which is the schedule.

Not all travelers will fit into these two categories. For example, a traveler may be traveling a particular route for the first time, but have experience riding similar routes in the system, and thus have a general idea of what to expect (e.g., “the bus will probably be a couple minutes late, due to traffic”). Thus, their service expectations may be some mix of schedule-based and experience-based.

For frequent travelers, service expectations should be based on trips taken under similar conditions—similar times of day, days (e.g. weekdays, holidays), weather, etc. This is because reliability should reflect unpredictable variation, not variation predictable by a traveler with reasonable knowledge. For example, a traveler usually traveling during the off-peak should not consider their usual route to be unreliable because they found it to be significantly slower in the peak period, as this is common for most urban bus routes.

**Reaction to Reliability**

First-time travelers will have little opportunity to react to the perceived unreliability, except by changing their general reliability expectations for the transit system as a whole (unless
they become frequent travelers on the particular route).

Abkowitz et al [3] posited that frequent passengers react to unreliability by altering their departure time to minimize travel time cost, as represented by the utility function $U_T = -\alpha T_{all} - \beta L$, where $U_T$ is the travel time cost, $T_{all}$ is the budgeted travel time, $L$ is the probability of being late, and $\alpha, \beta$ are the individual’s weights given to these factors. For a given level of reliability, a frequent traveler has an intuition (which may not be statistically accurate) of their probability of being late, $L$, for a given allocated travel time $T_{all}$. The passenger can reduce $L$ by departing earlier, thus reducing the probability of late arrival. However, doing so will cause their budgeted travel time to increase, causing the cost associated with $\alpha T_{all}$ to also increase. Thus, there is a trade off between increasing the likelihood of on-time arrival, and reducing allowed travel time.

The particular travel time a passenger will choose to minimize $U_T$ will depend on how much they value these competing costs—i.e., the values of $\alpha$ and $\beta$. A traveler valuing on-time arrival highly (e.g., someone traveling to the airport) will set their departure time earlier, incurring more cost from $\alpha T_{all}$, than someone with a low need for on-time arrival. In the case of an individual with no preferred time of arrival (e.g., leisure travel), an individual will not adjust their departure time at all in response to unreliability.

Most frequent travelers will budget a travel time larger than the median (i.e., their likelihood of on-time arrival will be greater than 50%). For such travelers, one can then define their buffer time as the difference between their budgeted and median travel times. The size of this buffer time is thus a quantifiable indicator of unreliability’s effect on the individual. If the passenger perceives an improvement in service reliability, a likely response will be to reduce their buffer time, and the reverse for a perceived decrease in reliability. As it is quantifiable, this buffer time can form a theoretical basis for measures of passenger-experienced reliability.

Agreement with Surveys and Studies

This model agrees with the key survey results of passenger attitudes towards reliability: it posits that passengers will adjust their departure times based on reliability, and they will react differently depending on their trip purpose (commute, leisure, etc.). Furthermore, it agrees with the empirical studies’ findings: that passengers are affected by unreliability through changes in their departure time decisions, rather than by the travel time variability.
in and of itself.

2.3 Comparison of Passenger and Operational Reliability

The model of passenger-experienced reliability developed in this section makes possible a formal comparison between passenger-experienced reliability, and reliability as captured by “traditional” operational reliability metrics. There are three main differences:

1. Operational reliability is based on timetable adherence, whereas passenger reliability is based on travel time variability.

2. Operational reliability is usually measured at the route or line-level, whereas passenger reliability is experienced at the origin-destination pair level.

3. Passenger reliability is experienced for the entire journey, but operational measures capture only portions of the journey.

Schedule-based vs. Predictability-based

As described in Section 2.1, operational reliability is generally measured in terms of adherence to schedules. Commuters and other frequent travelers, however, experience reliability as service variability. While schedule adherence and service variability are often correlated, they are not identical. If a service has a “faulty” schedule, it may consistently deviate from the schedule, but still have a compact travel time distribution.

Returning to the Abkowitz et al example, suppose the OTP standard for that route was -1/+4 minutes late. According to that standard, the bus would be very unreliable, with 0% OTP. However, as described before, for experienced travelers, this would be a highly reliable service, as it is very consistent. This situation can also apply with measures of headway regularity. Suppose for two subway lines, the scheduled headway is 5 minutes, and the acceptable headway threshold is 150% of the scheduled headway. On Line A, 25% of headways fall outside the threshold; of these late trains, however, more than half the headways are 15 minutes or more. On Line B, 35% of headways fall outside the threshold, but the longest headway is only 8.5 minutes. In terms of the headway regularity standard, Line A is better than Line B. However, from the passenger’s perspective, Line B would
probably be considered more reliable, as there is less probability of very long delays, which may entail longer buffer times to compensate.

**Route/Line vs. OD Pair**

Operational reliability is almost exclusively measured in terms of route/line level performance, because this is the level at which transit services are managed. Passengers’ reliability, however, is experienced in terms of a journey from a particular origin to a particular destination (i.e., OD pair). Route/line-level operational measures will be inadequate for journeys involving transfers (a large fraction of journeys in most large transit systems). Even for journeys beginning and ending on a single route/line, though, route/line-level reliability measures can substantially misrepresent the passenger’s experience, because service reliability on a single line/route can vary dramatically by location (CBD vs. suburbs) or direction (into CBD vs. out of CBD during “rush hour”).

**Journey Components vs. Full Journey**

For passengers, reliability is experienced in terms of their total travel time, from origin to destination, including all components of their journey: waiting time, in-vehicle time, transfer time, etc. The operational reliability measures described in Section 2.1, however, only measure single components of the passenger’s journey. OTP, headway regularity, and EWT capture only variation in waiting times, while running time adherence captures only in-vehicle time variation. For public transport services operating very close to capacity, non-vehicle factors such as denied boarding delays can form a significant portion of overall journey time variability.

Operational measures are also inadequate for measuring passenger reliability because they ignore how different reliability factors combine, negatively or positively, to affect the overall travel time variability (e.g., missing a timed transfer due to a late bus departure at one’s origin). Assessing combined reliability effects is especially important in situations where the effects of these factors are correlated. A classic example of this is “bunching” in high-frequency services, where longer-than-scheduled headways can cause longer running times, due to increased dwell times along the route (longer headways result in more passengers boarding, given a random arrival process, and more boardings lead to longer dwell times). Thus, in situations with serious bunching, there is a strong correlation between
waiting time and in-vehicle time variation for passengers’ journeys. An accurate passenger reliability assessment would need to capture this correlation, indicating a level of unreliability significantly “more than the sum of the parts”.
Chapter 3

Uses and Benefits of Reliability Measures

The discussion of passenger reliability in the previous chapter addressed the conceptual question of “What is passenger reliability?” From the pragmatic perspective of a transit operator, however, the more important question regarding the measurement of passenger-experienced reliability is: “What are the benefits?” This question is addressed in this chapter. It is organized according to the five broad applications of reliability measures identified in this research: Management of Reliability, Passenger Information, Reporting to Stakeholders, Cost-Benefit Analyses, and Long-Term Planning. Each reliability management application is first described in detail, based on the research literature and on practice. From these findings, the potential benefits of using passenger-experienced reliability measures are then identified and discussed for each application.

3.1 Reliability Management

3.1.1 Overview of Reliability Management

Reliability management can be defined as the system used by a public transport authority to maintain, and hopefully improve, the reliability of its services. This system has three main components:

1. Identification of reliability problems

2. Identification of causes of reliability problems
3. Implementation of service changes to address reliability problems

Both industry guidebooks, such as the TCRP Capacity and Quality of Service Manual (TCQSM) [21], and operator practices, such as those of NYCT [22], describe reliability measurement as playing an important role in each part of the overall process, as described below.

**Identification of Reliability Problems:** This first part of the reliability management system is the process for detecting unreliable service, and then assessing whether the unreliability is severe enough to justify intervention to improve it. Formal reliability measurement’s role in this process is clear: without it, there would be no objective way for the public transport authority to be informed about services being unreliable.

Once a public transport authority begins tracking reliability metrics for a given service, it can apply different decision rules to decide whether the level of unreliability merits action. These fall into three main categories: service standards, performance relative to peer services, and reliability trends:

- **Service Standards:** Measured performance is compared to a set reliability standard for that service. If the performance falls below the standard, the service is flagged. Compliance can be assessed in terms of a single reliability metric (e.g., fail if $\text{OTP} \leq 70\%$) or several (e.g., fail if $\text{OTP} \leq 70\%$ or $\text{Headway Regularity} \leq 70\%$).

- **Performance relative to peer services:** Service is flagged if its reliability is low relative to peer services (e.g., worst-performing line in the network).

- **Reliability Trends:** If reliability metrics for a service indicate significant deterioration in reliability over time (e.g., OTP decline of 50% relative to six-month average), the service is flagged. What constitutes a “significant” decline will depend upon the judgment of the operator, and attributes of the service. Using reliability trends to identify “problem routes” can permit a more proactive approach to managing reliability, helping identify problems before they cause large declines in service quality.

Notably, while reliability trend rules are based on changing reliability performance, the first two rules do not require reliability performance change to merit intervention; thus, these can flag transit services on the basis of “normal” operations.
Identification of Causes of Reliability Problems: Once reliability issues are identified, their causes must be “diagnosed”, so specific interventions can be developed. Reliability measures can assist this step in several ways. First, based on what specific reliability measures triggered the intervention, the possible categories of causes can be narrowed to some degree. For example, if OTP declines, but an incident delay indicator shows no change, the issues are unlikely to be related to incidents. When performance deteriorated can also help investigators identify the causes of unreliability (e.g., finding reliability declines coincide with snowstorms). Factors causing unreliability can also be identified by comparing patterns in reliability metrics with patterns in other service attributes. For example, New York City Transit (NYCT) [22] examined the relationship between headway regularity performance and passenger demand, and found that passenger demand explained 77% of the variation in headway regularity, indicating demand is a driver of poor performance.

Evaluation of Service Changes

After service changes have been implemented to improve reliability, their effectiveness needs to be evaluated: Did reliability improve? If so, to what degree? Did the improvements meet expectations? Were the increased costs (if any) justified by the reliability improvement?

To answer these questions, service reliability must be measured both before and after the service changes, and compared between the two time periods. In this role, reliability measurement should be done in a manner that isolates the effects of the service change as much as possible from exogenous factors (e.g., weather). Using reliability measurement for this purpose is emphasized in industry guidebooks [21, 23]. Public transport authorities often use this type of evaluation to determine whether pilot projects should be made permanent. For example, NYCT [22] used its reliability measurement program to evaluate the effectiveness of its pilot Select Bus Service (a BRT-type service), and signaling improvements on the Lexington Avenue subway line.

3.1.2 Benefits of Passenger-Experienced Reliability Measures

The proposed reliability management process can only identify, diagnose, and correct reliability problems detectable through the operator’s reliability metrics. If certain types of unreliability cannot be measured with an operator’s set of metrics, such reliability cannot be effectively measured. Thus, the overall effectiveness of a reliability management program is
largely determined by the range of unreliability sources captured by the operator’s metrics; if certain types of unreliability cannot be measured with an operator’s metrics, such unreliability cannot be effectively managed. Ideally, an effective reliability management process should use metrics that capture all service factors that affect the passenger’s experience of reliability.

In this context, passenger-experienced reliability measures can substantially improve the effectiveness of reliability management systems, by capturing those aspects of passenger-experienced reliability described in Section 2.3, not reflected in operational reliability measures, including:

- Denied boarding effects
- Congestion in stations (increased access/egress time)
- Correlation between delays on multiple aspects of a journey (e.g., bus bunching leading to long wait and long in-vehicle times)
- Impact of very long delays (OTP usually treats all delays over X minutes as similar)

By enlarging the scope of a reliability management program to include these reliability factors, incorporating passenger-experienced reliability metrics should result in a program more responsive to passenger’s experiences and needs.

This is problematic for the large number of public transport agencies dependent upon operational metrics. As shown in Section 2.1, there are a number of aspects of public transport service which, while integral to passenger expectations about service reliability, are not captured by operational reliability measures. This issue limits the ability of the reliability management program to respond to passenger needs. This problem can be addressed, however, with the inclusion of well-designed measures of passenger-experienced reliability. Such measures should allow public transport agencies to track trends in a number of types of service reliability that would generally be infeasible to measure with “traditional” reliability measures, in turn allowing these aspects of reliability to be managed for the benefit of passengers. Such “revealed” aspects of service reliability are described below:
3.2 Passenger Information

By providing passengers with detailed reliability information about the public transport services they use, public transport agencies can help passengers better plan their journeys. As described in Section 2.2, passengers often react to the presence of unreliability by adjusting their departure times so as to reduce the probability of late arrival to an acceptable level. Passengers may also choose to change their route on the basis of this perceived reliability. They do this, though, not based on a “calculated” distribution of journey times, but rather a rough “sense” of how reliable the services are.

To the extent that this intuition about reliability is inaccurate, passengers may incur unnecessary travel costs, due to their chosen departure times not being “optimal” given their value of budgeted travel time and likelihood of on-time arrival. If a passenger overestimates the service unreliability, they may budget more “buffer time” than they need to achieve their intended likelihood of on-time arrival. If a passenger underestimates the level of service unreliability, on the other hand, they may arrive late more often than planned. If provided reliability information can give passengers a better idea of their likelihood of late arrival, for a given budgeted travel time, passengers can then refine their service expectations, and better “optimize” their future travel, in terms of departure time and route taken.

This benefit to passengers can only be achieved, however, if reliability information is provided using metrics that directly reflect the passenger’s reliability experience. Such metrics are necessary to convey reliability effects in a manner that can be easily interpreted and understood by passengers. An operational metric like OTP, presenting reliability at the train level, tells passengers little about their real likelihood of late arrival, because the information is not presented in terms of passenger travel times, and ignores important sources of travel time variability, such as denied boardings.

Uniman [2] provides an example of how such reliability information could be provided. Figure 3-1 shows a hypothetical online journey planner for the London Underground. In addition to the expected arrival time being provided (the default for public transport journey planners), the “latest arrival” time is also provided, representing the latest time at which a passenger should expect to arrive at the destination (set at the 95th percentile travel time; refer to the documentation on the RBT for why this was chosen). There are, of course, other ways in which reliability could be presented to passengers, and other media by which
it could be conveyed (e.g., printed schedules, in-station displays, etc.).

Figure 3-1: Proposed presentation of reliability information in Journey Planner

<table>
<thead>
<tr>
<th>Route</th>
<th>Depart</th>
<th>Expected Arrival</th>
<th>Latest Arrival</th>
<th>Duration (up to)</th>
<th>Interchanges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>08:27</td>
<td>08:57</td>
<td>09:07</td>
<td>00:30 (00:40)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>08:20</td>
<td>08:50</td>
<td>09:00</td>
<td>00:30 (00:40)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>08:20</td>
<td>08:50</td>
<td>09:00</td>
<td>00:30 (00:40)*</td>
<td></td>
</tr>
</tbody>
</table>

*Service is currently disrupted – reported figures may not apply – expect severe delays

Besides the direct benefit to passengers described above, providing passengers with reliability information in terms of their own experience also has a potential indirect benefit: improving passengers’ overall perceptions of the public transport authority. Transport for London [18] notes that providing effective passenger information can improve passengers’ overall perceptions of the public transport authority. Providing good information to passengers indicates to them that the operator cares about their needs, and is trustworthy.

### 3.3 Reporting to Stakeholders

The primary purpose of reporting performance measures is to hold public transport agencies accountable to their outside stakeholders with respect to how well they are meeting the needs of their passengers. If these stakeholders determine the public transport authority’s performance is not adequately meeting the needs of public transport riders (or the public in general), such stakeholders will bring pressure to bear upon the authority to address these issues. This fundamental process should be the same regardless of the stakeholder’s role, whether it be a governmental oversight committee or an neighborhood activist group. For reporting service reliability, this process can be improved by incorporating passenger-experienced reliability measures. The general value for this application is clear: if the overall objective is to address passengers’ concerns regarding reliability, the metrics used to report reliability should accurately reflect, as much as possible, passengers’ reliability experiences. The specifics of how such measures can improve the reporting process depend on the relationship between the stakeholder and the public transport authority.

---

1Source: Uniman [2]
For stakeholders with direct authority over public transport agencies (boards of directors, regional transportation departments, government oversight committees, etc.), performance measures are used to assess the operator’s overall performance. If performance is deemed insufficient, such bodies may compel the public transport authority to make changes to improve future performance. Thus, it is critical that such measures reflect the interests of passengers and the public at large; otherwise, a public transport authority may be compelled to take actions not in the best interest of the public—or be exempted from actions it should take. This result would not only be detrimental to riders, but would also go against the mandate of such oversight bodies, which is to represent the interest of the general public. Thus, for oversight of reliability, it is beneficial that public transport agencies report it in terms of the passenger experience, rather than the operational experience.

For other stakeholders, with no direct power over the public transport authority (riders’ groups, neighborhood advocacy groups, etc.), performance metrics are used to present their concerns to the management, and inform the ensuing debate between the stakeholders and the operator over what (if any) actions should be taken to address these concerns. Reported reliability metrics that accurately reflect passengers’ reliability experience should be effective tools for such stakeholders to articulate their concerns about service reliability. Arguments made to public transport agencies in terms of metrics they already accept (e.g., “the reliability X of Line C through our neighborhood declined by 60% over the past six months”) are likely to carry much more weight than those based mostly on anecdotal evidence (e.g., “twenty people have come to us and complained that service has gotten worse”).

For both types of stakeholders, the extent to which reported reliability measures reflect passengers’ actual experiences affects the trust such stakeholders have in this process. If stakeholders perceive that the reliability being reported does not reflect the passenger experience, they may conclude that the public transport authority is not interested in passengers’ needs, or even worse, is actively trying to misrepresent its performance. Thus, the implementation of passenger-experienced reliability measures could enhance the credibility of this reporting and accountability process, as well as, possibly, that of the public transport authority itself.
3.4 Cost-Benefit Analysis

A fundamental part of any public transport authority’s planning process for infrastructure or operational investment is formal cost-benefit analysis. In this process, the effects of proposed investments are predicted and quantified, in terms of their costs and benefits to passengers and the public transport authority. These costs and benefits then inform the final decisions made (either by the operator, or by outside funding bodies) to accept or reject the proposed investments, on the basis of maximizing the net benefits (i.e., benefits minus the costs) to the public transport authority and its customers.

To evaluate projects expected to produce significant changes in service reliability, the public transport authority should have a means of quantifying the costs and/or benefits to passengers from this reliability change. Passenger-experienced reliability measures, by definition, are the best metrics to quantify the expected change in passenger-experienced reliability. These reliability measures could be used by themselves to inform the decision-making process, or be converted into monetary values by some “value of reliability” analogous to the “value of time” often used in cost-benefit analyses.

This need for passenger-experienced reliability measurement in cost-benefit analysis applies to decisions whether to pursue a single project (i.e., binary “yes or no” decisions), as well as the decision process for selecting a set of projects to implement from a larger set of potential projects. An example of the first scenario would be a decision whether to replace the signaling system on a subway line; such an investment may not be justifiable on capacity improvements alone, but may become attractive when the benefits of improved reliability are included. An example of the second scenario would be an operator deciding which two of their ten key bus routes should be given exclusive bus lanes; clearly the degree of unreliability on each of these routes should be a key factor in deciding which should receive the exclusive lanes.

3.5 Long-Range and Regional Planning

Passenger-experienced reliability measures can support long-range and regional planning in several respects. First, because passenger reliability can be measured at a very disaggregate, OD-pair level, such measures can support geographical studies of service quality provided by a public transport network. Such studies could focus on particular neighborhoods or
districts, important corridors, or areas with certain characteristics across the service area. A particularly important class of such studies are equity analyses; passenger reliability can be compared across geographic areas, or between areas with large low-income and/or minority populations and those without, to see if certain areas or groups receive significantly worse service quality than others.

Second, passenger-experienced reliability can be an input into studies of city and regional accessibility, i.e. where people can travel for given levels of travel cost. While such studies usually represent travel cost in terms of average travel time (to jobs, schools, etc.), this travel cost can be more accurately modeled by including the cost to passengers of unreliability. Finding this unreliability cost is, of course, best done through reliability measures that capture the passenger experience of reliability. Combined with its geographically disaggregate nature, passenger reliability measurement allows more accurate accessibility studies to be performed at a variety of geographic scales.

This ability of passenger-experienced reliability measures to improve travel cost estimates also makes them potentially useful for transport demand modeling, as passenger choices of where to travel, what mode to take, and what route to take are largely a function of the general costs of the various alternatives. If passenger reliability measures are found that can be effectively modeled and forecast, their inclusion could improve the accuracy of transport models (e.g., four-step models) used throughout transport planning practice.
Chapter 4

Design of Reliability Metrics

The benefits of measuring the passenger’s experience of reliability described in Chapter 3 can be achieved only through the use of well-designed reliability metrics. Such metrics should be consistent with the conceptual model of passenger-experienced reliability described in Chapter 2, and the practical needs of reliability measurement applications described in Chapter 3. To aid the design process for passenger-experienced reliability metrics, this chapter develops design objectives for such metrics based on these needs, and then evaluates previously-developed passenger-experienced reliability metrics against these objectives.

Section 4.1 outlines the design objectives, which are meant to shape the mathematical definition of the metrics. These can be considered “intrinsic” properties of a given metric, independent of the particular transit system context. Sections 4.2 and 4.3 review and evaluate previously-developed metrics, using the design objectives developed in Section 4.1. Section 4.2 reviews the Reliability Buffer Time (RBT), a measure that estimates the “buffer time” passengers may allocate due to service variability, as described in Chapter 2. Section 4.3 reviews several other existing passenger-experienced reliability metrics. The review identifies several important limitations of existing metrics, including exclusion of important unreliability sources and potential bias from passenger behavior, which motivate the development of this thesis’s two new metrics in Chapters 5 and 6. The RBT is reviewed in greater detail because it forms the basis for these two new metrics.
4.1 Design Objectives

This section describes the desired attributes of a passenger-experienced reliability metric. A metric’s ability to be successfully implemented, and produce the benefits described in the previous chapter, largely depends on the extent to which it achieves these design objectives. In practice, fulfilling all of the objectives well is infeasible, given that the requirements, as will be shown, are sometimes contradictory. Thus, cost-benefit trade-offs between competing requirements are unavoidable. Fortunately, such metrics do not need to fulfill all of these objectives to produce significant benefits.

In addition to informing the design of new metrics, these objectives can also be applied to evaluate existing measures, as in Sections 4.2 and 4.3. The requirements described in this section are grouped into five general objectives: representative of the passenger experience, meaningful for passengers and non-experts, comparable across times of day and services, time period flexibility, and service scope flexibility.

4.1.1 Representative of the Passenger Experience

A passenger-experienced reliability metric should represent the “typical” passenger’s experience of reliability, consistent with the model described in Section 2.2.3. This objective should be self-evident: the entire basis for measuring reliability with such measures is that they capture the passenger’s experience of reliability; otherwise, they provide minimal benefit over existing operational reliability measures.\(^1\) This need applies to all of the applications described in Chapter 3. This objective can be evaluated using the following criteria:

**Include all Sources of Unreliability:** The metrics should incorporate all sources of passenger-experienced unreliability. This includes both “operational” unreliability related to poor schedule adherence—long wait times due to longer-than-scheduled headways, long in-vehicle times due to longer-than-scheduled running times—as well as delays related to high passenger demand, such as denied boardings, and longer access (station entrance to platform walk), egress (platform to station exit walk), and transfer times due to crowding in stations.

\(^1\)Note: It is not expected that such metrics fully capture how the passenger perceives reliability, as perception is subjective, and affected by many outside factors (e.g., mood) not related to the physical experience of reliability.
Distinguish between Service Variability and Schedule Adherence: A metric should capture either travel time variability, or the difference between actual and scheduled travel times, depending on whether the metric is meant to measure the reliability of non-schedule-reliant (e.g., most commuters) or schedule-reliant (e.g., occasional users of a low-frequency service) travelers. Measuring reliability for both groups will thus generally require two separate measures. This division largely corresponds to the division between occasional and frequent travelers, with the latter more likely to rely on experience than schedules (including online journey planners).

Distinguish between early and late arrivals: For metrics measuring travel time variability, the numerical values should not weigh late and early arrivals equally, as the former is likely to be much more costly for the passenger than the latter. Thus, standard measures of variability such as the standard deviation are not appropriate, as they weigh values above and below the mean equally.

Control for Variation in Passenger Behavior: Travel time variability metrics should exclude, to the extent possible, variability due to passenger behavior. This is necessary to ensure the metrics primarily reflect the performance of the transit service, rather than outside factors. This includes both inter-personal variation, variation in behavior among individuals (e.g., slow vs. fast walkers), and intra-personal variation, variation in an individual’s behavior between trips (e.g., shopping at a store inside the paid fare area on occasion).

Control for Time of Day Variation in Average Travel Time: Metrics measuring travel time variability should control for systemic time of day trends in average travel time. This includes changes in the service schedule (e.g., frequency), and changes in passenger demand (e.g., peak vs. off-peak). The variability incorporated into reliability measures should only consist of the “unpredictable” variability that makes it difficult to plan journeys, rather than predictable changes that can be planned for.

Exclude “Extreme” Delays: The metrics should exclude the effect of “extremely” long delays that occur very infrequently (e.g., 1 in 100 trips). Such delays can significantly increase aggregate travel time variability, even though they affect only a small fraction
of travelers’ trips, and are thus unlikely to be a major factor of the typical passenger’s reliability experience. Abkowitz et al [3] recommend transport operators use a separate reliability measure for extremely long delays.

**Calculated at OD-Pair Level:** The metrics should be calculated, at the base level, for individual OD pairs. If higher-level metrics are desired, they should be derived from OD pair level measurements, accounting for the difference in passenger flow across the different OD pairs.

**No Bias towards Particular Passenger Demographics:** The metrics should neither over-represent, under-represent, nor exclude the reliability experience of a large demographic group of users (low-income, elderly, etc.). This issue is of particular concern for metrics using automatic fare collection data from “smart cards”, as such cards may have lower adoption rates among certain demographic groups.

### 4.1.2 Meaningful for Passengers and Non-Experts

The results of passenger-experienced reliability metrics should be meaningful for “non-experts”, including passengers, outside stakeholders, regulators, politicians, senior management—essentially everyone not deeply involved with the analysis of transit performance. Metrics that are not meaningful to non-experts will not be effective for passenger information purposes, or for reporting to outside stakeholders (who are usually “non-experts”). “Meaningful” is, of course, a vague term without further explanation; in this context, passenger-experienced reliability measures are considered to be “meaningful” if they meet the following (still subjective) requirements:

**Understandable:** The numerical results of the metrics should be intuitively understandable by passengers and other non-experts. They should be able to see the results of the reliability measure for particular OD pairs or routes/lines of interest and, with little prior explanation, understand two key points: (1) what the quantity physically represents (e.g., OTP is the percentage of trains departing within X minutes of the scheduled time), and (2) how its value generally relates to journeys (e.g., a higher OTP generally indicates a greater likelihood of on time arrival). Understandability is, of course, dependent upon the particular individual; what may be understandable to a university-educated government official
may be quite confusing to a passenger with little formal education. The challenge of making reliability metrics understandable is likely greater for passenger information than for reporting, as the recipients for the latter are likely to have better technical understanding, and put more effort into understanding the measures.

What particular attributes drive “understandability” in a metric is subjective, and involves a number of psychological factors outside the scope of this research. However, some tentative guidelines can be offered:

1. The metric should not be expressed in complex mathematical terms. However, this does not entail that the metric cannot be calculated using such operations. For example, the Excess Wait Time (see Section 2.1.2) is calculated in terms of a coefficient of variation, which is beyond the grasp of the average passenger; however, it is expressed in terms of an average (the average excess wait time), which most people understand.

2. The metric should be expressed in terms of time (e.g., 5 minutes) or a percentage, which are more easily understood than continuously varying dimensionless factors (e.g., coefficient of variation).

3. The metric should be expressed in terms of generally understood concepts, such as travel time, waiting time, or train (bus) trips.

**Objective:** The metric’s results should be considered objective, i.e. they cannot be used to “hide” bad performance from stakeholders. Towards this end, the metrics used should require little or no human judgment in their estimation. Such unwanted subjectivity can be explicit, or implicit in the form of arbitrary parameters used in the calculation, that were set based on subjective criteria.

**Useful for planning journeys:** To be effective in a passenger information role, reliability metrics should provide actionable information for planning journeys, that facilitates improved decisions about departure times and route choices. This is best done by expressing reliability in terms of factors that drive passenger decisions: total journey time, arrival time, departure time, etc. Metrics should also be, as noted by Transport for London [18], “future focused”: they should help predict future performance, rather than being useful only for describing past performance. Finally, metrics should be suitable for the communi-
cation media available to the operator: journey planners, printed schedules, mobile apps, etc.

4.1.3 Comparable Across Different Services and Times of Day

Passenger-experienced reliability metrics should be comparable across times of day and public transport services (train lines, bus routes, etc.). For example, one should be able to objectively say that passenger-experienced reliability on Line A is better (or worse) than that on Line B, or Route X is more reliable in the AM peak than the PM peak.

Comparability across services and times of day is essential for reporting in two respects. First, outside stakeholders often want to know relative performance of different services. Second, outside stakeholders often desire system-level performance indicators reported across multiple services (e.g., all subway lines). To calculate such system-level indicators, reliability measurements at the service-level must be “averaged” in a manner that accurately represents their relative reliability performance. Cost-benefit analyses may also require comparability, as reliability costs and benefits for large projects may need to be assessed across multiple services (e.g., an improved bus dispatching system).

This objective can be evaluated according to the following criteria:

Independent of schedule: A measure of travel time variability should be independent of the service schedule, i.e. scheduled headways and running times, because these will vary across services, and across different times of day for a single service (this is not applicable, of course, for schedule adherence metrics).

Absolute, not relative: If reliability metric values are to be compared across different services, they should be absolute measures of reliability, not ones relative to past performance of a particular transit service. This entails that a reliability value of X on Service A should represent the same quality of service as a reliability value of X on Service B.

4.1.4 Time Period Flexibility

Passenger-experienced reliability metrics should be flexible with respect to the time of day intervals and date periods they are calculable for. The former is the portion of the service
Time of day flexibility is needed because reliability can change significantly over the service day. Such variation is not adequately captured by simply dividing the day into “peak” and “off-peak”; finer temporal resolution is required. This is especially important during the peak hours, when there is often substantial within-peak service quality variance, much of it driven by dynamics in demand. Capturing these effects is critical for passenger information, because such information is only useful when relevant to when one is traveling. It is important for reporting purposes, because it gives stakeholders an idea of how reliability performance changes over the day. If performance indicators are provided at too broad time periods, an impression may be given that the operator is trying to hide “bad” reliability performance during certain times, by averaging it out with “good” reliability performance at other times. For reliability management, it is useful because the most serious reliability issues often occur during the peak hour or peak half-hour of the service day. Flexible time periods allow analysts to measure reliability specifically for these “peak of peak” periods.

Date period flexibility allows reliability measurement to be separated by type of service day (weekday, weekend, holiday). This is important for passenger information and reporting, because performance is generally very different across weekdays, weekends, and holidays. Flexible date periods also allow one to control for exogenous events affecting reliability, useful when determining root causes of reliability problems. For example, if bus service is affected by rain storms, the reliability could be calculated for rainy days, and compared to reliability for dry days.

Reliability metrics can be considered to have good time period flexibility if they:

- Are calculable for short time of day intervals across the entire service day. In particular, metrics should be calculable at 60-minute intervals or (preferably) smaller during the peak hours.
- Allow weekends, holidays, and other specific days (e.g., the date of a major incident.) to be excluded from calculation.
4.1.5 Service Scope Flexibility

Passenger-experienced reliability metrics should be calculable for service scopes beyond
the route/line-level. Such scopes include the OD pair level, service segment, and transfer
pattern level (e.g., trips starting on the Island Line and ending on the Tsuen Wan Line).
Service scopes can also be defined geographically—for example, all trips from Kowloon to
Hong Kong’s CBD.

The ability to calculate at the OD pair level is critical for representing the passenger’s
experience, as noted earlier. Service scope flexibility can also be useful for assessing reliabil-
ity issues affecting particular parts of a line/route (e.g., congestion in a particular station),
or across multiple lines/routes (e.g., high demand in the CBD). This feature is also use-
ful for high-level planning and analysis purposes. Activities including regional accessibility
studies, neighborhood/corridor reliability analyses, and equity analyses can require aggre-
gating reliability measures over a variety of service and geographic scales. The best units
of aggregation may not correspond to routes or lines, as a single transit service may cross
multiple geographic boundaries of interest (e.g., municipal borders).

Service scope flexibility is best realized when a metric is calculable for any combination
of OD pairs on a transit network. This should then allow the construction of aggregate
metrics representing the average passenger-experienced reliability for any part or combina-
tion of transit services, or geographic areas of interest. In keeping with the philosophy of
representing the reliability for the “typical passenger”, the aggregation process should ac-
count for the differences in passenger flow across different OD pairs. A straightforward way
to do this is to take the weighted average of OD level values with respect to the passenger
flow on each OD.

4.2 Reliability Buffer Time (RBT)

As described in Chapter 2, travelers often allocate extra “buffer time”, in addition to their
expected travel time, to ensure an acceptable probability of on-time arrival. Each passen-
ger’s buffer time for a particular journey (i.e., OD pair) will be determined largely by the
distribution of their previous travel times, which in turn is primarily a function of the reli-
ability of the transit service(s) used. If the service is very reliable, passengers’ travel time
distributions will usually be “compact”, requiring a small buffer time relative to their ex-
pected travel time. However, if a passenger frequently experiences delays, their travel time distribution will likely have a long “right tail”, requiring a large buffer time to compensate. Thus, a passenger’s buffer time is a good indicator of their reliability experience. Chan and Uniman [2] utilized this buffer time concept to propose a new reliability metric called the Reliability Buffer Time, or RBT[^2] [4].

### 4.2.1 Definition of the RBT

The RBT is defined as the difference between the $N$th and 50th percentile (i.e. median) passenger travel times over a specific OD pair and time period. It aims to represent the buffer time necessary for a typical passenger to achieve an $N$-percent probability of on-time arrival, where $N$ is an “upper percentile” much greater than 50—generally the 95th percentile, for reasons discussed later in this section. The median journey time represents passengers’ “expected” travel time; the median is used, rather than the mean, for two reasons [2]: (1) it is not sensitive to outliers or the right tail of the distribution, and (2) studies have shown [24] that the median travel time is a better predictor of passenger behavior than the mean. The mathematical formulation of the RBT is then given by:

$$\text{RBT} = (\text{TT}_{N\%} - \text{TT}_{50\%})_{\text{OD, Time Interval, Date Period}}$$  \hspace{1cm} (4.1)

where $\text{TT}_{N\%}$ and $\text{TT}_{50\%}$ indicate the $N$th percentile and median of this distribution, respectively. The subscript *Time Interval* indicates the time-of-day interval (e.g., 8-8:30 AM, 4-7 PM, etc.) over which passenger trips are aggregated, based on their station entry times. Likewise, the subscript *Date Period* indicates the set of days over which passenger trips are aggregated (e.g., weekdays in April 2014).

The RBT has a direct basis in the Buffer Time Index (BTI) metric developed to measure roadway congestion, defined as the difference between the 95th percentile and average travel times for a particular road section, divided by the average travel time, as given below [25]:

$$\text{BTI} = \frac{\text{TT}_{95\%} - \text{TT}_{Avg}}{\text{TT}_{Avg}}$$  \hspace{1cm} (4.2)

The Buffer Time Index has been used in the United States by the Federal Highway Administration (FHWA), Minnesota Department of Transportation (MnDOT), and Washington

[^2]: Note: Chan [4] refers to the RBT in her thesis as the Reliability Factor (RF).
State Department of Transportation (WSDOT) to assess and monitor traffic congestion in urban areas[26].

**Spatial Aggregation**

The RBT, as defined above, applies only to an individual OD pair. For many management purposes, reliability metrics defined at the line- or route-level are desired. Chan and Uniman defined the line/route-level RBT as the average of the OD-level RBTs for all OD pairs beginning and ending on the line (route), weighted by the estimated passenger flow for each OD pair, \( f_{OD} \), as given by

\[
RBT_{Line} = \frac{\sum_{OD\in Line} f_{OD} \cdot RBT_{OD}}{\sum_{OD\in Line} f_{OD}}
\]  

(4.3)

This approach gives greater weight to the reliability of OD pairs used by more passengers, producing an aggregate RBT more representative of the typical passenger’s experience than one calculated from a simple average of the OD-level RBT values for the line. By grouping the trips by OD pair, the method above also controls for differences in average travel time across OD pairs, travel time variation not experienced by the passenger.

**4.2.2 Upper Percentile \( N \)**

The 95th percentile was chosen by Chan and Uniman as the recommended value of \( N \), to address three major considerations: relevance to passengers, realistic for operators, and data feasibility. If a passenger budgets the \( N \)th percentile travel time, they should expect to arrive late (assuming a consistent level of reliability) once every \( \frac{100}{100-N} \) trips, as shown in Figure 4-1 (the figure annotations indicate the equivalent delay frequency for a typical commuter). For \( N=95 \), this equals one late arrival every 20 trips, roughly equivalent to one late arrival per month. This is probably acceptable for most commuters. Budgeting significantly less time (e.g., \( N=80 \)) would likely raise the frequency of late arrival unacceptably (for a fixed level of reliability), while budgeting significantly more time (e.g., \( N=99 \)) would incur extra travel time costs, with little perceived benefit.

For operators, the \( N \)th percentile should represent a level of delay that can feasibly be controlled by the operator. There will always be some random, unpredictable delays.
that occur even in the best-run transport systems—passenger sickness, extreme weather, etc.—beyond the ability of the operator to control. Reliability metrics that mainly reflect these “extreme”, uncontrollable delays will be of little value in managing operations. Thus, the upper percentile $N$ should be set low enough to exclude such factors from the RBT; the 95th percentile is considered by Chan and Uniman to be sufficient in this regard.

From the statistical perspective, it is more difficult to estimate accurately the “extreme” values of a distribution (i.e., $N \rightarrow 100$), than values closer to the median. Thus, an RBT based on a very high $N$ is likely to be more prone to random errors than one based on a lower $N$. The 95th percentile is considered in Chan and Uniman to be possible to estimate with acceptable accuracy.

Two methods of calculating the RBT have been proposed. The first, developed by Chan, estimates the RBT from AFC (smart card) data, while the second, developed by Ehrlich and extended by Schil, derives it from AVL data. These two methods, the AFC-based RBT and AVL-based RBT, respectively, are reviewed below.

### 4.2.3 AFC-Based Calculation Method

The AFC-based calculation method obtains the distribution of passengers’ journey times directly from AFC data, using the difference between entry gate and exit gate transaction times to calculate individuals’ trip times. This method automatically and accurately ac-
counts for denied boardings and station congestion, as such factors will affect passengers’
gate-to-gate travel times. It also requires the transit service use both entry and exit fare
transactions, so that total journey times can be derived from AFC data. The RBT, then, is
simply the difference between the 95th and 50th percentiles of this distribution. Calculating
an AFC-based line-level RBT is straightforward, as $f_{od}$ can be measured directly for each
same-line OD pair, for any given time interval and date period.

### 4.2.4 AVL-Based Calculation Method

The AVL-based calculation method was developed by Ehrlich [5] as a means of calculating
the RBT for London Bus services. Because passengers do not “tap out” when alighting
from the bus, their travel times cannot be measured directly from AFC data; thus, the
AFC-based calculation method cannot be used. The objective of the AVL-based method,
then, is to model the passenger travel time distribution using only headway and vehicle
travel time information derived from AVL data.

The basic approach taken by Ehrlich is to calculate the travel time distribution of
passengers on each bus trip during the time period of interest, and then find the average
weighted by the number of passengers on each trip. This can be written using cumulative
distribution functions (CDFs) as

$$
P(T \leq t) = \frac{\sum_{i \in N} P(T_i \leq t) \cdot f_i}{\sum_{i \in N} f_i}
$$

(4.4)

where $T$ is the total travel time, $P(T \leq t)$ is the time period passenger travel time CDF, $N$
is the set of vehicle trips in the target time period, $f_i$ is the estimated number of passengers
on trip $i$, and $P(T_i \leq t)$ is the travel time CDF for passengers on vehicle trip $i$. To estimate
the bus trip-level travel time distributions, the method makes the following assumptions:

1. Passengers arrive randomly.

2. The rate of arrival is constant over the time interval of interest.

3. All passengers board the first vehicle to arrive (i.e., no denied boardings)

The method models the total travel time as the sum of two components: waiting time
$W$, and in-vehicle time $V$. Given the passenger arrival assumptions, it can be shown [5]
that the probability a given passenger on bus trip $i$ with preceding headway $H_i$ will have a waiting time of $w$ is given by the probability density function (PDF):

$$P(W_i = w) = \begin{cases} 
\frac{1}{H_i} & 0 \leq w \leq H_i \\
0 & w > H_i 
\end{cases} \quad (4.5)$$

The in-vehicle time $V_i$ for a particular trip is deterministic, being the time it takes the bus to move from the origin stop to the destination stop. Thus, the total travel time distribution for trip $i$ is simply the waiting time distribution “shifted” forward in time by $V_i$:

$$P(T_i = t) = \begin{cases} 
0 & t < V_i \\
\frac{1}{H_i} & V_i \leq t \leq V_i + H_i \\
0 & t > V_i + H_i
\end{cases} \quad (4.6)$$

This concept is illustrated by Figure 4-2. Integrating equation 4.6 with respect to time yields the following CDF:

$$P(T_i \leq t) = \begin{cases} 
0 & t < V_i \\
\frac{t - V_i}{H_i} & V_i \leq t \leq V_i + H_i \\
1 & t > V_i + H_i
\end{cases} \quad (4.7)$$

If it is assumed the arrival rate over the time period is constant, it follows that the number of passengers traveling on each vehicle trip $i$ is proportional to the preceding headway $H_i$. This allows the time period journey time CDF in Equation 4.4 to be rewritten as

$$P(T \leq t) = \sum_{i \in N} P(T_i \leq t) \cdot H_i \quad (4.8)$$

From Equations 4.7 and 4.8, a numerical passenger travel time CDF can be calculated from a set of headways and vehicle travel times. Denoting this $F_T(j)$, the RBT can then be derived from the inverse of this function:

$$\text{RBT(AVL)} = \left( F_T^{-1}(0.95) - F_T^{-1}(0.5) \right)_{\text{OD, Time Interval, Date Period}} \quad (4.9)$$
The process for aggregating the OD-level AVL-based RBT to the line-level depends on the data sources available to calculate $f_{od}$. If AFC data is available with both entry and exit transactions, $f_{od}$ can be measured directly, as with the AFC-based RBT. If such data is unavailable, static OD-matrices derived from surveys or manual observations can be used. As a given OD pair’s demand relative to total line demand is unlikely to shift greatly over a several-year time span (unless major route/line changes occur), the use of recent static OD matrices is unlikely to introduce appreciable error into the line-level AVL-based RBT.

4.2.5 RBT Dependence on the Scheduled Headway

An important consideration when assessing the RBT metric is its dependence on the scheduled headway. As noted in Section 2.2, for high-frequency services (i.e., headways 10 minutes or less) most passengers arrive at random at departure locations (train platform, bus stop, etc.). This arrival time variation leads directly to waiting time variation, when combined with discrete departure times: passengers arriving immediately before a departure will have no wait time, those arriving immediately after will wait the full headway, and all others will wait somewhere in between (this includes passengers who arrive when a train is in the station—while they do not need to wait on the platform, they still have “wait time” between when they board the train and its departure).

This passenger arrival-driven wait time variation is a major factor in the overall travel

\[ \text{Figure 4-2: Journey time PDFs for 3 bus trips}^{3} \]

\[ \text{Source: Ehrlich [5]} \]
time variability, and thus the RBT. It is implicitly incorporated with the AFC-based calculation method, through passengers’ actual journey times, and explicitly in the AVL-based method, through the bus trip wait time distribution (Equation 4.5). This factor’s magnitude is largely a function of the average headway, which is generally equivalent to the service’s scheduled headway. Therefore, regardless of the calculation method, the RBT is dependent on the scheduled headway.

This effect can be explicitly quantified in terms of the “minimum buffer time”, the typical passenger buffer time if the service was run perfectly on schedule, and all travel time variability was due to arrival-driven variation. Denoted the “minBT”, this can be defined in terms of the scheduled headway $h_{sch}$ and upper percentile $N$ as:

$$\text{minBT} = \left( \frac{N - 50}{100} \right) h_{sch}$$

(4.10)

For $N = 95$, this simplifies to 45% of the scheduled headway, or

$$\text{minBT} = 0.45 h_{sch}$$

(4.11)

The minimum buffer time, and arrival-driven wait time variation in general, is a factor not only for the RBT, but for all metrics that quantify the typical passenger’s buffer time.

### 4.2.6 Evaluation

**Representative of the Passenger’s Experience**

The RBT’s key strength is it is based on the journey buffer time, passengers’ primary means (according to the model proposed in Section 2.2.3) of planning for service reliability, based on past reliability experience. This implies that the RBT is representative of passengers’ reliability experience. Besides its specific relation to the buffer time, the RBT is generally representative of the passenger experience in that it measures travel time variability, rather than schedule adherence, and is thus suitable for non-schedule-reliant passengers. Furthermore, the RBT, by measuring only the travel time distribution above the median, focuses on variability causing passengers to be late, rather than that causing early arrival. Finally, the RBT’s basis on the entire journey time should ensure it captures all unreliability sources.

The RBT’s potential representativeness is limited in practice, however, by the meth-
ods currently available to calculate it. The AFC-based method is problematic because passengers’ recorded travel times are a function not only of service quality, but also their individual behavior: choice of entry/exit locations, slow vs. fast walking speed, etc. These factors should be generally consistent for an individual, but will vary significantly across individuals. Thus, by aggregating travel times across many individuals, the AFC travel time distribution used to calculate the RBT includes “cross-passenger” travel time variation not relevant to individual passengers. This will tend to increase the journey time distribution’s variance, leading to RBT values that over-estimate the typical passenger’s 95th percentile buffer time.

The AVL-based method, in contrast, is unaffected by “cross-passenger” variation, because it does not use passengers’ travel times directly. This method’s limitation is its inability to measure sources of unreliability not captured by AVL data: denied boardings due to overcapacity, congestion in stations, etc. These sources of delay can be significant in systems that are operating near capacity, so ignoring them can seriously misrepresent the passenger’s experience of reliability.

**Meaningful for Non-Experts**

The RBT’s basis on the buffer time also implies it should be meaningful to passengers and other non-experts. First, it is provided in terms of a travel time concept—the idea of budgeting extra time to account for unreliability—most travelers are familiar with, and is not mathematically complex. Second, it is “future focused”, providing information (how much time to budget, how often one will be late) helpful for planning future trips.

A potential problem with this application, however, is that most passengers probably do not precisely know their “expected” travel time for a given journey, and thus will not have a “baseline” to add the buffer time to. This can be resolved by providing the RBT alongside the 50th percentile travel time as proposed by Uniman, as shown in Figure 3-1.

**Comparability Across Services and Times of Day**

The ability to compare RBT values across services and times of day is limited, due to the RBT’s dependency on the scheduled headway. This makes it difficult to objectively compare reliability across services or times of day with different scheduled headways. For example, a RBT of 4 minutes for a service with 8 minute headways is exceptionally good service, as
it is very close to the minimum possible buffer time (3.6 min). However, a 4-minute RBT for a service with 2 minute headways could indicate reliability problems, as the service’s minimum buffer time is only 0.9 minutes.

**Time Period and Spatial Scope Flexibility**

The RBT is very flexible in terms of time periods for calculation. The RBT can be calculated for any time interval, for a given OD pair, for which at least 20 trips were taken. For a reasonably popular OD pair, this quota can be met even with short time-of-day intervals (30 minutes or less). In terms of spatial scope, the RBT is also quite flexible, because it is calculated at the OD-pair level. An aggregate RBT can then be calculated for any desired combination of OD pairs, with the average weighted by the OD flow.

**4.3 Other Passenger-Experienced Reliability Metrics**

Several other measures of passenger-experienced reliability have been developed in the literature, both in academia and industry. In the interest of brevity, this section is not a complete review of such measures; instead it focuses on those that have seen practical implementation, or extensive development in the research literature. Such measures are, in the author’s view, the most promising in terms of broader application in public transport practice. For each metric, its development and definition is reviewed, and its strengths and limitations assessed, utilizing the design objectives from Section 4.1. This review concludes with a summary of the common limitations of the reviewed metrics and the RBT, identifying opportunities for improved metrics to advance the state of the art.

**4.3.1 Excess Reliability Buffer Time (ERBT)**

**Description**

The ERBT was proposed by Uniman [2] as an extension of the RBT, with two objectives. First, to control for the “minimum variability” related to scheduled headway included in the RBT, limiting its comparability across services and time periods. Second, to separate service variability into “recurrent” and “incident-related” variability, to help analysts better understand the causes of unreliability; because the RBT includes all sources of variability, it provides little information in this regard.
The OD-level ERBT is calculated by classifying service performance on each day of a date period as being either “recurrent” or “incident-related”. The classification is done using the 95th percentile OD pair travel time as a performance indicator for each day, chosen because of its sensitivity and relation to the RBT. A stepwise regression method is then applied to find the days (defined as the “incident-related” days) that exhibit significantly higher 95th percentile travel time than other (“recurrent” performance) days. The ERBT is then defined as the difference between the RBT calculated for all days, and the RBT calculated only for days with recurrent performance, or “baseline” RBT reflecting “typical” performance, as given:

\[
ERBT = (RBT_{Overall} - RBT_{Recurrent})_{OD, Time Interval, Date Period} \quad (4.12)
\]

The OD-level ERBT can be aggregated to the line-level in the same manner as the RBT, by taking the average weighted by OD flow.

**Evaluation**

The ERBT seems to achieve its first goal. By excluding recurrent travel time variation from the metric, the ERBT should control for the minimum travel time variability, allowing for greater comparability across services and time periods. In terms of separating out “incident-related” delays, however, Frumin [1], in a critique of the ERBT, notes that:

...it is not clear that the world can be so easily and neatly divided into recurrent or incident-related conditions...incidents on a continuum of severity occur constantly. It is often unclear whether or not some perturbation to the service is normal and what in fact constitutes an “incident.”

Thus, there is a risk that the classification of unreliability in the ERBT calculation may mostly reflect assumptions and judgments “baked into” the classification process, rather than clear differences in service.

The exclusion of “normal” travel time variability from the ERBT also reduces its utility as a general measure of passenger-experienced reliability, in two ways. First, it removes information about non-incident related sources of unreliability that affect “normal” conditions—denied boardings due to under-capacity, dwell time variation, etc. Second, it removes the ERBT’s direct relation to the passenger’s buffer time, limiting its connection...
to passengers’ actual experience. This makes the ERBT more difficult to understand by the public, and less useful for planning journeys.

4.3.2 Excess Journey Time (EJT)

Description

The Excess Journey Time, or EJT, can be defined conceptually as the difference between the average passenger’s actual journey time, and their journey time had the service operated according to the timetable—in essence, the average time per trip “lost” due to unreliability.

This metric was first defined and used by the London Underground in 1999 [27], as an extension of the EWT metric (see Section 2.1.2). In this method, the EJT is the sum of the excess time for each journey component: wait time, in-vehicle time, access time, egress time, and transfer time. Train headways and running times (from AVL) are used to calculate excess wait and in-vehicle time, with excess wait calculated as in Equation 2.2. Denied boarding delays and excess access, egress, and transfer times are estimated using a combination of manual sampling at major stations and passenger flow models. The EJT is directly estimated at the OD pair level; line- or network-level EJT is calculated by averaging the constituent OD-pair EJTs, weighted by survey-based estimates of the passenger flow on each OD, similar to the line-level RBT (4.2.1). This formulation of EJT has been in use continuously since its development, as one of the Underground’s key performance indicators.

Frumin developed an improved formulation of the EJT (drawing on Chan’s [4] work on the EJT), in which the EJT is estimated for each individual trip, rather than for each OD pair. Passengers’ actual journey times are measured directly from AFC data (unavailable when London Underground first developed the EJT), rather than modeled. A path-finding algorithm then estimates the “scheduled” time for each trip, based on the timetable, origin, destination, and entry time. The algorithm assumes scheduled access and egress times are negligible. For transfer trips, the algorithm chooses the shortest path in terms of scheduled travel time. The OD-pair level EJT is the average individual EJT on that OD pair (for a given time period). From this, the line- or network-level EJT can be calculated by the average weighted by OD flow measured from the AFC data, rather than from a static OD matrix. This method improves upon the London Underground method by eliminating the uncertainty and potential bias in the models, manual sampling, and static OD matrix used to
estimate passenger crowding effects. Frumin also shows, through a mathematical derivation, that this method’s results should be independent of passenger incidence behavior.

Evaluation

The EJT has a number of strengths as a reliability measure:

- While complex to calculate, it is simple to understand conceptually.
- By examining the passenger’s entire journey, rather than just a single component, it includes all sources of unreliability. The exclusion of scheduled access and egress times in the Frumin method could bias the results, but this issue could be addressed by incorporating the Underground’s scheduled access and egress time estimates into the model used to calculate scheduled trip times.
- Frumin’s EJT, being independent of passenger arrival behavior, can potentially be a single method used for both high-frequency (i.e., random arrivals) and low-frequency (i.e., schedule-based arrivals) transit services.
- EJT can be calculated for trips involving transfers, and aggregated over any combination of OD pairs, providing excellent service scope flexibility.

The most significant limitation of the EJT is that is calculated in terms of schedule adherence, rather than absolute service variability. This raises the following issues:

- The EJT is of limited use for helping non-schedule-reliant passengers (e.g., commuters) plan their journeys.
- It is difficult to compare reliability objectively across schedule changes using the EJT, as the schedule itself affects the EJT values. For example, if scheduled running times or headways are increased, but actual operated service does not change, EJT will decrease, giving the appearance of a reliability improvement.
- Similarly, it is difficult to compare reliability objectively across services, if some routes have “padded” schedules, while others do not—the EJT differences may be more reflective of timetable differences, than differences in operated service.

The second key limitation of the EJT is that the estimates of scheduled access, egress, and interchange time do not reflect variation in passengers’ walking speeds, or entry and
exit gate locations. If a passenger walks slowly, or uses an entry gate farther from the platform, their trip’s EJT will be skewed upwards, regardless of the service provided. This effect could be significant in subway networks, such as the MTR, with very large stations, requiring significant walking time. This is problematic, as reliability metrics should only reflect the operator’s performance, and not individual passenger behavior.

4.3.3 Passenger Journeys on Time

Definition

The Passenger Journeys On Time, or PJOT, is a reliability metric developed by MTR as one of its key performance indicators. It estimates the percentage of passengers not affected by a 5 or more minute train delay, for a given subway line. It is calculated using the following equation:

\[
\text{PJOT} = \frac{P - P_D}{P} \times 100\% \tag{4.13}
\]

- \( P \) is the total passenger trips on the line, during the time period of interest.
- \( P_D \) is the total number of passengers affected by a train delay of 5 minutes or more, during the time period of interest.

\( P \) is derived from AFC data. \( P_D \) is derived from AFC data and train incident records indicating which trains suffered more than 5 minute delays; MTR’s train loading estimation tool (described in Section 6.2) is then applied to estimate the number of passengers on the delayed trains. MTR calculates the PJOT both at the line-level and the network level, and uses it to set annual reliability performance targets.

Evaluation

The key strength of the PJOT is its independence of the timetable and passenger incidence behavior. This makes the PJOT suitable for comparisons between services, or longitudinal studies of a single line’s reliability over schedule changes. There are several important limitations to the PJOT, however:

- Because it only captures incident delay, the PJOT ignores delays to passengers from other sources—denied boardings, long dwell times, congestion in stations, etc.
- The PJOT does not distinguish between trips delayed five minutes, and trips delayed by much more than five minutes. However, a 15 or 20 minute delay is likely to be perceived by passengers as far more onerous than a five minute delay, and will have a much larger impact on their perception of reliability.

- PJOT values are not helpful for planning journeys, because they are not provided at the OD-pair level, and are not provided in terms of travel time.

- The PJOT is only calculable at the line-level—thus, it is not flexible in service scope.

4.3.4 Conclusions from Review of Metrics

The main conclusion from the review of existing passenger-experienced reliability metrics is that no existing metric meets all the metric design objectives outlined in Section 4.1. In particular, each of the metrics reviewed is susceptible to one or more of the following concerns:

- Values potentially biased by passenger behavior (EJT, AFC-based RBT)

- Exclusion of important sources of unreliability perceived by passengers (AVL-based RBT, PJOT, ERBT)

- Limited comparability across different services and time periods (EJT, RBT)

- Difficult to relate to direct passenger experience (PJOT, ERBT)

These common limitations form the motivation for the development of new passenger-experienced reliability metrics presented in Chapters 5 and 6.
Chapter 5

Individual-based RBT

This chapter introduces, analyzes, and evaluates the first of the two new passenger-experienced reliability metrics developed in this thesis, the Individual-based Reliability Buffer Time, or IRBT. The objective is to develop an AFC-based measure of a typical passenger’s (95th-percentile) buffer time, not affected by cross-passenger variation—essentially, to remove passenger behavior bias from the AFC-based RBT. This approach is taken for two reasons. First, buffer time measures capture service variability, rather than schedule adherence. This is more relevant to MTR’s customers, as MTR does not publish schedules for its services (which are all high-frequency). Second, two buffer time measures have already been developed in the literature, the AFC-based and AVL-based RBT. Such a starting point allows the IRBT development to focus on solving the particular issues with these metrics, rather than on fundamental conceptual problems.

Section 5.1 defines the IRBT, including a variant, termed the Excess IRBT, that controls for scheduled headways. Section 5.2 compares the IRBT and AFC-based RBT, using OD and line-level results from MTR. Section 5.3 tests the sensitivity of the IRBT to delays, scheduled headways, and passenger demand. Section 5.4 discusses the IRBT’s most significant limitation, its requirement for a large group of frequent travelers for calculation, which results in significant restrictions on the time periods for which it can be calculated. Finally, Section 5.5 presents an overall evaluation of the IRBT, with respect to the design objectives for passenger-experienced reliability metrics presented in Section 4.1.
5.1 IRBT Definition

This section describes the background and defines the IRBT. The IRBT is based on the individual passenger’s travel time distribution for a particular OD pair, rather than the distribution across all OD pair passenger trips, as with the AFC-based RBT. This section thus begins (5.1.1) by examining the properties of the individual’s OD pair travel time distribution, using examples from MTR’s AFC data, and demonstrates the existence of “cross-passenger bias” in the complete OD pair AFC travel time distribution. Section 5.1.2 defines the IRBT, as a means of “averaging” the travel time distributions of many individuals, at the OD and line-level. Section 5.1.3 describes an AFC-preprocessing step to remove “outliers”.

5.1.1 Basis for the IRBT: The IBT

Theoretical Background

As explained in section 4.2.3, given an OD pair, time-of-day interval, and date period, one can obtain the set of travel times across all users from AFC data. Similarly, for a specific individual, OD pair, time-of-day interval, and date period, one can find the individual’s travel time distribution. An individual’s travel time variability can then be quantified, just as for the complete distribution, as the difference between their $N$th and 50th percentile travel times. To distinguish this individual travel time variability from the RBT, it is denoted by the IBT:

$$\text{IBT} = (\text{TT}_{N\%} - \text{TT}_{50\%}) \text{ID, OD, Time Interval, Date Period}$$

where $ID$ indicates an identifier for an individual, such as a smart card number.

The factors affecting both travel time distributions can be grouped into two categories: provided service, which includes service reliability, and passenger behavior. The latter includes factors such as normal walking speed, preferred station entry/exit location (locations can have different faregate to platform walk distances, affecting access and egress times), and familiarity with the system.

As noted in Section 4.2.6, these factors can vary significantly across passengers. It is expected, then, that behavioral factors will be a large component of overall travel time
variability. However, for *individuals*, these factors are unlikely to vary significantly from trip to trip. For example, most passengers are unlikely to walk slowly one day, and fast the next, or change their station access/egress paths significantly day-to-day. It follows that the behavioral component of the typical individual’s travel time variability should be small, relative to the service-driven component. Thus, measures of an individual’s travel time variability, such as the IBT, should almost entirely reflect their service experience, whereas AFC-based RBT values will reflect some (unknown) combination of passenger behavior and service variability. Assuming passenger behavior and service quality are not strongly correlated (and there is no reason to believe otherwise), this should result in RBT values being substantially higher than the corresponding average IBT values—the RBT, in effect, “over-estimating” the typical individual’s buffer time.

**Examples from MTR**

The above hypotheses were tested by analyzing individuals’ actual travel times from MTR AFC data. Sets of travel times and the corresponding AFC smart card numbers were obtained for 12 selected OD pairs and time periods, chosen to be representative of typical commuting on the MTR network, including the morning and evening peak hours, and both direct and transfer routes. Trips were then grouped by card number for each OD pair and time period. Such groups of trips are assumed to be trips taken by an individual, on that OD pair during the time period. Next, the 95th percentile IBT was calculated for all individuals (card numbers) with more than 20 trips (the minimum to calculate a 95th percentile RBT). A two-month date period was used for all OD pairs, to obtain a large number of such frequent travelers for each OD pair. This produced a set of IBT values for each OD pair/time period, from which an empirical IBT distribution was calculated. For comparison, the RBT was also calculated for each OD pair and time period.

The resulting IBT distributions all follow the general pattern exhibited in Figure 5-1 (for the Wong Tai Sin to Central OD pair, for trips starting between 8-9 AM in September and October 2012): roughly normally distributed, with significant right skew. It should be emphasized that the IBT distribution shown is *not* a travel time distribution, but rather a distribution of travel time variability. Furthermore, for every OD pair, the large majority of IBT observations were significantly lower than the corresponding RBT (indicated by the dashed line), consistent with expectations; the RBT exceeded the 85th-percentile IBT for
every sampled OD pair.

Figure 5-1: Representative IBT Distribution

Significant variation was found in the IBT values, with coefficients of variation in the 0.35-0.55 range for the sampled OD pairs. Some IBT variation is likely due to differences in the degree of “regularity” of individuals’ behavior; for example, some travelers may occasionally shop inside the paid area (increasing their access/egress times), while others may never do so. However, most is probably due to “random” variation in passengers’ departure times and days of travel\(^1\). This will lead to some passengers, by chance, experiencing fewer incidents/delays than others (e.g., someone is on vacation when a major incident occurs); such “lucky” passengers in this regard will likely experience less travel time variation—and thus, have lower IBTs—than the “unlucky” ones. Given this significant IBT variance, it follows that a random individual’s IBT may not be representative of the typical travel time variation for passengers on that OD pair.

Therefore, to quantify the “typical” passenger’s reliability experience using the IBT, it is necessary to calculate the entire IBT distribution, and then choose a measure of the central tendency that represents the “typical” passenger’s experience. Due to the significant right skew of the IBT distributions, the median is likely to be better than the average for this purpose.

\(^1\)Very few passengers will travel within a short time interval every day in an extended date period, due to holidays, errands, sick days, etc. The distribution of these “skipped days” should have no correlation with service quality, and so can be considered “random”.

68
5.1.2 Definition of the IRBT

The definition of the IRBT follows directly from the above results. For a given OD pair, time-of-day interval (denoted $T$), and date period (denoted $D$), the IRBT is defined as:

$$IRBT = \text{median}\left(\{IBT_I\}_{OD,T,D}\right)$$

(5.2)

where $\{IBT_I\}_{OD,T,D}$ is the set of 95th percentile IBTs for all individuals $I$ with 20 or more trips over the OD during time interval $T$ and date period $D$, with the IBT calculated as in equation 5.1. $I$ can thus be formally denoted, for the set of all travelers $t$, as

$$I = \{t \mid N_t \geq 20\}_{OD,T,D}$$

(5.3)

where $N_t$ is the number of trips taken by traveler $t$ over $OD$ during $T$, $D$. This definition effectively sets a 20-day minimum length for the date period $D$, given the low likelihood of passengers traveling more than once per day on the same OD pair, within a short time interval. The 95th percentile IBT is chosen for the reasons elaborated in section 4.2.2: good relevance for passengers (1 delay per month is tolerable for most passengers), realistic for operators (excludes “extreme” events and uncontrollable behavior), and data feasibility (reasonable sample size).

The IRBT, as defined, appears to fulfill its stated objectives. Like the AFC-based RBT, it expresses travel time variance in terms of the average passenger’s buffer time. However, by calculating this travel time variance at the individual level, rather than at the aggregate level, it removes the effects of cross-passenger variation. In effect, the IRBT is the average (median, technically) of every frequent traveler’s actual day-to-day reliability experience.

Spatial Aggregation of Performance

A line-level IRBT can be calculated, with respect to the OD-level IRBT, in the same manner as the line-level RBT (Section 4.2.1). Denoting by $Line$ the set of all OD pairs beginning and ending on the target line, and $IRBT_{OD}$ the IRBT for a particular OD pair, the line-level IRBT is defined as

$$IRBT_{Line} = \frac{\sum_{OD \in Line} f_{OD} \cdot IRBT_{OD}}{\sum_{OD \in Line} f_{OD}}$$

(5.4)
where Line is the set of same-line OD pairs and $f_{od}$ is the total OD pair demand for a given OD pair measured from AFC data. Total OD demand is used, rather than the number of captured frequent travelers (i.e., size of set $I$), because the number of frequent travelers per OD pair is not necessarily proportional to the total travelers per OD pair; some OD pairs are more “frequent traveler-heavy” than others. As the overall aim is to represent the typical passenger’s reliability experience, it is preferable to weight OD IRBT values by the total travel demand.

This method can be extended to any desired level of spatial aggregation, by replacing Line with the set of OD pairs comprising the desired spatial scope. Examples include:

- **Network-level**: all OD pairs in the network
- **Line-direction**: all same-line OD pairs in one line direction (e.g., Island Line east-bound)
- **Transfer pattern**: all OD pairs beginning on one line (or line-direction), and ending on another (e.g., Island Line to Kwun Tong Line).
- **Line segment**: all OD pairs beginning and ending on a particular line section

### 5.1.3 Outlier Removal

The influence of passenger “regularity” behavior on the IBT, and thus the IRBT, is detrimental to the IRBT’s effectiveness as a reliability metric, as it should ideally measure only the service performance. To help mitigate this effect, a “pre-processing” step can be added to the IRBT calculation to remove “outliers” from the AFC data which are, almost certainly, not a result of service reliability.

In the pre-processing step, trips are removed from the AFC data if the travel time meets the interquartile range method criteria for being considered an extreme outlier [28]:

$$TT > Q3 + 3 \times IQR \quad \text{or} \quad TT < Q1 - 3 \times IQR$$

where $TT$ is the travel time, $Q1$ and $Q3$ are the first and third quartile travel times, and $IQR$ is the interquartile range ($Q3 - Q1$). Outlier removal is applied separately for each day of the date period, so the trips considered were taken under similar service conditions; this should result in the identified outliers being almost entirely behavior-related, rather
than service-related. This process was applied for all IRBT values shown in this thesis. The effect of this pre-processing on the IRBT result is generally relatively small, however, and can be skipped in the IRBT calculation process if necessary.

5.1.4 Excess IRBT

The IRBT, as described, should include all sources of travel time variation due to unreliability. However, the IRBT also captures a major source of travel time variability not related to reliability: "random" variation in passengers’ arrival times. As described in Section 4.2.5, this passenger arrival-driven variability’s magnitude is largely a function of the scheduled headway $h_{sch}$, and can be expressed as a “minimum buffer time”, denoted minBT. For $N = 95$, this is equal to 45% of the scheduled headway.

For the IRBT, the minBT represents the minimum possible IRBT, at both the OD and line-level, given the headway of the service being evaluated—essentially, the IRBT achieved if the service was operated with “perfect” reliability. From this, one can define “Excess IRBT” as the difference between the IRBT and minBT, as follows

\[
\text{Excess IRBT} = \text{IRBT} - \text{minBT}
\]  

(5.5)

The Excess IRBT should, in theory, represent the portion of passenger travel time variability caused by service unreliability, and the minBT the schedule-driven “baseline” travel time variability that will be present regardless of the service reliability. This results in two potential advantages of the Excess IRBT over the IRBT: First, the value should more directly measure the reliability performance of the operator. Second, by controlling for the scheduled headway, the Excess IRBT is much more amenable to comparison across lines, or across times of day for a given line. Put another way, the IRBT is a relative reliability measure, with respect to the scheduled headway, whereas the Excess IRBT should be an absolute reliability measure, independent of the schedule.

The Excess IRBT, however, has a potentially major disadvantage: it does not relate directly to the passenger’s experience. The IRBT represents the typical passenger’s buffer time, a concept with direct relation to how passengers might plan their journeys given unreliability. The Excess IRBT, on the other hand, represents a more abstract concept, less related to passengers’ experiences—and thus, less meaningful to passengers and the public.
Thus, while the Excess IRBT has great potential as an analytical tool, it is not ideal as a general-purpose reliability metric, especially one to be shared publicly.

5.2 IRBT and AFC-Based RBT: Results from MTR

The section demonstrates the calculation of the AFC-based RBT and IRBT for MTR OD pairs, lines, and transfer flows, and compares the results of the two metrics. The findings show that, in accordance with expectations, the IRBT’s buffer time estimates are significantly smaller than the RBT’s, at all spatial scales. This analysis is divided into three parts: the first compares the two metrics’ results at the OD pair-level, in terms of both magnitude and trends over time; the second similarly compares the two metrics at spatially aggregate levels; the third summarizes the findings, and relates them to the definition of the IRBT.

5.2.1 IRBT/RBT Results: OD Pair-Level

Figures 5-2, 5-3, and 5-4 show the AFC-based RBT and IRBT results for three selected MTR OD pairs: Wan Chai to Quarry Bay, Kowloon Bay to Lam Tin, and Wan Chai to Jordan (see the MTR map, Figure 1-1). The first two are single-line OD pairs, while the third involves a single transfer at a very congested station, Admiralty. The (a) subfigures show the RBT and IRBT for hourly time intervals (7-8 AM, 8-9 AM, 9-10 AM, etc.), for a single date period, September 1 to October 15, 2012. The (b) subfigures show the IRBT and RBT for a single time interval, 6-7 PM, over rolling 6-week date periods (e.g. 8/5-9/15, 8/12-9/22, 8/19-9/29, ...); the x-axis dates indicate the end of each 6-week period.

For all OD pairs, and all times of day, the IRBT values are consistently and significantly lower than the RBT. The magnitude of this RBT–IRBT difference varies greatly across OD pairs, and time of day for a single OD pair, ranging from about 30 seconds to over 6 minutes. The RBT–IRBT difference is fairly consistent, however, across date periods (for a given OD pair and time interval).

There is no clear pattern in the RBT–IRBT difference’s variation by time of day, but its variation among OD pairs can potentially be explained by variation in total in-station walking distance (platform access, platform egress, transfer) among OD pairs. As walking distance increases, walking speed variation should induce greater total travel time variation.
This agrees with the results; the Wan Chai and Quarry Bay stations require more walking distance than Kowloon Bay and Lam Tin, corresponding to the greater RBT–IRBT differences in Figure 5-2 compared to Figure 5-3. This can also explain the even greater difference seen in Figure 5-4, as that OD pair’s transfer adds a transfer walking segment not needed for the other OD pairs.

Figure 5-2: Wan Chai to Quarry Bay RBT and IRBT (Island Line)

Figure 5-3: Kowloon Bay to Lam Tin RBT and IRBT (Kwun Tong Line)

While the magnitudes of the IRBT and RBT may be significantly different, the time of day and date period trends in their values are generally similar. For example, in Figure 5-4a,
both the IRBT and RBT register a significant “spike” from 6-7 PM, despite the two metrics’ absolute magnitudes being 4-5 minutes apart at the time. This indicates that trends in the IRBT, which reflect frequent travelers’ reliability experience, are generally representative of trends in the overall passenger population’s reliability experience. This is an important finding, as the number of frequent travelers captured by the IRBT is typically orders of magnitude smaller than the overall passenger population size captured in the AFC-based RBT (for the results in the (a) subfigures, the average RBT group size was 4314, and the average IRBT frequent traveler group size was 32), which raises concern about the IRBT’s ability to represent the “typical” passenger’s experience. This issue of representativeness, and strategies to address it, is discussed in more detail in Section 5.4.

The RBT–IRBT difference analysis above can be generalized by looking at the statistical properties of 230 pairs of IRBT and RBT values for various OD pairs and times of day (176 non-transfer, 54 transfer). The average RBT–IRBT difference was found to be 2.11 minutes, or 43% of the RBT. This provides strong evidence for the hypothesis that the IRBT will estimate lower travel time variability, compared to the AFC-based RBT. The difference was also found to be highly variable, with $\sigma=1.83$ minutes and $\sigma=16\%$, respectively. When broken down by transfer OD pairs and non-transfer OD pairs, it was found that the RBT–IRBT difference was significantly higher for transfer OD pairs, indicating that transfers significantly increase cross-passenger travel time variation. These results are summarized in Table 5.1, with 95% confidence intervals specified for averages.
Table 5.1: RBT–IRBT Difference Statistics

<table>
<thead>
<tr>
<th>OD Type</th>
<th>Average</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minutes</td>
<td>% of RBT</td>
</tr>
<tr>
<td>All ODs</td>
<td>2.11 (±0.24) min</td>
<td>43% (±2%)</td>
</tr>
<tr>
<td>Non-transfer</td>
<td>1.67 (±0.20) min</td>
<td>39% (±2%)</td>
</tr>
<tr>
<td>Transfer</td>
<td>4.14 (±0.53) min</td>
<td>57% (±3%)</td>
</tr>
</tbody>
</table>

5.2.2 IRBT/RBT Results: Aggregate Level

Figures 5-5, 5-6, and 5-7 show examples of spatially aggregate IRBT and RBT results from MTR. Figures 5-5 and 5-6 show line-level IRBT and RBT results for two of MTR’s busiest lines, the Island Line and the Kwun Tong Line, while 5-7 shows the aggregate transfer-level RBT and IRBT for all OD pairs beginning on the Island Line and ending on the Tsuen Wan Line, one of the most common transfer movements in the MTR system. As with the figures in the previous section, the (a) figures show the IRBT and RBT at hourly time-of-day intervals, for a fixed date period, while the (b) figures show results for a single time interval (6-7 PM) over rolling 6-week date periods. These values are aggregated over 182, 210, and 196 OD pairs, respectively.

![Figure 5-5: Island Line RBT and IRBT](image)

(a) Aug.-Sept. 2012, hourly
(b) 6-7 PM, rolling 6-week periods

The spatially aggregate-level results show a IRBT-RBT relationship similar to that found at the OD-pair level. The RBT is consistently higher than the IRBT, across all time-of-day intervals and date periods. As with OD pairs, the RBT-IRBT difference varies more
across hours of the day than across different date periods (this is particularly striking in Figure 5-6). The RBT-IRBT difference is significantly higher for the aggregate transfer flow than for the line-level results, which is consistent with the earlier finding (in Table 5.1) that the RBT-IRBT difference is generally higher for transfer OD pairs than for non-transfer OD pairs. Further following the OD pair patterns, the aggregate IRBT and RBT exhibit similar time interval and date period trends, despite their difference in magnitude. This suggests that the IRBT-RBT correlation seen at the OD pair level is representative of OD pair RBTs and IRBTs across the MTR system.
5.2.3 Conclusions on IRBT/RBT Relationship

The above results support the hypothesis behind the design of the IRBT, that the AFC-based RBT significantly overstates the level of unreliability experienced by passengers, by not controlling for cross-passenger variation when estimating the typical passenger’s travel time variability. The difference between a given pair of RBT and IRBT values, according to the theory in Section 5.1.1, should effectively be the magnitude of cross-passenger variation present in the RBT, and excluded from the IRBT. Thus, the results in Table 5.1 indicate that on average, more than 40% of the AFC-based RBT’s value is due to cross-passenger variation, a factor not part of the individual passenger’s reliability experience. Furthermore, the magnitude of cross-passenger variation appears to be related to walking distance and the presence of transfers, consistent with the hypothesis that most cross-passenger travel time variation is due to walking behavior variation. Based on these findings, it can be concluded that the IRBT will generally be a more accurate estimator of the typical passenger’s reliability experience than the AFC-based RBT.

5.3 Sensitivity to Operational and Demand Factors

A theoretical advantage of the IRBT over “traditional” operational reliability measures is that it should capture both operational and passenger demand-related sources of travel time variability affecting the passenger’s travel experience. The objective in this section is to test if the IRBT can achieve this in practice; towards this end, the section explores the IRBT’s sensitivity to incidents (i.e. train delays), in Section 5.3.1; the scheduled headway, in Section 5.3.2; and passenger demand (as measured by line passenger flow), in Section 5.3.3. This type of analysis is critical in evaluating the utility of an “untested” new metric; a reliability metric that is not responsive to “obvious” sources of unreliability is, clearly, an ineffective metric.

5.3.1 Sensitivity to Incidents

To analyze the IRBT’s sensitivity to major incidents (breakdowns, signal problems, passenger sickness, etc.), the AM peak hour (or half hour) IRBT was calculated for a number of OD pairs and line-directions, for 6-week rolling date periods over 5-6 months. The MTR incident logs were then referenced to find all major incidents that occurred on the lines
Figure 5-8: IRBT and Major Incidents, Island Line WB OD Pairs

Figure 5-9: IRBT and Major Incidents, Island Line WB and EB

Figure 5-10: IRBT and Major Incidents, Kwun Tong Line WB and EB
of interest, during the time-of-day intervals and date periods. Selected results are plotted in Figures 5-8, 5-9, and 5-10. Figure 5-8 shows the IRBT results for several westbound Island Line OD pairs, and the dates of major Island Line westbound incidents, for the 8-9 AM time interval. Figure 5-9 shows 8-9 AM line-direction level IRBT results for the Island Line westbound and eastbound, alongside the same Island Line westbound incident dates. Figure 5-10 similarly shows the 8-8:30 AM line-direction IRBT results for both Kwun Tong Line directions, along with major incident dates for the Kwun Tong Line westbound.

Incidents were found to correspond to an IRBT increase of about 20-45 seconds (about 20-35%) after the incident, at both the line and OD level. This effect persists for all date periods containing the incident. For the 6-week periods shown, this results in an elevated IRBT for 6 consecutive periods; if the date period is shorter (or longer), the duration of an incident’s effect changes accordingly. When multiple major incidents occur in the same date period, as is the case for the Island Line westbound with incidents on Oct. 3 and Nov. 1, and Kwun Tong Line westbound on Dec. 7, Jan. 7, and Jan. 29, the incident effects appear to be additive. This effect can be explained by the fact that only a fraction of frequent travelers are affected by a particular incident, since such travelers do not generally travel every day in the date period. Such “incident-affected” passengers should have higher IBTs than “unaffected” passengers. The more incidents that occur within a particular date period, the higher the fraction of “incident-affected” passengers, resulting in a higher median IBT, and thus, higher IRBT. This effect is also consistent with a priori expectations of passengers’ reliability experience; i.e. more frequent major incidents results in greater unreliability.

A noteworthy result seen in Figure 5-9 is that the two westbound Island Line incidents, in addition to affecting the westbound Island Line IRBT, also caused a smaller, but noticeable, increase in the eastbound IRBT. This indicates that major westbound delays had “knock on” effects causing minor delays to eastbound trains.

These findings confirm that the IRBT, as predicted, is sensitive to incidents. Operational incidents should cause passengers’ 95th percentile travel times to rise, thus increasing the IRBT. The magnitude of the response is only moderate, however, with a major incident adding only 25-35% to the “baseline” IRBT. This indicates that while incidents can significantly affect the IRBT, the majority of the IRBT represents variation not due to major incidents.
5.3.2 Sensitivity to Scheduled Headways

As described in Section 4.2.5, for high-frequency transit services, a significant portion of the IRBT should be due to the random arrival times of passengers at platforms. Furthermore, this portion, denoted the “Minimum Buffer Time”, or minBT, should theoretically be equal to 45% of the scheduled headway (Equation 4.11). Thus, it is expected that the IRBT will exhibit significant correlation with the scheduled headway. To test this, the IRBT was calculated at half-hour time intervals for two line-directions: Island Line westbound and
Tsuen Wan Line northbound (date period: Sept.-Oct. 2012). The scheduled headways for these line-directions were then used to calculate a corresponding minBT for each half-hour interval (if headways changed within a half-hour interval, the average headway was used).

The results, in Figures 5-11a and 5-12a, indicate a very strong relationship between the IRBT and scheduled headway. Statistically, the correlation between the minBT and IRBT is significant \( r = 0.767 \). The relationship appears to be strongest in the off-peak, when the two metrics not only exhibit similar trends, but also very similar magnitudes—in fact, the off-peak IRBT and minBT values in Figures 5-11a and 5-12a are almost identical. This suggests that almost all off-peak passenger travel time variation is attributable to passengers’ random arrival times.

During the morning and evening peak hours there is a much greater gap between the IRBT and minBT values. This is shown explicitly in Figures 5-11b and 5-12b, which plot the difference between the IRBT and minBT, termed the Excess IRBT (see Section 5.1.4), for each half-hour period. For both line-directions, the Excess IRBT has clear peaks around 8:00-10:00 and 16:00-20:00, while dropping to near zero from 10:00-16:00 and 20:00-23:30. This indicates that the connection between the IRBT and scheduled headway (via the minBT) is much weaker during the peak than off-peak, and that passenger travel time variation is largely driven by other factors during the peak, such as passenger demand.

**Broader Line Reliability Implications**

As described in Section 5.1.4, the Excess IRBT should be an absolute measure of passenger-experienced unreliability, independent of the schedule. Thus, the Excess IRBT being significantly higher during the peak times of day indicates that passenger-experienced unreliability on the Island and Tsuen Wan Lines was consistently greater during the peak, compared to the off-peak. This agrees with expectations; higher passenger demand should lead to more variable platform waiting times (because of denied boardings) and access/egress times (because of station congestion), even if good operational reliability is maintained. This result is thus encouraging from the standpoint of assessing the “validity” of the IRBT metric.

It should be noted that the Excess IRBT results, in absolute terms, are indicative of an exceptionally high level of reliability being provided by MTR. During the off-peak, the Excess IRBT ranges from 0 to 0.3 minutes—essentially perfect service, from the passenger’s standpoint. Even in the peak of the peak, the Excess IRBT does not exceed 1.3 minutes, an
amount that would still be barely noticeable, from the standpoint of the average passenger. Excess IRBT figures for the average North American system, the author suspects, would be much greater than seen for MTR.

5.3.3 Sensitivity to Passenger Demand

The IRBT, as a reliability metric, should reflect the substantial delays to passengers that occur when passenger demand nears a transit system’s design capacity—denied boardings, station congestion, and long dwell times. Thus, it is expected that passenger demand will have a significant influence on IRBT results. MTR’s system is a good environment to test this, as the busiest MTR lines regularly approach their design capacity during the peak hours. Such an analysis is presented below, examining the IRBT’s sensitivity on MTR to time of day demand variation, as well as long-term demand variation.

Time of Day Demand Variation

To analyze the IRBT’s sensitivity to passenger demand, the average passengers per hour per direction, or PPHPD\(^2\), was estimated for the same line-directions and time periods shown in Figures 5-11 and 5-12, at the peak load point (i.e., busiest station-to-station link) for each line\(^3\). For reference, the line capacity in PPHPD was also calculated throughout the day, based on the scheduled headways and MTR official train capacity (2000 per 8-car train, as of 2013). PPHPD was chosen as the passenger demand measure because it is a commonly-used measure of passenger flow and line capacity. The results are plotted in Figures 5-13 and 5-14, alongside the Excess IRBT from Figures 5-11b and 5-12b. The Excess IRBT is used, rather than the IRBT, because it allows for objective comparisons across periods with different scheduled headways.

The results indicate a strong correlation between Excess IRBT and average passenger flow, especially during the morning and evening peak periods\(^4\). Statistically, the correlation

\(^2\)The PPHPD is defined as the number of passengers carried by a line/route past a point in one direction, per hour.

\(^3\)The passenger flow data was derived from MTR’s train load estimate data. For more information about MTR’s train load estimate algorithm, see Section 6.2.

\(^4\)Note: The apparent misalignment between the Island Line westbound PM peak passenger capacity and peak passenger flow is explained by a combination of two factors: First, eastbound peak PM passenger flow is significantly higher than westbound, but occurs essentially at the same time (e.g., the peak half-hour for both directions is 18:30-19:00). Second, trains enter service from a depot near the eastern terminal, about 30 minutes from the western terminal. Thus, to serve the westbound demand effectively and efficiently, trains are entered into service eastbound about 30 minutes before peak demand begins, and start being taken out
between the average passenger flow and Excess IRBT was found to be 0.799, for both line-directions. This relationship implies that the “basic” IRBT is also very sensitive to passenger demand. Between the strong IRBT-minBT correlation in the previous section, and the strong Excess IRBT-passenger flow correlation demonstrated here, it appears that the large majority of time-of-day IRBT variation (at least at the line-direction level) can be explained by scheduled headways and passenger demand.

Figure 5-13: ISL WB Excess IRBT, Average Passenger Flow (at Peak Load Point), and Passenger Capacity

Figure 5-14: TWL NB Excess IRBT, Average Passenger Flow (at Peak Load Point), and Passenger Capacity

The IRBT-passenger demand relationship can also be explored in terms of passenger

of service about 30 minutes before peak demand ends, with the reduced frequency still sufficient to serve the lower eastbound peak demand.
capacity utilization, i.e. the ratio of passenger demand to line capacity. In theory, capacity utilization should better predict denied boarding effects than raw passenger demand, as denied boardings should only occur when demand nears (or exceeds) capacity; even moderate demand can cause denied boardings if service frequency is inadequate. Figures 5-15 and 5-16 show line-direction Excess IRBT and capacity utilization at the peak load point, using the data from Figures 5-13 and 5-14. It is apparent that, while the capacity utilization appears to peak at the same time as the IRBT, the correlation is much weaker than between the Excess IRBT and average passenger flow at the peak load point.

A potential explanation is that denied boardings should only occur when capacity utilization is very high; thus, trends in capacity utilization below some “threshold” will have little effect on the IRBT (probably around 75%—because passengers are not evenly dis-
tributed in cars, denied boardings can occur at capacity utilization levels significantly below 100%). For Figures 5-15 and 5-16, the scheduled headways are such that the highest capacity utilization happens to occur during the highest absolute demand, and low capacity utilization during low absolute demand, resulting in the high capacity utilization effects better matching passenger demand than the utilization itself.

Another potential cause is that a significant portion of demand-related delay comes from station congestion, rather than denied boardings. Station congestion, related to the capacity of stairwells, corridors, and escalators, should be independent of scheduled headways, and thus is better explained by absolute passenger demand than line capacity utilization.

**Long-Term Demand Variation**

It is also useful to test the IRBT’s sensitivity to longer-term demand variation, such as variation in demand over the year. Such information is important for a system with consistent ridership growth over time, such as the MTR, in order to anticipate the effects of future growth on passengers’ experienced reliability.

To test the effects of long-term demand variation, the 8-9 AM and 6-7 PM IRBT was calculated for rolling 9-week periods over the course of an entire year (July 2012-July 2013), for both line-directions on the Island Line and Kwun Tong Line. Longer 9-week date periods were used for this analysis (compared to the 6-week periods used previously), in order to better control for short-term reliability trends (e.g., incidents), and highlight longer-term trends; this topic is discussed further in Section 7.4. The intra-line ridership (i.e., number of trips beginning and ending on the line) was then calculated for the same line-directions and time periods. This was used as a proxy for the total line passenger demand, because insufficient MTR train load data was available to calculate the average passenger flow for the entire year; the two should be highly correlated, so the substitution is not unreasonable. With these two data sets, trends in line-direction IRBT and passenger demand were compared over an extended period of time.

This comparison found no significant correlation \( (r = 0.0013) \) between long-term IRBT variation and long-term passenger demand variation, as seen in Figure 5-17. The magnitude of the long-term demand variation observed was also quite small, compared to the time-of-day variation, no more than ±10% from the mean. This suggests that the IRBT, while clearly sensitive to large changes in passenger demand (especially when a line nears
capacity), is fairly insensitive to small changes in passenger demand. This finding seems reasonable, as minor demand variation shouldn’t significantly increase the level of demand-related delays, unless the baseline demand was near a “critical flow” where denied boardings and/or station congestion begins if exceeded.

![Figure 5-17: ISL 8-9 AM IRBT and intra-line demand, by direction](image)

### 5.4 Frequent Traveler Group Size Limitations

The IRBT’s use is constrained by the need to obtain a sufficient number of frequent travelers, or “frequent traveler group size”, to ensure the IRBT value accurately reflects the transit service provided. As shown in Section 5.1.1, IBTs for a particular OD pair can vary substantially among passengers, due to passenger behavior and random variation in passengers’ travel days and times. For a large frequent traveler group size, as shown in Figure 5-1, these factors effectively “average out”, and the median IBT—i.e., the IRBT—should reflect the reliability of the transit service provided. However, if the group size is too small, these factors can significantly affect the IRBT, thus leading to IRBT patterns not reflective of the service reliability.

The purpose of this section is to investigate how frequent traveler group size requirements limit the time periods and services for which the IRBT can be calculated. The analysis begins, in Section 5.4.1, with establishing a rough minimum group size for the OD-level and line-level IRBT on MTR, by quantifying the “noise” in the IRBT at various frequent
traveler group sizes. In Section 5.4.2, these standards are applied to OD pair and line IRBTs calculated for various time-of-day intervals and date periods, for one of MTR’s busiest lines (Island Line); it is found that the minimum group size significantly constrains the times and locations for which OD pair and line IRBTs can be calculated with acceptable accuracy.

5.4.1 Frequent Traveler Group Size Standards

Methodology

Establishing appropriate minimum frequent traveler group sizes is not a straightforward task, because there is no clear statistical method to quantify the “uncertainty” in IRBT values or trends as a function of the group size, through something like $t$-tests. This is because the “sample” of frequent travelers is not a random sample, but rather the subset of travelers traveling more frequently than the “general population”, and because the IRBT is a measure of the median, not the mean. Furthermore, such uncertainty would need to be quantified in terms of IRBT magnitude and relative IRBT trends, as both are of interest for reliability measurement applications.

Given these issues, a more subjective, and thus less rigorous, approach is taken to establish appropriate frequent traveler group size minimums. The method used to estimate the OD size minimums is as follows:

1. Calculate the IRBT for a number of OD pairs and time intervals with large group sizes (i.e. $\geq 100$ passengers), for rolling date periods.
2. Re-calculate the selected OD pair IRBTs using a set of random samples from the IBT distribution of different sizes.
3. Calculate the Root Mean Square Deviation (RMSD) and correlation $r$ between the actual IRBT values and the IRBT values for each of the sample sizes.
4. Select the minimum sample size for which the RMSD is acceptably low, and the correlation is acceptably high, for the ODs examined. This becomes the minimum frequent traveler group size standard for similar OD pairs.

The idea behind this method is that the actual IRBT values provide a reference against which the sample IRBT values can be compared, with the differences between the actual and sample IRBTs then representing the “error” introduced by the small group sizes. This
error, it is hypothesized, is equivalent to “noise” in the IRBT occurring when there are insufficient frequent travelers over a target OD pair. The magnitude of this error should be reflected in the RMSD, while the error in relative trends should be captured by the correlation. By choosing acceptable maximum RMSD and minimum correlation values, and setting the group size standard as the minimum to meet these “thresholds”, one is able to establish standards for both absolute and relative error.

The choosing of the RMSD and correlation thresholds will, of course, be subjective, depending on the particular application requirements, and the analyst’s judgment. For this analysis, a maximum RMSD of 30 seconds, and minimum correlation of 0.85, is proposed. A 30-second difference should be imperceptible to passengers, while a correlation of 0.85 should be sufficient to capture important reliability trends.

The method for setting minimum frequent traveler group sizes for line-level IRBTs is essentially the same as for OD pairs. The only difference is the sampling process used: IBTs are randomly sampled across all intra-line OD pairs (i.e., the population consists of passengers from all OD pairs). Larger sample sizes are also used, in accordance with the need to represent a variety of reliability conditions. The sampled IBTs are then grouped by OD, and used to calculate a line-level IRBT in the same manner as the “normal” line-level IRBT. Because of the much larger number of passengers present at the line-level, this test can be applied to essentially any line or line-direction, rather than just the busiest ones, as with OD pairs. Given the greater importance of line-level metrics for reporting and reliability management, RMSD and correlation standards should be tighter at the line-level than at the OD-level. For this analysis, the line-level standards of 15 seconds and 0.95 are proposed, respectively.

**OD Pair Results**

To establish a frequent traveler group size standard for OD pairs on MTR, the IRBTs for nine OD pairs from 8-9 AM and 6-7 PM (the morning and evening peak hours) were analyzed, over 16 6-week periods (Sept.-Dec. 2012). All were heavily traveled, resulting in large frequent traveler group sizes ranging from 112 to 559 on average per 6-week period. All were non-transfer OD pairs on the Island, Kwun Tong, or Tsuen Wan Lines. In addition to the “normal” IRBT, the IRBTs were calculated for frequent traveler sample sizes of 15, 30, 45, 60, and 75. The upper sample size was set to 75 because most OD-level frequent
traveler group sizes fall below this value; 15 was set as the lower bound, because it seemed unreasonable to expect accurate results below this value. Results for two OD pairs are plotted in Figure 5-18 (for legibility, only samples sizes of 15, 30, and 60 are shown).

The RMSD and correlation results are presented in Table 5.2. As expected, as the passenger sample size increases, the RMSD decreases, and the correlation increases. All sample sizes meet the RMSD threshold of 30 seconds, but the minimum correlation of 0.85 is not met at a sample size of 15. The poor correlation at the 15 sample size is apparent in Figure 5-18, where significant divergence is seen between the sampled IRBT and actual, to an extent unacceptable for practical use. Based on these findings, a group size standard of 30 is proposed for the OD pair IRBT. This standard could certainly be refined, however, with an examination of many more OD pairs, and more sample size gradations. This analysis should be considered more a “proof of concept” of a minimum frequent traveler group size setting process, rather than a definitive recommendation for MTR.

**Line-Level Results**

To establish a line-level frequent traveler group size standard, the line-direction IRBTs were calculated and analyzed for six line-directions: Island Line both directions, Kwun Tong Line both directions, and Tsuen Wan Line both directions. All were calculated from 8-9 AM, for the same 6-week date periods used for the OD pair analysis. As expected, the line frequent traveler group size is much greater than at the OD pair level, averaging about 8000 the six
Table 5.2: Root-Mean-Square Deviation and Correlation between OD-pair IRBT and Sampled OD-pair IRBT

<table>
<thead>
<tr>
<th>Passenger Sample Size</th>
<th>RMSD (sec)</th>
<th>Correlation $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>14.1</td>
<td>0.68</td>
</tr>
<tr>
<td>30</td>
<td>7.9</td>
<td>0.89</td>
</tr>
<tr>
<td>45</td>
<td>6.8</td>
<td>0.93</td>
</tr>
<tr>
<td>60</td>
<td>5.4</td>
<td>0.95</td>
</tr>
<tr>
<td>75</td>
<td>4.5</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 5-19: Line-level IRBT, Normal and for Limited Passenger Sample Size

line-directions. Besides the “normal” IRBT, the line-direction IRBTs were calculated at frequent traveler sample sizes ranging from 100 to 1000, with 100 assumed to be the lower bound on a feasible minimum group size. Above 300, no significant deviation from actual was seen. Figure 5-19 shows the Island Line westbound and Kwun Tong Line eastbound results for sample sizes of 100, 200, and 300.

The line-direction level RMSD and correlation results are presented in Table 5.3. As with the OD IRBT, as the passenger sample size is increased, the RMSD decreases and the correlation increases for the sampled line-direction IRBT. All sample sizes meet the RMSD threshold of 15 seconds, and the minimum sample size to meet the correlation threshold on 95 is 150. Thus, a minimum line/line-direction frequent traveler group size of 150 is proposed—a value that should be viewed with the same caveats as the previous OD pair recommendation.
### Table 5.3: Root-Mean-Square Deviation and Correlation between Line-direction IRBT and Sampled Line-direction IRBT

<table>
<thead>
<tr>
<th>Passenger Sample Size</th>
<th>RMSD (sec)</th>
<th>Correlation $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>8.4</td>
<td>0.94</td>
</tr>
<tr>
<td>150</td>
<td>7.3</td>
<td>0.96</td>
</tr>
<tr>
<td>200</td>
<td>5.2</td>
<td>0.98</td>
</tr>
<tr>
<td>250</td>
<td>4.5</td>
<td>0.99</td>
</tr>
<tr>
<td>300</td>
<td>3.9</td>
<td>0.99</td>
</tr>
</tbody>
</table>

An important point about this result is that for a line-direction level sample size of 150, the average group size per OD pair is only 1.4. This indicates that it is not necessary for every, or even a substantial minority, of OD pair IRBTs to meet the OD pair minimum group size requirement when averaged to calculate an aggregate-level IRBT. Evidently, then, having a large group size effectively “averages out” behavior and chance factors, regardless of which OD pairs the individuals are traveling between.

#### 5.4.2 Group Size Constraints on Time Periods

A number of factors affect the frequent traveler group size for an OD pair or line, including:

- **Time of Day:** Far more frequent travelers will be found during the peak commuting times, immediately before and after the “normal” work day, than during midday and evening off-peak times. Not only is the total volume of passengers larger during peak times, but the proportion of travelers who are frequent travelers is also likely to be greater.

- **Time Interval Length:** Longer time-of-day interval lengths (30 minutes, 1 hour, etc.) should result in larger group sizes, for two reasons. First, as the interval length increases, more trips will be captured simply because a larger portion of the day is captured. Second, longer time-of-day intervals will capture frequent travelers with more variable departure times, whose departure times would be split if time intervals were smaller. For example, if a traveler’s departure times are randomly distributed from 8:20 AM to 8:40 AM over 30 days, and the time-of-day intervals are 8-8:30 AM and 8:30-9 AM, only 15 trips will be captured by each interval, a number insufficient to calculate an IBT. However, if the time interval is 8-9 AM, all 30 trips will be captured, and their IBT can be calculated.
- **Date Period Length:** Assuming passengers travel between a particular OD pair once per day, 20 travel days are required per passenger to calculate an IBT. For the “ideal” commuter traveling weekdays, this can be achieved in 4 weeks. However, most frequent travelers are more irregular in their commuting habits; they take vacations and sick days, arrive early (or late), work from home occasionally, etc. Thus, for most frequent travelers, a date period longer than 4 weeks will be required to obtain 20 travel days. As the date period’s length is increased, more irregular travelers will be included in the frequent traveler group.

- **Location/Direction:** Some OD pairs will be more popular with frequent travelers (e.g., residential to CBD) than others. For lines/routes serving the CBD, the relative popularity of OD pairs depends on the time of day, with inbound OD pairs most trafficked in the morning, and outbound OD pairs in the evening.

These factors, combined with the frequent traveler group size limits discussed in Section 5.4.1, will impose limits on the time periods and services for which the IRBT can be calculated. To demonstrate these limits in the MTR context, group sizes were analyzed for a single (heavily-traveled) line-direction, Island Line westbound. For a set of four time interval lengths (15 min., 30 min., 1 hour, 2 hours), four date period lengths (4 weeks, 5 weeks, 6 weeks, 8 weeks), and two general times of day (morning peak vs. afternoon off-peak), the group size was calculated at the line-direction level and for all intra-line OD pairs individually, for weekly rolling date periods over a four-month period (Aug.-Dec. 2012).

For each time of day, time interval length, and date period length, the average group size was calculated for the line direction and each intra-line OD pair. These group sizes were then compared against the standards proposed for MTR (30 for OD pairs, 150 for lines/line-directions). The results are summarized in Tables 5.4 and 5.5; Table 5.4 shows, for each interval length, date period length, and time of day combination, the percentage of intra-line OD pairs with group sizes above the OD pair minimum, while Table 5.5 shows the raw line-direction group size for each combination, with the cells in green indicating the combinations above the line minimum.

An important observation that can be made from these results is that the line-level IRBT can be calculated over a much greater range of times than the OD-level. The line-level IRBT can be calculated with acceptable accuracy in almost all scenarios in the peak,
and several in the off-peak. There exist no scenarios, however, where all OD pair IRBTs achieve sufficient frequent traveler group size, and only in 6 out of 32 scenarios are a majority of OD pairs above the threshold. This has the potential to impose serious limits on the application of the IRBT at the OD-level. It is also consistent with the findings from the previous section, that on a per-OD pair basis, the line-level 150 minimum is much easier to achieve than the OD pair-level 30 minimum.

It should be noted that the OD pairs meeting the minimum group size for a given time period are generally those with the highest demand; thus, the percentages in Table 5.4 underestimate the fraction of intra-line passenger trips whose reliability can be estimated using the OD pair IRBT. These percentages are shown for the AM peak in Table 5.6, calculated by summing the total passenger trips (i.e., not exclusively frequent passengers).
over the OD pairs meeting the minimum group size, and dividing the sum by the total intra-line demand. This is potentially beneficial for applications where only a representative selection of OD pair IRBTs are required, rather than all OD pairs; some high-level planning applications could fall into this category.

Table 5.6: Percentage of Passenger Trips Represented by Island Line WB OD pair IRBTs meeting min. group size

<table>
<thead>
<tr>
<th>Time Interval Length</th>
<th>Date Period Length</th>
<th>4 weeks</th>
<th>5 weeks</th>
<th>6 weeks</th>
<th>8 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Peak (7:30-9:30 AM)</td>
<td>15 min</td>
<td>0%</td>
<td>9%</td>
<td>30%</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>30 min</td>
<td>3%</td>
<td>44%</td>
<td>56%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>1 hour</td>
<td>20%</td>
<td>82%</td>
<td>86%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td>2 hours</td>
<td>24%</td>
<td>86%</td>
<td>90%</td>
<td>93%</td>
</tr>
</tbody>
</table>

The table also shows several group size patterns across time periods, that affect the ability to calculate OD pair and line/line-direction IRBTs:

- **Time of Day**: *Meeting group size requirements is much more difficult in the off-peak.* There were no time periods in the off-peak where more than a few OD pairs had sufficient group sizes, and most had zero acceptable OD pairs, whereas 10 out of 16 in the peak had 20% or more. At the line-direction level, 13 time periods met the minimum group size in the AM peak, compared with 5 in the off-peak.

- **Date Period Length**: *Four weeks is generally an insufficient date period length.* At 4 weeks, sufficient group sizes can be attained for no more than 15% of OD pairs, and at the line-direction level only for 2 hour time intervals during the peak. For date periods 5 weeks or longer, however, sufficient group sizes can be attained, during the peak, for the line-direction at all time intervals, and 20% or more of OD pairs for all time intervals of 30 minutes or longer.

- **Time Interval Length**: Time periods of an hour or greater are required to obtain sufficient group sizes for a majority of OD pairs in the peak. However, only 15 minutes is necessary in the peak to obtain a sufficient line-direction level group size.

These specific findings, it should be noted, stem only from a single line-direction on a single system; the limitations on lines in other systems could be quite different. However, some of the general patterns observed—the IRBT being more constrained by frequent traveler
group size during the off-peak, and for OD pairs compared to lines—are likely to hold in
general for high-frequency, high-demand transit services such as the MTR.

5.5 Evaluation of the IRBT

The definition and analysis of the IRBT presented in this chapter provides a sufficient
basis to evaluate this metric, against the design criteria for passenger-experienced reliability
metrics proposed in Section 4.1, as grouped into five general requirements: Representative of
the Passenger Experience, Meaningful for Passengers and Non-Experts, Comparable Across
Services and Times of Day, Time Period Flexibility, and Service Scope Flexibility.

Representative of the Passenger Experience

The design of the IRBT effectively satisfies almost all the criteria for being representative
of the passenger experience:

- Its basis in complete passenger journey times ensures it includes all sources of unreli-
  ability, including delays from denied boarding and congestion.

- By excluding passenger trips below the median from analysis, the metric distinguishes
  between early and late arrivals, the former being of less concern to the passenger, and
  also controls for average travel time variation across time of day.

- The IRBT is always calculated at the base level for OD pairs, relating to actual
  passenger journeys, with aggregate metrics simply being averages of OD pair-level
  results.

- By excluding passenger travel times above the 95th percentile from the analysis, the
  IRBT (as implemented for \( N=95 \)) should exclude “extreme” delays that occur very
  infrequently, and are thus unlikely to affect the “average” passenger’s perception of
  reliability.

- The IRBT, with its basis in individuals’ travel time variability (the IBT), substan-
  tially controls for variation in passenger behavior. Thus, the IRBT almost exclusively
  reflects the attributes of the transit service, rather than exogenous factors—so long
  as there is a sufficient group size to “average out” IBT variation due to behavior. At
  low frequent traveler group sizes, behavior can bias results.
The only significant concern with the IRBT, with respect to the criteria laid out in Section 4.1.1, is that it only captures the reliability experience of frequent travelers, taking more than 20 trips on a given OD pair. This could be problematic if such travelers, as a class, have a significantly different reliability experience than passengers in general, which would conflict with the need for no bias towards particular passenger demographics. More work is required to determine whether or not this effect is significant.

**Meaningful to Passengers and Non-Experts**

The IRBT’s performance is mixed in terms of being meaningful to passengers and non-experts.

A strength of the IRBT is that it is expressed in terms of travel times; specifically, the “extra” travel time a passenger needs to budget for unreliability—the buffer time. Thus, the IRBT in theory relates directly to passengers’ reliability experience, as described in Section 2.2.3. However, the buffer time concept is unlikely to be immediately understood by the public, or even most knowledgeable stakeholders; some explanation will be required to communicate the concept. For management, advocates, and other highly motivated stakeholders, this is a reasonable requirement. Once understood, the IRBT should be more meaningful to them than operational measures such as OTP. Even a 5 minute explanation, on the other hand, may be more than the average passenger is willing to spend learning about a new reliability metric. Thus, effective advertising and public communications will likely be necessary to get public acceptance of a metric like the IRBT.

The IRBT is fairly objective, as it is based on passengers’ actual travel times, and is not dependent on many subjective parameters. The only significant parameter to be set is the upper percentile $N$, which could be set excessively low to “hide” bad performance. However, this should not be an issue if the 95th percentile is used, as recommended here and in the previous literature on the RBT.

The IRBT is of limited use for directly helping passengers plan journeys. Theoretically, one could publish OD pair IRBTs, with the recommendation that passengers add this time to their expected travel time as a buffer time. However, most passengers are unlikely to precisely know their “expected” travel time for a given OD pair, as most do not carefully monitor their travel times. Thus, most passengers will not have a clear “reference” travel time to add the IRBT to, to get a recommended “budgeted” travel time. For journey
planning purposes, it is thus better to provide an absolute range of travel times, based on the overall OD travel time distribution, as proposed by Uniman [2] (see Figure 3-1).

**Comparable Across Services and Times of Day**

The IRBT cannot, in general, be directly compared across different transit services, or between different times of day for the same service, as the IRBT magnitude is greatly dependent on the scheduled headway, as shown in Section 5.3.2. However, this limitation can be overcome by using the Excess IRBT, which, as described in Section 5.1.4, controls for the scheduled headway. The Excess IRBT is effectively an “absolute” measure of reliability, which can be compared across services and times of day.

**Time Period Flexibility**

In theory, the IRBT can be calculated for essentially any time of day interval and set of days, giving it excellent time period flexibility. However, this is limited in practice by the need to obtain a sufficient frequent traveler group size, as detailed in Section 5.4. A minimum of five weeks of input data is required in almost all cases, which may be a longer period than desirable for many reliability monitoring or reporting applications. Furthermore, the ability to calculate the IRBT during the off-peak is limited, effectively restricted to the line-level at 2-hour intervals. On the other hand, during the peak periods—usually the time of day of greatest concern for reliability monitoring—the line-level IRBT can be calculated for time of day intervals as short as 15 minutes, allowing for detailed examination of trends within the peak periods.

**Service Scope Flexibility**

The IRBT, like the RBT, is theoretically calculable for any combination of OD pairs. However, the ability to calculate the IRBT at the OD pair level is significantly limited by minimum frequent traveler group sizes, especially during off-peak times of day, when the IRBT cannot be calculated for almost all OD pairs. This issue is likely to preclude the use of the IRBT for journey planner applications, where information is usually provided at the specific OD pair level, and often for different times of day, which would include the off-peak.
Chapter 6

Platform-to-Platform RBT

This chapter defines and evaluates the second metric proposed in this thesis, the Platform-to-Platform Reliability Buffer Time, or PPRBT. The PPRBT’s goal, like the IRBT’s, is to measure passengers’ reliability experience by estimating the typical passenger’s buffer time, and do so more effectively than existing measures; towards this end, the PPRBT aims to exclude the cross-passenger variation present in the AFC-based RBT, while including the passenger congestion effects and transfer unreliability not captured by the AVL-based RBT.

The PPRBT has an additional objective: to estimate the buffer time without using the travel times of individual users. There are several motivations for this. First, such a metric will not depend on a large sample size of frequent travelers, which imposes significant time period and location restrictions (as seen for the IRBT in Section 5.4). Second, and more importantly, metrics not dependent on individuals’ travel times could potentially be used for transit services without exit fare transactions—a category including essentially all public buses, and a large number of metro systems (e.g., New York City Subway).

This chapter begins, in Section 6.1, with the definition of the PPRBT metric. The conceptual starting point for the metric’s design is the AVL-based RBT, as defined by Ehrlich [5] and Schil [6]. This approach is taken because the AVL-based RBT, by being based on AVL data, effectively removes cross-passenger variation, and is clearly not dependent on observation on individuals’ travel times. The PPRBT’s design improves upon the AVL-based RBT, however, by incorporating passenger demand effects (denied boardings, station congestion, etc.) into the estimated travel time estimation, and being applicable to OD pairs and aggregate service scopes with transfers. This design, however, comes with a limitation:
the inability to measure access/egress time variation.

The remainder of the chapter explores the properties of the PPRBT, as applied to the MTR system. Section 6.2 describes the process used to implement the PPRBT algorithm in the MTR context, including a description of the MTR-specific data sources used as inputs to the model, and the assumptions incorporated into this data. Section 6.3 presents PPRBT results from the MTR system, and compares them with corresponding IRBT results as a means of validating the PPRBT model. In this validation, potential issues with the PPRBT’s ability to accurately capture congestion effects are identified. Finally, Section 6.4 presents an evaluation of the PPRBT as a passenger-experienced reliability measure, both with respect to the guidelines laid out in Section 4.1, and with respect to the IRBT.

6.1 PPRBT Design

This section describes the design of the PPRBT. An overview is given in Section 6.1.1, and the various data sources required are given in Section 6.1.2. The PPRBT algorithm itself is presented in five parts: Section 6.1.3 describes the model assumptions; Section 6.1.4 describes passenger waiting time simulation; Section 6.1.5 describes the algorithm for calculating the PPRBT for single-stage journeys; Section 6.1.6 extends the algorithm to journeys requiring transfers; and Section 6.1.7 describes the process of aggregating OD-level PPRBTs to the line (or other spatially aggregate) level.

6.1.1 Design Overview

The PPRBT, like the RBT, is based on an estimate of the overall travel time distribution—for a given OD pair, time-of-day interval, and date period—with the buffer time estimated as the difference between the 95th and 50th percentiles. The heart of the PPRBT methodology is the process for estimating this travel time distribution, in the absence of direct observation of passengers’ travel times.

The travel time distribution is estimated by simulating the journeys of a large number of “virtual passengers”. Each simulated passenger arrives at a specific time at the origin platform (or bus stop), and boards the first available train (or bus) headed towards their destination; for transfer trips, this is repeated for each leg of the journey. Vehicle arrival and departure times from stations/stops are those of the actual trips operated during the
time period, as recorded by the AVL system. Thus, the virtual passengers are subject to all of the vehicle delays, bunching, and capacity restrictions experienced by actual passengers.

The difference between the passenger’s arrival time at the origin platform (or bus stop) and their alighting time at their destination platform then becomes their “platform-to-platform” travel time. Aggregating the platform-to-platform travel times of all virtual passenger produces a platform-to-platform time distribution, which forms the travel time distribution for calculating the PPRBT—hence the full name of the measure, Platform-to-Platform Reliability Buffer Time. The simulation thus excludes access and egress times. This approach is taken due to the significant cross-passenger variation in access and egress times: excluding these elements removes unwanted cross-passenger variation from the (simulated) travel time distribution.

A major component of the travel simulation, and probably the most important contribution of the PPRBT, is a method of estimating the delays that occur when a passenger is unable to board the first train (bus) to arrive, due to insufficient capacity—i.e., denied boardings. To model denied boardings, the simulation estimates, for each arriving train (bus), the available passenger capacity, and the number of passengers waiting to board. Taking the difference yields the number of denied boardings, who must wait for the following train (bus). This process is modeled as a first-in, first-out (FIFO) queueing process. The denied boarding simulation adds significant complexity, as it must account for, in addition to passengers traveling on the target OD pair, those at the origin and those boarding at prior stations traveling to all downstream destinations.

A fundamental assumption of the algorithm is that, for the target OD pair, all passengers take the same route. This assumption is adopted for two reasons: to simplify the simulation, and to remove route choice as a source of cross-passenger variation. The set of train (bus) line-directions used for this route are termed the “target” services.

6.1.2 Data Inputs

The following set of data inputs is used to calculate the PPRBT:

**Origin Departure Times, In-Vehicle Times:** The operated transit service is input as two data sets: vehicle departure times from the origin, and vehicle arrival times at the
destination. Each vehicle’s origin to destination running time\(^1\) can be derived from this data. For trips requiring transfers, headways and in-vehicle times are calculated for each stage of the journey, given the transfer locations and different services used. For example, if a journey from A to C involved taking Line 1 from A to B, and Line 2 from B to C, headways and running times would be obtained for Line 1 from A to B, and Line 2 from B to C. This data is obtainable directly from AVL data.

**OD Passenger Demand:** The total number of passengers traveling the target OD pair is needed at 15-30 minute intervals, grouped by passengers’ arrival time at the origin (e.g., all passengers arriving on Sept. 1 from 8-8:15 AM). This data defines the arrival rate of “virtual passengers” throughout the specified time period. These small intervals are needed to capture rapid changes in demand, which can occur during the peak hours.

The method used to calculate this OD passenger demand depends on the automatic data sources available. If linked entry and exit AFC transactions are available, the demand can be observed directly. If only entry transactions are available, it may be possible to estimate the OD demand using origin-destination inference (ODX) algorithms, such as those developed by Gordon [29]. If such methods cannot be applied, then “static” OD matrices can be used, which estimate the OD flows at intervals for a “typical” day. Such static OD matrices, which are less accurate, are typically derived from manual passenger surveys.

**Vehicle Capacity:** The vehicle capacity, denoted \(C\) in this chapter, is the estimated maximum number of passengers that can board a particular vehicle before boardings are denied. For rail services, actual vehicle capacity depends on not only the vehicle characteristics (size, seating configuration, number of cars, etc.), but also the evenness of loading among cars in a single train; if train loading is very uneven, denied boardings can occur at some cars, while others have remaining capacity, reducing the effective vehicle capacity. Unlike vehicle characteristics, the evenness of loading can vary greatly from station to station, even on a single line; thus, \(C\) is best specified for rail transit at the line and station level, rather than just the line-level.

Ideally, \(C\) values should be derived from station- and line-specific observations of denied boardings. If this is infeasible, an alternative is the “official” train (bus) capacity used

\(^1\)i.e., the time between when the train or bus departs the origin and arrives at the destination.
for service planning by most operators. However, such official capacities are not station-specific, and not usually based on actual observation of passenger loads, and thus may not represent the true capacity. Thus, if the official capacity is used, it should be considered not a “true” value, but only as an initial value of a parameter to be calibrated, possibly on a station-by-station basis. If multiple vehicle types are used on a single service, $C$ may need to be further specified at the trip-level.

**Incoming Vehicle Load, Alightings:** Estimating the available passenger capacity for each train (bus) at the origin (or transfer point) requires two passenger demand inputs: (1) the estimated incoming passenger load, denoted $L$, and (2) the estimated number of alightings, denoted $E$. Given the estimated vehicle capacity, the remaining passenger capacity of any given vehicle $R$ is, trivially, $R = C - (L - E)$.

If Automatic Passenger Count (APC) data is available, incoming passenger load and alightings can be measured directly for each trip. If APC is unavailable—the more common scenario—they can be estimated using static, ODX-based, or full AFC-based OD matrices, and OD route choice data. The accuracy of the derived values will then depend on the accuracy of the input OD matrices and route choice information. Section 6.2 describes the process used to estimate the vehicle load and alightings for the MTR system using AFC data.

**Station Demand, Transfer Demand:** The total passenger demand for the target service(s) is required at the origin, and each transfer point (if applicable). This is given by two data sets: station demand, the demand from passengers entering at the origin/transfer location, and transfer demand, the demand from passengers transferring from other services at the location. Both are specific to the service direction serving the OD pair.

These concepts can be illustrated by an MTR example: the Admiralty to Fortress Hill OD pair, in Figure 6-1. This OD pair is served by the Island Line (blue) eastbound, and the origin is a transfer station with the Tsuen Wan Line (red). The station demand, then, is the total passengers entering the system at Admiralty and boarding the Island Line eastbound, and the transfer demand is the total passengers transferring from the Tsuen Wan Line to Island Line eastbound at Admiralty.

Like the OD passenger demand, station and transfer demand is needed at 15-30 minute
intervals throughout the time period, to capture the dynamics in demand. This data can be derived from OD matrices and route choice information; Section 6.2 describes this process for MTR. In some situations it may be possible to measure station inflow directly from AFC entry transactions—at terminal stations, or on many proof of payment systems, for example.

**Source Service Arrival Times:** To simulate the “clustered” platform/stop arrival times of passengers transferring to the target service at the origin or transfer locations, which correspond to the discrete vehicle arrival times of non-target services that contribute to transfer demand (e.g., the Tsuen Wan Line in Figure 6-1), these “source service” vehicle arrival times are needed at the origin and all transfer locations. This can be obtained from AVL data.

**OD Route Choice:** Route choice information is needed for every OD pair in the network served by more than one feasible route. This information should contain, for every OD pair, a set of feasible route choices, and the estimated percentage of passengers choosing that route. This information is used to: (1) set the “primary” route for calculating the PPRBT for transfer OD pairs, and (2) calculate the incoming vehicle load, alightings, station entries, and transfer demand values. Route choice information can be estimated through passenger surveys (intercept, mail-in, or online), or by route choice models derived from such surveys.
Algorithm Notation

The data inputs are denoted in the rest of this chapter using the following notation, for a target OD pair. Demand data provided at 15-30 minute intervals is specified using interval number \( n \), which counts intervals sequentially from the start of the time of day interval on the first date period day. For example, if the time-of-day interval is 8-9 AM over a 30 day date period, and demand is given at 30 minute intervals, these intervals would be \{8-8:30 Day 1, 8:30-9 Day 1, 8-8:30 Day 2, 8:30-9 Day 2, 8-8:30 Day 3, 8:30-9 Day 3, \ldots \}, and denoted by \( n = \{1, 2, 3, 4, 5, 6, \ldots \} \). For transfer OD pairs, vehicle trip data is specified separately for each service used, and station-specific data specified at the origin and each transfer point.

\[
\begin{align*}
    t & \quad \text{time elapsed from start of period} \\
    C & \quad \text{effective vehicle capacity at origin (or transfer location)} \\
    t_n & \quad \text{start time of interval } n \\
    I_n & \quad \text{origin or transfer station inflow during interval } n \\
    f_{OD,n} & \quad \text{OD passenger demand during interval } n \\
    z & \quad \text{train (or bus) trip number, for target service, counting from start of time period} \\
    t_d(z) & \quad \text{departure time of trip } z \text{ from origin (or transfer station)} \\
    V_{OD}(z) & \quad \text{running time of trip } z \text{ from origin station to destination station, or trip leg} \\
    L(z) & \quad \text{incoming vehicle passenger load for vehicle trip } z \text{, at origin (or transfer location)} \\
    E(z) & \quad \text{alightings for vehicle trip } z \text{, at origin (or transfer location)} \\
    k & \quad \text{“source service” identifier} \\
    X_{n,k} & \quad \text{transfer inflow during interval } n \text{, for route } k \text{, at origin (or transfer location)} \\
    y_k & \quad \text{train (or bus) trip number, for source service } k \\
    t_{a,k}(y_k) & \quad \text{arrival time of source service vehicle trip } y_k \text{, at origin (or transfer location)}
\end{align*}
\]

6.1.3 Model Assumptions

The PPRBT algorithm depends on a series of assumptions about arrival and boarding behavior, to simplify the problem and improve its tractability:

**Passenger Arrival Behavior:**

1. Passengers enter origin/transfer stations at a constant rate during each 15-30 minute interval \( n \).
2. Passengers transferring from source services arrive in groups, according to the vehicle arrival times $t_{a,k}(y_k)$ for that route. The number of passengers arriving in each group is assumed to be proportional to the preceding headway for each arrival $y_k$.

3. Group arrivals are not simultaneous, but spread over a short interval $s$ (about 30-60 seconds), a function of the vehicle layout and transfer facility characteristics.

**Explanation:** Assumption (1) is reasonable for most high-frequency services, where passengers should arrive randomly, and demand should not change dramatically within a short (15-30 minute) period. Thus, this simulation model is not valid for long-headway services, where many passengers time their arrivals based on the schedule. Assumption (2) is reasonable in the absence of information about the passenger loading on vehicle trip $y_k$.

Assumption (3) accounts for the fact that all passengers cannot exit a vehicle simultaneously, and will have different walk times from the arrival platform to the departure platform, introducing some level of variation in their arrival times at the following platform.

**Boarding Behavior:**

4. Waiting passengers form a queue, in order of arrival at the platform (or bus stop).

5. Trains are treated as a single vehicle, with a single capacity $C$, and a single queue (rather than a queue for every door).

6. If, for a given vehicle departure, the number of waiting passengers $P$ exceeds the remaining capacity $R$, then the first $R$ passengers in the queue will board the vehicle, and the remaining $P - R$ will be denied boarding, and wait for the next departure.

**Explanation:** Assumptions (4), (5), and (6) describe a first-in, first-out (FIFO) queueing process for the entire bus or train. This assumption is clearly unrealistic; typically, waiting passengers do not form queues, and board in a somewhat random order. Furthermore, even if passengers queue at every door (as is the case on MTR), the division of waiting passengers into queues at each train door is unlikely to result in “first in, first out” order over the entire train platform. However, for most transit systems, it can be assumed passengers arriving later in a headway are more likely to be denied boarding. Thus, the “single car, single queue” model is a reasonable starting assumption for most transit systems, even if FIFO queueing does not take place explicitly.
6.1.4 Waiting Time Simulation (All Destinations)

For a given OD pair, the first PPRBT calculation step is to estimate the wait time distribution for all passengers boarding the target service at the origin, headed to all destinations. This is done by deriving passenger platform/stop arrival and departure time estimates, denoted \( a(p) \) and \( d(p) \), as a function of the order \( p \) in which the passenger arrived from the start of the simulation period (i.e., the \( p \)th passenger to arrive).

**Arrival Time Function Definition**

From the discrete station demand \( I_n \), and assumption (1), a continuous station demand rate function \( \lambda_I(t) \) can be calculated as,

\[
\lambda_I(t) = \frac{I_n}{t_{n+1} - t_n}, \quad t_n \leq t < t_{n+1} \tag{6.1}
\]

Similarly, a continuous transfer demand rate function \( \chi_k(t) \) can be defined for each source service \( k \) as follows,

\[
\chi_k(t) = \frac{X_{n,k}}{t_{n+1} - t_n}, \quad t_n \leq t < t_{n+1} \tag{6.2}
\]

To transform continuous transfer arrivals into grouped transfer arrivals (in accordance with assumptions (2) and (3)), the group size function \( \gamma_k(y_k) \) is defined, representing the number of passengers alighting from source service arrival \( y_k \), and transferring to the target route. It is found by integrating \( \chi_k \) over the preceding headway for arrival \( y_k \), as follows:

\[
\gamma_k(y_k) = \int_{t_{a,k}(y_k-1)}^{t_{a,k}(y_k)} \chi_k(t) \, dt \tag{6.3}
\]

The grouped arrival rate function is then defined as,

\[
\lambda_k(t) = \begin{cases} 
\frac{\gamma_k(y_k)}{s} & t_{a,k} \leq t < t_{a,k} + s, \quad \forall y_k \\
0 & \text{otherwise}
\end{cases} \tag{6.4}
\]

indicating that \( \gamma_k(y_k) \) transfer passengers arrive after each source service arrival \( y_k \), over a short interval of length \( s \). This calculation is illustrated in Figure 6-2.

Given the station arrival rate \( \lambda_I(t) \), and transfer arrival rates \( \lambda_k(t) \) for all source services...
Figure 6-2: Transfer Arrival Rate $\lambda_k(t)$

$k$, the total passenger arrival rate $\lambda(t)$ can be expressed as follows,

$$\lambda(t) = \lambda_I(t) + \sum_{k} \lambda_k(t) \quad (6.5)$$

The cumulative passenger arrivals as of time $t$, denoted $A(t)$, can be then be expressed as

$$A(t) = \int_{0}^{t} \lambda(t) \, dt = \int_{0}^{t} \lambda_I(t) + \sum_{k} \lambda_k(t) \, dt \quad (6.6)$$

The arrival time function $a(p)$ can then be defined as the inverse of the cumulative passenger arrival function, as follows,

$$a(p) = A^{-1}(p) \quad (6.7)$$

Equation 6.7 relies on the fact that $p$ represents different attributes when applied to the individual, and the whole system. For the individual, $p$ represents the order in which the individual arrived. For the system, $p$ represents the total (i.e. cumulative) number of boarding passengers to have arrived, as of time $t$. This relation works because $A(t)$ and $a(p)$ are calculated in terms of continuous passenger arrivals, rather than discrete arrivals—i.e., can be evaluated for “fractional passenger” values such as $p = 123.6$. This simplification implies that, for any time $t$, there is a unique value of $A(t)$. Thus, $A(t)$ is a continuous
function, which can be inverted to obtain $a(p)$.

**Departure Time Function Derivation**

For a given train (bus) $z$ leaving the origin station, the number of passenger boardings (headed to any destination), denoted $b(z)$, is given by

$$b(z) = \min\{R(z), P(z)\}$$  \hspace{1cm} (6.8)

where $R(z)$ is the remaining vehicle capacity, and $P(z)$ is the number of passengers waiting (in a queue, per assumption (4)) to board the target transit service. This statement succinctly defines the boarding scenarios. If there is enough remaining capacity ($R \geq P$), all waiting passengers will board the vehicle; otherwise ($R < P$), passengers will board until the vehicle is filled to capacity $C$, per assumption (6).

$R$ can be expressed in terms of input data, as $R(z) = C - L(z) + E(z)$: capacity minus incoming passenger load, plus the extra space made available by alighting passengers. $P(z)$ can be expressed as the difference between the cumulative arrivals prior to trip $z$’s departure, $A(t_d(z))$, and the total passenger boardings prior to $z$, $\sum_{i=1}^{z-1} b(i)$, as given,

$$P(z) = A(t_d(z)) - \sum_{i=1}^{z-1} b(i)$$  \hspace{1cm} (6.9)

Given this information, $b(z)$ can then be fully defined as a recursive function, in terms of input data and the previously calculated function $A(t)$, as follows:

$$b(z) = \min \begin{cases} A(t_d(z)) - \sum_{i=1}^{z-1} b(i) \\ C - L(z) + E(z) \end{cases}$$  \hspace{1cm} (6.10)

Equation 6.10 is the heart of the simulation method, determining which passengers are denied boarding, and forces the simulation to be run in a sequential, trip-by-trip process, due to its recursive nature. Given $b(z)$, a cumulative departures function $D(z)$ can be defined, for trips 1 through $z$, as

$$D(z) = \sum_{i=1}^{z} b(i)$$  \hspace{1cm} (6.11)
Given the FIFO boarding process assumptions ((4) and (6)), for a given passenger with arrival order \( p \) the vehicle trip they departed from, denoted \( z^*(p) \), is defined by

\[
z^*(p) = \min (z \mid p \leq D(z)) \tag{6.12}
\]

The minimum indicates that passengers board the first available vehicle, while the condition of \( p \leq D(z) \) implies that, in accordance with assumption (6), passengers do not “cut in line”, or board vehicles that arrive before the passenger arrives. Given \( z^*(p) \), \( d(p) \) can then be defined in terms of the departure time of trip \( z^*(p) \), as follows

\[
d(p) = t_d(z^*(p)) \tag{6.13}
\]

**Waiting Time Estimate**

Once \( a(p) \) and \( d(p) \) have been derived, the estimated waiting time, including denied boarding delay, is simply the difference between the departure and arrival times, as given

\[
w(p) = d(p) - a(p) \tag{6.14}
\]

**Simulation Time Period**

The simulation process above should, at a minimum, be run for the PPRBT time-of-day interval \( T \), for all days in the date period. However, if the probability of denied boardings occurring immediately before the beginning of \( T \) is significant, it is preferable to begin the simulation earlier, so queue lengths and denied boardings are not underestimated at the beginning of the time-of-day interval.

**6.1.5 PPRBT Calculation: Single Leg**

After the wait time function \( w(p) \) is calculated, several additional steps are required to derive the PPRBT for single-stage journeys (i.e., no transfers), described below:
1. Derive OD Arrival Function

Given discrete OD passenger demand $f_{od,n}$, the cumulative OD arrival function $A_{od}(t)$ can be derived similarly to $A(t)$. First, an OD arrival rate function can be calculated as,

$$\lambda_{od}(t) = \frac{f_{od,n}}{t_{n+1} - t_n}, \quad t_n \leq t < t_{n+1} \quad (6.15)$$

Integrating $\lambda_{od}(t)$ then produces the cumulative OD arrival function,

$$A_{od}(t) = \int_0^t \lambda_{od}(t) \, dt \quad (6.16)$$

Inverting $A_{od}(t)$ then yields, as in equation 6.7, the OD pair passenger arrival time function:

$$a_{od}(p_{od}) = A_{od}^{-1}(p_{od}) \quad (6.17)$$

2. Derive OD Waiting Time Function

To derive the waiting time distribution for passengers traveling a particular OD pair, $w(p)$ is transformed from a function of $p$ to a function of $p_{od}$, the order in which the OD pair passengers arrive at the origin—i.e., the OD equivalent of $p$. First, a function for $p$ as a function of $p_{od}$ is calculated as follows,

$$p = A(a_{od}(p_{od})) \quad (6.18)$$

Let $z_{od}^*(p_{od})$ be the vehicle trip passenger $p_{od}$ departed on; $z_{od}^*(p_{od})$ can be defined by substituting $A(a_{od}(p_{od}))$ into the equation for $z^*(p)$, 6.12, resulting in

$$z_{od}^*(p_{od}) = \min(z \mid A(a_{od}(p_{od})) \leq D(z)) \quad (6.19)$$

Following Equation 6.13, the OD departure time $d_{od}(p_{od})$ can be given by

$$d_{od}(p_{od}) = t_d(z_{od}^*(p_{od})) \quad (6.20)$$

From $a_{od}(p_{od})$ and $d_{od}(p_{od})$, the OD waiting time function can then be defined as:

$$w_{od}(p_{od}) = d_{od}(p_{od}) - a_{od}(p_{od}) \quad (6.21)$$
3. Derive Platform-to-Platform Time Function

For a given passenger \( p \), their platform-to-platform travel time (for a single-stage journey) is equal to the sum of their waiting time \( w_{od}(p_{od}) \), and their “in-vehicle time”, denoted \( v_{od}(p_{od}) \). This in-vehicle time is equivalent to the origin to destination running time \( V(z) \) of the particular vehicle trip \( z_{od}^{*}(p_{od}) \) they board. From this result, \( v_{od}(p_{od}) \) can be defined as

\[
v_{od}(p_{od}) = V(z_{od}^{*})
\]  
(6.22)

The platform-to-platform time function \( TT_{od}(p_{od}) \) can then be defined as

\[
TT_{od}(p_{od}) = w_{od}(p_{od}) + v_{od}(p_{od})
\]  
(6.23)

This can then be rewritten in terms of \( z_{od}^{*} \) using Equations 6.13 and 6.22 as

\[
TT_{od}(p_{od}) = t_{d}(z_{od}^{*}) + V(z_{od}^{*}) - a_{od}(p_{od})
\]  
(6.24)

4. CDF and PPRBT Calculation

An empirical platform-to-platform travel time CDF can be derived by calculating \( TT_{od}(p_{od}) \) for a large, evenly spaced set of \( p_{od} \) values such that \( a_{od} \) falls within the target time-of-day interval and date period. This set, \( p \), represents all passenger trips taken on the OD pair, during the time period of interest. Given this information, the empirical platform-to-platform travel time CDF and the PPRBT can be defined as

\[
F_{od}(x) = P\left( TT_{od}(p) \leq x \right)
\]  
(6.25)

\[
PPRBT_{od} = F_{od}^{-1}(0.95) - F_{od}^{-1}(0.5)
\]  
(6.26)

If there are relatively few travelers for the target OD pair, \( TT_{od}(p_{od}) \) can be sampled at fractional intervals (e.g. \( p = 0.5, 1, 1.5, 2, \ldots \)), to obtain a larger sample.

The process for calculating the single-leg PPRBT can be summarized by the pseudocode Algorithm 6.1, taking \( t_{d}(z), C, L(z), E(z), A(p), A_{od}(p_{od}) \), and \( z_{max} \) as inputs, with \( z_{max} \) being the last vehicle trip during the time period.
Algorithm 6.1 Algorithm for single-leg PPRBT calculation

\[
D(0) \leftarrow 0 \\
\text{for } z \leftarrow 1 \text{ to } z_{\text{max}} \text{ do} \\
\quad P(z) \leftarrow A(t_d(z)) - D(z - 1) \\
\quad R(z) \leftarrow C - L(z) + E(z) \\
\quad b(z) \leftarrow \min(P(z), R(z)) \\
\quad D(z) \leftarrow D(z - 1) + b(z) \\
\text{end for} \\
\text{for all } p_{\text{od}} \in p \text{ do} \\
\quad a_{\text{od}} \leftarrow A^{-1}_{\text{od}}(p_{\text{od}}) \\
\quad z^*_{\text{od}} \leftarrow \min(z \mid A(a_{\text{od}}(p_{\text{od}})) \leq D(z)) \\
\quad TT_{\text{od}}(p_{\text{od}}) \leftarrow t_d(z^*_{\text{od}}) + V(z^*_{\text{od}}) - a_{\text{od}}(p_{\text{od}}) \\
\text{end for} \\
TT_{50} = \text{percentile}(TT_{\text{od}}, 50) \\
TT_{95} = \text{percentile}(TT_{\text{od}}, 95) \\
\text{PPRBT}_{\text{od}} = TT_{95} - TT_{50} \\
\]

6.1.6 Extension to Transfer OD pairs

To calculate the PPRBT for transfer OD pairs, travel time distributions are calculated for each leg (i.e., trip aboard a single vehicle) of the OD route independently, and then combined via a simplified transfer model.

OD Path Simulation

The first step is to define the OD path, the sequence of transit services and transfer points used to complete the journey. For OD pairs with multiple feasible paths the algorithm selects the OD path with the highest fraction of OD trips (i.e. the most popular path), based on the input route choice data. For a given OD path, each journey leg is denoted by \( l \in \{1, 2, \ldots, N\} \), where \( N \) is the total number of legs in the OD path.

With the OD path defined, the waiting time simulation is run for each leg \( l \)'s beginning station/stop (i.e., the origin and every transfer location). Let the resulting \( A(p), D(z), \) and \( t_d(z) \) for each leg \( l \) be denoted \( A_l(p), D_l(z_l), \) and \( t_{d,l}(z_l) \), respectively, where \( z_l \) indicates the vehicle trip number for the leg \( l \) target service.

Transfer Model

The following describes the passenger transfer model used. Let \( e_{\text{od},l}(p_{\text{od}}) \) be the arrival time of the vehicle carrying passenger \( p_{\text{od}} \) for leg \( l \), at the leg’s end station. If the passenger
then transfers, the end location of leg \( l \) is also the departure location for the next leg \( l+1 \). The arrival time at the waiting location for leg \( l+1 \) can then be denoted \( a_{od,l}(p_{od}) \)—i.e., \( a_{od}(p_{od}) \) for leg \( l \). From this, the relation between \( a_{od,l}(p_{od}) \) and \( e_l(p_{od}) \) is defined as

\[
a_{od,l+1}(p_{od}) = e_{od,l}(p_{od}) + \tau_l(p_{od})
\]  

(6.27)

where \( \tau_l(p_{od}) \) is the transfer time between the leg \( l \) alighting location and waiting location (e.g., platform) for leg \( l+1 \). \( \tau_l(p_{od}) \) is defined as a uniform random variable over the interval \([\tau_{l,min}, \tau_{l,max}]\), modeling the transfer time variability due to alighting time variation (one’s alight time varies depending on one’s location in a car), and congestion in stations (which affects walking time). \( \tau_{l,min} \) and \( \tau_{l,max} \) are set as the minimum and maximum feasible times to make the particular transfer \( l \) at a “normal” walking speed, respectively. Walking speed variation is intentionally not modeled in \( \tau_l(p_{od}) \), because that transfer time variability is due to passenger behavior, not service reliability.

**Platform-to-Platform Time Distribution**

For passenger \( p_{od} \) and leg \( l \), the alighting time \( e_{od,l}(p_{od}) \) is equal to the sum of their leg departure time, \( d_{od,l}(p_{od}) \), and in-vehicle time, \( v_{od,l}(p_{od}) \), as follows

\[
e_{od,l}(p_{od}) = d_{od,l}(p_{od}) + v_{od,l}(p_{od})
\]  

(6.28)

Following Equations 6.20 and 6.22, \( d_{od,l}(p_{od}) \) and \( v_{od,l}(p_{od}) \) can be defined as functions of \( z_{od,l}^*(p_{od}) \), the particular vehicle trip \( z_l \) the passenger took for leg \( l \): \( d_{od,l}(p_{od}) = t_{d,l}(z_{od,l}^*) \), and \( v_{od,l}(p_{od}) = V_l(z_{od,l}^*) \). Thus, \( e_{od,l}(p_{od}) \) can be rewritten as a function of \( z_{od,l}^*(p_{od}) \) as

\[
e_{od,l}(p_{od}) = t_{d,l}(z_{od,l}^*) + V_l(z_{od,l}^*)
\]  

(6.29)

The vehicle trip for each leg \( z_{od,l}^*(p_{od}) \) is defined in the same manner as the vehicle trip for single-stage journeys \( z_{od}^*(p_{od}) \), following Equation 6.19, as

\[
z_{od,l}^*(p_{od}) = \min (z_l \mid A_l(a_{od,l}(p_{od})) \leq D_l(z_l))
\]  

(6.30)

where \( A_l(p) \) and \( D_l(z_l) \) are outputs for the leg \( l \) waiting simulation. Then, from Equation

114
6.27, \( z^*_{od, l}(p_{od}) \) can be redefined in terms of the alighting time for the previous leg \( l - 1 \) as
\[
 z^*_{od, l}(p_{od}) = \min (z_l | A_l(e_{od,l-1}(p_{od}) + \tau_l(p_{od})) \leq D_l(z_l)) \tag{6.31} 
\]

Combined, Equations 6.29 and 6.31 fully define \( e_{od,l}(p_{od}) \) as a recursive function for \( l \geq 2 \), in terms of the simulation outputs \( A_l(p) \) and \( D_l(z_l) \), vehicle departure times \( t_{d,l}(z_l) \), vehicle running times \( V_l(z^*_{od,l}) \), and random variable \( \tau_l(p_{od}) \). For \( l = 1 \), \( e_{od,1}(p_{od}) \) can be defined as in 6.30 for \( a_{od,1}(p_{od}) \), with \( a_{od,1}(p_{od}) \) calculated as in 6.17 based on the origin arrival rate.

With \( e_{od,l}(p_{od}) \) defined, the platform-to-platform travel time function can then be defined as the difference between the passenger’s alighting time at their destination, at the end of the final leg \( N \), and their arrival time at the origin, as given
\[
 TT_{od}(p_{od}) = e_{od,N}(p_{od}) - a_{od,1}(p_{od}) \tag{6.32} 
\]

At this point, the PPRBT can then finally be calculated in the same manner as for single-stage journeys, according to Equations 6.25 and 6.26. The process for calculating the PPRBT for transfers can be summarized by the pseudocode Algorithm 6.2, taking \( t_{d,1}(z_l) \), \( C_l, L_l(z_l), E_l(z_l), A_l(p), A_{od,l}(p_{od}), \) and \( z_{max}(l) \) as inputs, with \( z_{max}(l) \) being the last vehicle trip for leg \( l \) during the time period.

### 6.1.7 Spatially Aggregated PPRBT

A spatially aggregated PPRBT can be calculated, like the RBT and IRBT, by taking the weighted average of the OD PPRBTs with respect to the estimated total OD pair demand \( f_{od} \). \( f_{od} \) can be derived from the input OD matrices. For example, a line-level PPRBT can be derived similarly to the line-level IRBT (Equation 5.4) as follows
\[
 PPRBT_{Line} = \frac{ \sum_{OD \in Line} f_{od} \cdot PPRBT_{od} }{ \sum_{OD \in Line} f_{od} } \tag{6.33} 
\]
6.2 PPRBT Application to MTR

This section describes how the generic PPRBT algorithm described in Section 6.1 was adapted to the specifics of the MTR system. There are two parts to this setup: the definition of the input data, and the implementation of the PPRBT algorithm in MATLAB and SQL.

The PPRBT data inputs, specified in Section 6.1.2, were defined using the available MTR data sources as follows:

**Origin Departure Times, In-Vehicle Times, Source Service Arrival Times:** Vehicle trip data for both target and source transit services comes from MTR’s AVL system. This data is generally accurate and complete.

**Vehicle Capacity:** The vehicle capacity $C$ is based on MTR’s 2012 planning capacity of 250 per subway car, or 2000 for a typical 8-car train. These values are derived from MTR surveys of passengers’ “acceptable” levels of crowding, rather than actual denied boarding observations, and are thus likely to underestimate the “true” capacity of MTR’s trains; most passengers are likely to tolerate marginally “unacceptable” levels of crowding in order...
to avoid being denied boarding. To mitigate this potential error, the PPRBT, for all OD pairs, is calculated for three $C$ values: 2000, 2250, and 2500 per train. Because $C=2000$ is likely to be an underestimate, the “alternative” values of $C$ used are higher than MTR’s official value. $C$ is specified at the network-level, due to the absence of any station-level data on denied boardings, or train load distribution. Thus, this PPRBT implementation is unable to account for variation in effective capacity across stations, due to factors such as uneven train loading.

**OD Route Choice:** MTR’s planning route choice estimates, derived from survey data, were used as the route choice input. There are several concerns about the estimates’ accuracy. First, they differ substantially from the results of a more recent MTR route choice survey [30]. Second, they assume route choice splits are constant over the day, which is not necessarily true if certain lines or stations become very congested in the peak periods.

**OD Passenger Demand:** OD passenger demand comes directly from AFC data, grouped at 15-minute intervals by entry time. MTR’s AFC data records both entry and exit transaction times down to the second, so this data is assumed to be accurate.

**Train Load, Transfer Demand:** The train load and transfer demand are based on outputs of MTR’s “passenger assignment” algorithm. This algorithm takes OD matrices from AFC data, the route choice data, average station-to-station running times, and estimated transfer times as inputs, and then estimates, for every passenger trip, the time the passenger traversed the intermediate links and transfers between their origin and destination. The algorithm then counts the passengers traveling over every line link and making each transfer movement during each 15-minute interval, yielding 15-minute link flow and transfer demand.

From the 15-minute output link flow, denoted $F_n$, the passenger load $L$ for a given train $z$ is found by dividing $F_n$ proportionally by the headway $h(z)$, as follows: $L(z) = \frac{1}{15} h(z) F_n$.\(^2\) If $L(z) > C$, the excess passengers $L - C$ are “rolled over” to the next arriving train. This simple model attempts to capture the correlation between headways and passenger load, for successive trains on a given route.

\(^2\)If the headway falls in two 15-minute periods, the average link flow is used.
The resulting train load and transfer demand are subject to any errors present in the route choice data, as well as the error introduced by using average running and transfer times, rather than actual times (affected by delays, denied boardings, etc.). Furthermore, the model \( L(z) = \frac{1}{15} h(z) F \), while perhaps an acceptable first-order representation, is clearly not strictly accurate, as headways can vary along a given route or line (e.g., “bunching”).

**Station Demand, Alightings:** The station demand and alightings are derived from AFC-based OD matrices and the route choice data. The OD matrices provide the number of passengers entering the origin (or transfer point) station, every 15 minutes. The route choice data is then applied to find the fraction of these entering passengers who boarded the target OD line-direction (as opposed to other lines, or the same line in the opposite direction); this yields the 15-minute station demand \( I_n \). Similarly, exit-based OD matrices provide the total passengers exiting the origin (or transfer point) station, every 15 minutes. The route choice data is then applied to find the fraction of these exiting passengers who alighted from the target OD line-direction, yielding 15-minute total alightings \( A_n \). Alightings for each train trip are then assumed to be proportional to the preceding headway of the target service, as with the train load, producing \( E(z) = \frac{1}{15} h(z) A_n \). The resulting station demand and alightings are subject to the route choice data errors. Further error is introduced into the alightings by the assumption that \( E(z) \) is proportional to the headway.

The PPRBT calculation is implemented using two software packages: SQL and MATLAB. SQL is used to store the input data, and generate the station demand and alightings from the route choice and raw AFC data. MATLAB is used for the algorithm calculations, and is chosen for its ability to connect easily to SQL databases. If an open-source implementation is desired, the code should be easily portable to Python. Calculating the line-level PPRBT at hourly intervals for a six-week period takes about a minute on a standard PC.

This input data is probably the best that can be practically expected for a large metro system such as MTR. MTR’s AVL data provides accurate headways and in-vehicle travel times. Its AFC data, combined with detailed route choice estimates, is able to provide load and passenger demand estimates based on actual OD flows during the time periods of interest, rather than relying on static OD matrices. The route choice data, while likely not entirely accurate, is probably more comprehensive than that available to most large
transit operators. Likewise, better train load and alightings estimates could theoretically be obtained using APC systems—but such systems are not in common use on metro trains.

The “weak link” in the MTR input is the train capacity parameter $C$, which is not based on actual observation of denied boardings—as is typical for metro operators. Even relatively small errors in the assumed capacity can cause significant under- or over-estimation of denied boarding effects on the PPRBT. This issue could be addressed by deriving empirical, station-level capacity estimates based on actual counts of train load and denied boardings. Such an effort, though, would likely be very labor-intensive, and thus costly.

6.3 PPRBT Results

This section presents PPRBT results for the MTR system, and assesses these results from two different perspectives. First, does the PPRBT respond as expected to the passenger demand and train capacity, given the PPRBT simulation model’s assumptions? Are the results consistent across locations and PPRBT types? Towards this end, both the OD and line-level PPRBT are evaluated for three values of the capacity parameter $C$: 2000, 2250, and 2500 per train. This approach allows assessment of the PPRBT’s sensitivity to $C$, which for MTR is not known accurately a priori. The PPRBT results are also compared with trends in line passenger flow; as flow nears capacity, denied boardings are expected to occur, which should be reflected in higher PPRBT values.

Second, the PPRBT is assessed in terms of validity: How well do the PPRBT results agree with passengers’ actual experiences? To assess this, the PPRBT results are compared with the IRBT, which (for sufficient passenger group sizes) should be representative of actual passengers’ reliability experiences, and be of similar magnitude to the PPRBT.

This following analysis is presented in three parts. Line-level results are presented first in Section 6.3.1, to illustrate general PPRBT behavior with respect to time of day and demand. Section 6.3.2 explores results for various OD pairs, including transfer OD pairs, expanding the analysis to examine trends with respect to location and other OD characteristics. Finally, general conclusions about the PPRBT’s performance are made in 6.3.3.
6.3.1 Line-Level Results

Figures 6-3 and 6-4 present PPRBT and IRBT results for two line-directions, Island Line westbound and eastbound; Figure 6-3 shows the PPRBT and IRBT at hourly intervals for a fixed date period (September-October 2012), while Figure 6-4 shows trends in the 8-9 AM and 6-7 PM PPRBT and IRBT over 6-week moving date periods (over Sept.-Dec. 2012). In addition, Figure 6-3 shows the average line-direction passenger flow during each hour, measured in passengers per hour per direction (PPHPD). The Island Line is used as the example because it regularly experiences denied boardings, and its passenger flow follows a predictable pattern: high demand westbound (inbound towards Hong Kong’s CBD) and low demand eastbound during the morning peak, and vice versa during the evening peak.

The PPRBT results are generally consistent with the model’s assumptions. As expected, the line-direction PPRBT is most sensitive to capacity during each line-direction’s peak flow hour (8-9 AM eastbound, 6-7 PM westbound), when denied boardings should be most frequent. Furthermore, as the capacity is increased from 2000 per train to 2500, the PPRBT decreases. This makes intuitive sense: as capacity is increased, more passengers can board each train; for a fixed passenger demand, this will result in fewer passengers being denied boarding due to insufficient capacity. Fewer denied boardings, in turn, will lead to more reliable journey times, and thus lower PPRBTs.

For all other “non-peak” times of day, the PPRBT values are essentially identical for all three values of $C$, and the PPRBT plots collapse to a single line. Presumably, during these times there is insufficient demand for denied boardings to occur, for all $C \geq 2000$. During these times, the PPRBT responds as expected to the scheduled headway, dropping when the scheduled frequency increases (which should reduce wait time variability). This can be seen from 5-6 PM for the Island Line westbound, and 8-10 AM eastbound.

These results also indicate strong correlation between the line-level PPRBT and IRBT. For the PPRBT at hourly intervals (Figures 6-3a and 6-3b), the average difference was only 5.9 seconds (for all $C$ values). This correlation is strongest during off-peak hours, with the metrics’ magnitudes and trends similar both over the day, and for long-term trends (as seen in Figures 6-4b and 6-4c). During peak hours, while the two metrics’ magnitudes are generally similar (the gap never being greater than 35 seconds for any $C$ value), the two metrics exhibit significantly different trends over time, as shown in Figures 6-4a and 6-4d.
Figure 6-3: Line-level PPRBT and IRBT, and Avg. Line Flow, by time-of-day

Figure 6-4: Line-level PPRBT and IRBT, for rolling 6-week periods
This is a potential issue for reliability monitoring—the PPRBT could indicate reliability gains where passengers experience reliability deterioration, and vice versa.

Surprisingly, the off-peak PPRBT was occasionally slightly higher than the IRBT, particularly during the evening hours. It was expected to be lower than the IRBT, given that the PPRBT excludes station access and egress time variability, and no denied boardings should be taking place during the off-peak. This may indicate a flaw in the model’s uniform arrival rate assumption; for example, if passengers often run to catch trains before they depart, the platform arrival rate will not be uniform, and will exhibit a tighter waiting time distribution. Further research is required to resolve this question.

Another important finding is that there is no clear, consistent “best fit” $C$ value that best brings the line-level IRBT and PPRBT into agreement, which could have been considered a line-level effective vehicle capacity. This is evident in Figures 6-4a and 6-4d; for some date periods, the IRBT is below the $C=2500$ PPRBT, while for others, it is nearly equal to the $C=2000$ PPRBT. A potential explanation for the absence of an effective best fit $C$ could be the existence of variation in $C$ across stations and directions, as hypothesized in Section 6.1.2. This explanation is tested in the following section, by analyzing the OD-level PPRBT results.

### 6.3.2 OD-Level Results

To assess the OD-level PPRBT, the PPRBT and IRBT were calculated and compared for a set of 13 OD pairs (8 non-transfer and 5 transfer) that are generally representative of travel on the “core” of the MTR network. Figure 6-5 presents OD PPRBT and IRBT results for four of these OD pairs—two transfer, two non-transfer—at hourly intervals over the day. In Figure 6-5, the IRBT is only shown for the hours where a sufficient frequent traveler group size (30 frequent travelers, see Section 5.4.1) was obtained for the IRBT results to be significant. For the two non-transfer OD pairs (subfigures (a) and (b)), Figure 6-5 also shows the average line flow at each OD pair’s origin station, in the target direction. For the transfer OD pairs, the average line flow is shown separately for the origin and transfer station in Figure 6-6. Figure 6-7 shows long-term PPRBT trends for six OD pairs, for fixed time-of-day intervals; the first three OD pairs (subfigures a–c) experiencing high line-direction demand, and the second three OD pairs (subfigures d–f) experiencing low line-direction demand.
Figure 6-5: OD PPRBT and IRBT, and Avg. Line Flow, by time-of-day

Figure 6-6: Avg. Line Flow for each OD leg, by time-of-day
Figure 6-7: OD PPRBT and IRBT, for 6-week rolling date periods
The OD PPRBT results’ response to demand and capacity $C$ was generally similar to the line-direction PPRBT’s: the highest sensitivity to $C$ occurred during the times of peak line flow, with lower $C$ resulting in higher PPRBT values, consistent with expectations. The magnitude of this response varies greatly among OD pairs. OD pairs originating near the peak load point of the target line-direction—the point where denied boardings are most frequent—exhibit a substantial “spike” in the PPRBT during the peak hour(s) at lower $C$ values. OD pairs originating at less-trafficked locations (where denied boardings should be rare) exhibit little response to demand and $C$ values. This pattern is exemplified by the Mong Kok to Central (a) and Kwun Tong to Yau Tong (b) OD pairs. The former exhibits a significant spike in the PPRBT for $C=2000$ from 8-9 AM, where line flow peaks at 35,000 PPHPD. The latter experiences a peak line flow of only 25,000 PPHPD, and shows no spike in PPRBT.

For transfer OD pairs, the PPRBT was responsive to line flow at both the origin and transfer stations. When large peaks in demand occur at transfer and origin stations simultaneously, the effects appear to be additive, resulting in much larger PPRBT spikes than for high demand on a single line OD pair. This behavior can be seen for the Tsim Sha Tsui to Causeway Bay OD pair (Figure 6-5d). Between 8-9 AM, both origin and transfer point line flows are high, causing a 4-minute jump in the $C=2000$ PPRBT, and a 1-minute jump in the $C=2250$ PPRBT. During 6-7 PM, line flow is only high at the transfer point, producing a much smaller PPRBT response.

The OD PPRBT’s response to the scheduled headway was also as expected. In the absence of denied boardings, when scheduled frequency increased, the PPRBT decreased, as seen in Figure 6-5b.

The OD PPRBT’s relationship to the IRBT was found to differ substantially between the peak and non-peak times for each OD pair. During non-peak hours—when the PPRBT is insensitive to $C$, and denied boardings are infrequent—the OD PPRBT is generally a good estimator of the IRBT. The non-peak OD PPRBT followed time-of-day trends in the IRBT fairly well (for the limited hours for which the IRBT can be calculated), as seen in Figure 6-5, but did not consistently follow date period IRBT trends, as evidenced by Figures 6-7d and 6-7e. Even when the OD PPRBT trends deviated from the IRBT, the gap was never larger than 40 seconds, a small difference from the passenger’s standpoint.

For the ODs’ peak demand times, on the other hand, the PPRBT was found to be an
unreliable estimator of IRBT. For some OD pairs (about half of those sampled), the two metrics’ long term trends were found to be similar during the peak hour, as in Figures 6-7a and 6-7b. For such scenarios, a “best fit” $C$ should exist that brings the PPRBT and IRBT values into close alignment. For Mong Kok to Central, the “best fit” $C$ appears to be about 2000 per train, while for Wong Tai Sin to Central, this value appears to be greater than 2500. If the IRBT is assumed to be “accurate”, this best fit $C$ should be equivalent to the actual effective train capacity at the origin (and/or transfer) station, which, as discussed in Section 6.1.2, should depend on station-specific factors such as loading evenness, boarding behavior, etc.

For many other OD pairs, however, little IRBT-PPRBT correlation was observed during the peak demand hour, as for example in Figure 6-7c. For such scenarios, there exists no “best fit” $C$ that will consistently bring the PPRBT into agreement with the IRBT; rather, there is irresolvable disagreement between the two metrics about the typical passenger’s reliability experience. This results in unavoidable large discrepancies between the PPRBT and IRBT, regardless of the $C$ parameter chosen, for the particular OD pair. For OD pairs highly sensitive to the $C$ value, this can lead to the PPRBT under- or over-estimating travel time variability by several minutes, an amount noticeable to both passengers and analysts.

### 6.3.3 Conclusions from PPRBT Results

Overall, the PPRBT results appear to be consistent with the model assumptions with respect to headways, demand, and capacity. When demand is low relative to capacity, the PPRBT decreases during the morning and evening “rush hours”, due to higher-frequency service reducing passengers’ waiting times. On the other hand, when demand is high relative to capacity, clear “spikes” in the PPRBT are observed, due to denied boardings.

With respect to the IRBT, the PPRBT results are not as consistent. For non-peak times, when demand does not approach capacity, the PPRBT and IRBT are generally close, both at the OD level (including transfer ODs) and line-level. However, during peak periods, the PPRBT is an unreliable estimator of the IRBT both in terms of magnitude and trends over time, especially at the OD level. The causes of this discrepancy are uncertain, but likely stem from both the fundamental design of the PPRBT, as well as problems with the execution of the PPRBT on MTR.

The PPRBT’s design, by modeling only the platform-to-platform travel time, ignores
access and egress time variation. During peak periods, congestion can occur on escalators, stairs, and corridors within stations, potentially adding several minutes to one’s access and egress times, compared to “free-flow” conditions. Thus, access/egress time variability is likely a significant contributor to passengers’ overall travel time captured by the IRBT, but not the PPRBT. The PPRBT is also susceptible to error introduced by its simplistic model of the passenger boarding process—the “one car, one queue, first-in first-out” model, which is clearly not how passengers actually board trains on MTR. It is not known whether this error would tend to bias the estimated buffer time upward or downward in situations where denied boardings occur.

In terms of PPRBT execution, uncertainty in the input demand data (due to potential route choice errors) and capacity parameter $C$ introduces significant uncertainty into the model’s estimates of denied boardings, and thus, peak period PPRBT. The capacity uncertainty is of particular concern, given the high sensitivity of many OD pairs’ PPRBT to this parameter. For example, increasing $C$ from 2250 to 2500, an 11% increase, causes the 8-9 AM Wong Tai Sin to Central PPRBT (Figure 6-5c) to increase by more than 3.2 minutes, or 84%. Without a means to calibrate $C$, for many OD pairs the PPRBT cannot be calculated with any certainty during the peak hours on MTR.

6.4 Evaluation of the PPRBT

This section offers an evaluation of the PPRBT as a passenger-experienced reliability metric, based on its technical definition and results presented in this chapter. The evaluation is presented in two sections: first, an evaluation against the design criteria for passenger-experienced reliability metrics described in Section 4.1, and afterwards, a set of general recommendations and conclusions about the potential use of the PPRBT.

6.4.1 Evaluation Against Design Objectives

Representative of the Passenger Experience

Theoretically, the PPRBT should meet almost all criteria for being representative of the passengers’ experience. It should capture all sources of unreliability, except for walking time delays due to congestion in stations. Its definition as the difference between the 95th and 50th percentile travel times both distinguishes between early and late arrivals, and
excludes “extreme” delays. It is calculated at the OD level, with aggregate metrics being simply averages of OD-level results. Because it is not based on individuals’ travel times, it controls for variation in passenger behavior, and has no bias towards particular passenger demographics.

In practice, however, the PPRBT does not seem to represent the passenger experience consistently. As described in Section 6.3.3, while the PPRBT and IRBT are generally consistent in the off-peak, the two diverge significantly during peak hours. This is problematic, because the peak hours are precisely when the IRBT is most representative of the passenger experience of reliability, as that is when frequent travelers make up the largest portion of travelers; thus, the discrepancy probably indicates error in the PPRBT, rather than the IRBT. Furthermore, the peak hours are, for most transit operators, the time of most importance for measuring reliability, as they are the period when the most “stress” is placed on the system, both operationally (frequency is at a maximum), and in terms of passenger demand.

As noted in 6.3.3, much of this “error” is due to flaws in the input data, rather than the algorithm itself. However, as noted in 6.2, the input data available from MTR is likely to be the best that can be expected from any transit operator. Therefore, the PPRBT’s inability to estimate passenger-experienced reliability accurately under such circumstances is a severe limitation to its potential application in the “real world”, and, in effect, a significant design problem with the metric.

**Meaningful to Passengers and Non-Experts**

In terms of meaningfulness to passengers and non-experts, the PPRBT’s performance is similar to the IRBT. Like the IRBT, the PPRBT is expressed in terms of the reliability buffer time, the extra time a passenger needs to budget for unreliability. Thus, the PPRBT should, in theory, relate directly to passengers’ reliability experience. The PPRBT could be difficult for non-experts to understand, however, and thus would likely require explanation.

The PPRBT should also be reasonably objective. Besides the upper percentile \( N \), the only other significant parameter to be set is the capacity \( C \). While \( C \) may be difficult to set accurately, it does have an objective definition, as the maximum number of passengers who can typically board the train.

Like the IRBT, the PPRBT is unlikely to significantly help passengers plan their jour-
neys, as most passengers do not have a clear “reference” travel time to add the PPRBT to as a buffer time. Furthermore, the PPRBT does not include access and egress time variation, which can be significant.

**Comparable Across Services and Times of Day**

Like the IRBT, the PPRBT cannot be directly compared across transit services and times of day, because it is strongly influenced by the scheduled headway. This problem can be overcome, however, by calculating an “Excess PPRBT” analogous to the Excess IRBT described in Section 5.1.4, by taking the difference between the PPRBT and the minBT. This Excess PPRBT should be an absolute measure of reliability, and thus comparable across all services and times of day.

**Time Period Flexibility**

One of the potential advantages of the PPRBT over the IRBT is time period flexibility. Because it is not dependent on a large sample of frequent travelers, it can be calculated accurately during the off-peak hours, and for short date periods well below the effective 5-week minimum for the IRBT. Furthermore, the metric can be calculated at short (e.g., 15-minute) time-of-day intervals, even in the off-peak, without compromising accuracy. This flexibility is achievable equally at the OD and line-level, unlike the IRBT.

However, while the PPRBT can be calculated for essentially arbitrary time periods, it cannot be calculated accurately during periods of peak passenger demand, as described in 6.3.3. Thus, in this context, the PPRBT is effectively limited to the non-peak hours—the reverse of the IRBT.

**Service Scope Flexibility**

Another advantage of the PPRBT over the IRBT is service scope flexibility. Unencumbered by frequent traveler group size limitations, it can be calculated consistently for all OD pairs, even during the off-peak. This suggests that, if issues with representing peak hour reliability were resolved, the PPRBT could be an acceptable metric for use in a journey planner, or similar passenger information function.
6.4.2 PPRBT Conclusions and Recommendations

Overall, the PPRBT has great potential as a passenger-experienced reliability metric. It has been shown, through application to the MTR system, to achieve acceptable accuracy at estimating the typical passenger’s buffer time for a wide range of times and locations, at both the OD and line-level. Furthermore, it is able to achieve this without direct passenger travel time data, providing the PPRBT not only with greater time period flexibility, but the potential to be implemented on the many transit systems without the high quality, closed-system AFC data required by the IRBT.

However, the PPRBT also suffers from a significant limitation: the inability to consistently and accurately measure passengers’ travel time variability during periods when denied boardings take place. The PPRBT, then appears to fail in achieving one of its design objectives, to be able to accurately model the effect of denied boarding delays on passengers. That said, the PPRBT results show the metric has the potential to model denied boardings, as the PPRBT model consistently predicts their occurrence at times of high demand, and responds appropriately, directionally, to changes in the assumed capacity. It is thus possible that, with additional research, the PPRBT could effectively measure denied boardings’ effect on passengers’ reliability experience.

Given the current limitations of the PPRBT, and the ability to calculate the IRBT, the PPRBT as currently formulated is not recommended for general use on the MTR system. For MTR, where denied boardings are common, and a significant concern for both passengers and management, it is necessary for any passenger-experienced reliability metric to accurately capture this phenomenon. However, for transit systems or services where denied boardings are not common, the PPRBT (or a simplified version without denied boardings) could be an effective measure of the passenger’s experience of reliability.
Chapter 7

Implementation of the IRBT

Thus far in this thesis, discussion of new passenger-experienced reliability metrics has focused on their design, and their inherent properties. However, the benefits of such metrics can only be realized in practice if they are implemented in an effective, well-thought-out process. Even the best-designed metric, if poorly implemented, will fail to produce actionable information for the transit agency and its stakeholders; or even worse, produce misleading results. There are four key considerations for effectively implementing a new transit service reliability metric:

- How should the new metric’s use be integrated with the agency’s other metrics, to complement their strengths and weaknesses?

- Transit Service Scope: Will the metric be calculated at the OD level, line level, or some other grouping of OD pairs? Will it be calculated for single-line journeys, or journeys involving transfers?

- Calculation Schedule: Will the metric be calculated periodically? If so, how frequently?

- Time Periods: What time-of-day intervals (e.g., hourly) and date periods will be used for the calculation?

There is no single “right answer” for these questions, for a given metric. The appropriate decisions depend on the characteristics of the transit services being evaluated, the particular metrics already used by the transit agency, and the specific application for the new metric.
This can entail a single metric being calculated differently depending on the application, even within the same operator.

This chapter aims to develop an effective implementation process for the IRBT. Section 7.1 describes how the IRBT can be best integrated with an operator’s currently-used reliability metrics, using MTR as an example. Sections 7.2-7.4 discuss strategies for setting the service scope, calculation schedule, and time periods, respectively. Not covered is implementation for journey planner or similar passenger information applications, because, as noted in Section 5.5, the IRBT is not recommended for such applications. Section 7.5 demonstrates the IRBT implementation process developed in Sections 7.1-7.4 in a case study: an assessment of the effects of the 2014 Hong Kong political demonstrations on MTR’s reliability.

7.1 Integration with Existing Reliability Metrics

Effective integration with existing reliability metrics is necessary for successful implementation of the IRBT. Such integration is achieved by determining, among the IRBT and the operator’s current reliability metrics, the metric, or set of metrics, that best meet the design needs for each of the operator’s reliability measure applications. *This analysis will determine which applications the IRBT should be used for, and which ones it should not.* With such integration, the IRBT can provide better information than if implemented in isolation. Without this analysis, however, the IRBT could be implemented in applications it is ill-suited for, while being passed over for applications where it could provide valuable insights not obtainable with other metrics.

The following process is proposed for effectively integrating the IRBT and an operator’s existing reliability metrics:

1. Assess the metric design requirements for each of the operator’s reliability applications.
2. Assess the design characteristics of current reliability metrics, and compare them with the IRBT’s design characteristics.
3. For each reliability application, find the reliability metric (IRBT included), or metric combination, whose design attributes best match the application’s design needs.

This section describes this process in detail, and demonstrates its application using the MTR
system and its two current reliability metrics as an example. To facilitate understanding the MTR example, this section begins with a review of MTR’s current reliability metrics.

7.1.1 Current MTR Reliability Metrics

MTR currently uses two main reliability metrics for most reliability measurement applications: the Passengers Affected Ratio, or PJOT, and the number of train delays $\geq X$ minutes.

**Passenger Journeys on Time (PJOT):** The PJOT, described in more detail in Section 4.3.3, is an estimate of the percentage of total passenger trips to the number of passenger trips not affected by a train delay $\geq 5$ minutes. Total passenger trips are derived from AFC data, while the delayed trips are estimated as the average train load multiplied by the number of delayed trains (from AVL data). It is generally calculated at the line-level.

**Number of delays $\geq X$ minutes:** This metric is actually a set of four similar metrics: the number of train delays lasting X minutes or more, where X is 2, 5, 8, or 20 minutes. These four thresholds effectively classify train delays into four levels of severity. These numbers are used, like the PJOT, for reliability monitoring, internal reporting, and performance target setting.

7.1.2 Assessment of Metric Design Requirements for Applications

The metric design requirements for most reliability measurement applications can be described by the following five general design criteria, adapted from the design objectives in Section 4.1:

- *Sources of unreliability captured:* What sources of unreliability is the metric sensitive to? (incidents, general train delay, denied boardings, station congestion, etc.)

- *Passenger delay or train delay?* Is the metric defined by delays to trains, or passengers’ delay and/or travel time variability? (e.g., OTP measures train delay, IRBT measures passenger delay)

- *Comparability across services and times of day:* Can the metric’s values be compared across services in the network, and different times of day, or are they dependent on
each service’s schedule?

- **Service scope flexibility**: What transit service scopes is the metric calculable for? (OD level, line-direction level, line level, etc.)

- **Ease of identifying unreliability causes**: How easy is to identify the specific causes of reliability trends when they appear? (e.g., increase in incidents vs. increase in demand)

These should not be considered the “definitive” set of design requirements for reliability measurement applications, however. Depending on an operator’s preferences and applications, different sets of characteristics may be better—for example, if only line level reliability measurement is of interest, service scope flexibility will not be important.

The main considerations in assessing a particular application’s requirements for each of these design criteria are explained below, using examples from the reliability measurement applications described in Chapter 3.

**Sources of Unreliability Captured**: The sources of unreliability to be captured for a particular application depend on two considerations. First, what unreliability sources significantly impact passengers using the services of interest? This can vary across transit systems. For example, MTR passengers are greatly affected by denied boardings and station congestion in the peak, but these delays are absent in a lightly-used bus network. Second, what reliability sources does the audience care about? When the audience is the general public—passenger information, public reporting, etc.—all factors that cause significant delays to passengers are of concern. However, if the audience is an internal group working on a specific reliability issue—e.g., incident recovery—only a few unreliability sources will be of interest.

**Passenger Delay or Train Delay Measure?**: Passenger unreliability measures are strongly recommended when an application falls into either of two categories: First, when passengers are the primary audience (e.g., public reporting), because passengers will better understand reliability information that relates directly to their journey experiences. Second, when passenger costs of unreliability need to be quantified (e.g., equity analysis, cost-benefit analysis), because passenger delay measures will better reflect passengers’ reliability costs than train delay measures.
Train delay measures are recommended for applications dealing directly with train operations management (scheduling, dispatching, etc.): such applications are usually a part of an operator’s reliability management program. If an application doesn’t fall into either of these two categories, either passenger or train delay measures can be used if they meet the other design criteria.

Comparability Across Services and Times of Day: This requirement arises for applications requiring reliability comparison across transit services and schedules. Such applications include:

- Equity analysis across areas served by different services (e.g., “transit reliability in Area A is worse than in Area B”).
- Reliability cost-benefit analyses for projects affecting reliability across multiple services.
- Analysis of reliability trends across services and schedules as part of reliability management, including evaluating the effects of a schedule change, comparing reliability over the service day, or comparing operational practices across lines.

Service Scope Flexibility: Reliability metrics should be calculable at the service scope at which actions will be taken for the application. For example, with reliability management most interventions (e.g., schedule changes) take place at the line-level, so measures should be defined at this level. Passenger information’s purpose, however, is to aid passengers’ travel decisions at the OD pair level, and is thus best provided at that level.

Ease of Identifying Unreliability Causes: An application’s need for easy identification of unreliability causes from trends in metrics depends on the audience’s involvement in operations management. For public-facing applications, this capability is not necessary. The public does not generally care why service is unreliable; they see that as the operator’s problem to solve. However, if the audience is the internal group responsible for day-to-day management of service operations, this ability would be critical. Senior management, as an audience for internal reporting, fall in between; they are likely to be interested in the general causes of unreliability problems, so resources can be directed to solve them, but not in the low-level operational details of the unreliability causes.
7.1.3 Assessment of Current Metrics

The assessment of a public transport operator’s current reliability metrics is best accomplished using the same general design criteria used to assess the application metric design requirements. Thus, if the five general criteria proposed in the previous section are used to evaluate application needs, they should also be used to evaluate an operator’s current metrics. Such an assessment is demonstrated below for the MTR’s current reliability metrics:

Sources of Unreliability Captured: Both the PJOT and number of \( \geq X \) minute delays are, by definition, only sensitive to incident delays. Neither are, like the IRBT, explicitly sensitive to denied boardings, station congestion, or interaction between multiple sources of unreliability.

Passenger Delay or Train Delay?: The number of \( \geq X \) minute delays is defined in terms of train delay. The PJOT, on the other hand, is meant to estimate the frequency of significant delays to passengers (from a single source, major incidents)—thus, it is a passenger delay measure, similar to the IRBT.

Comparability Across Services and Times of Day: Neither the PJOT nor number of \( \geq X \) minute delays are dependent upon the timetable, so they should be comparable across MTR lines and times of day. The IRBT, as discussed in Section 5.5, is not normally comparable across services, because it is affected by the scheduled headway. However, if calculated in the form of the Excess IRBT, which controls for the scheduled headway, the IRBT can be compared across lines and time periods.

Service Scope Flexibility: Both of MTR’s reliability metrics are calculable at the line-direction level, line level, and aggregations of multiple lines (e.g., entire network level). Thus, these metrics cannot be calculated for spatial scopes such as OD pairs, transfer flows, and geographic segments (e.g., into CBD, out of CBD), which are available with the IRBT.

Ease of Identifying Unreliability Causes: The focus of the PJOT on incident delays makes “diagnosing” trends in this measure straightforward—they are always due to changes in incident frequency and/or severity. This process is even simpler for the number of incidents \( \geq X \) minutes, which are explicitly defined in terms of incident frequency and severity.
Identifying causes of trends in IRBT is significantly more challenging because of the IRBT’s sensitivity to many more sources of unreliability. Deducing the specific source, or combination of sources, driving an IRBT trend from this multitude of potential causes is no small task. This task is made harder by the spatial and temporal correlations among these factors, making “controlling” for particular factors difficult. For example, denied boardings, station congestion, and long dwell times are all correlated, because they are all related to peaks in passenger demand.

7.1.4 Matching Metrics to Applications

Once the application design and metric design needs assessments are complete, matching application needs to available metrics should be fairly straightforward. The objective should be to find the set of metrics that meet the application’s design needs, at the lowest “cost”. For example, if a single metric can meet all the application’s needs, only that metric is needed; using more metrics would increase implementation effort, but with little benefit. Such matching can give three results with respect to the IRBT:

- **IRBT Alone Recommended**: The IRBT alone is sufficient to meet the application’s needs, or, if several metrics can meet these needs, the IRBT is the most effective.

- **IRBT Not Recommended**: Existing metric(s) can meet all the application’s needs, and do so more effectively than the IRBT. Using the IRBT would only add interpretation cost, without additional benefit; thus, the IRBT is not recommended.

- **IRBT in Combination with Existing Metrics**: The application’s needs can be met most effectively through a combination of the IRBT and existing metrics.

This matching process is demonstrated below for a hypothetical MTR application, using MTR’s current reliability metrics (see Section 7.1.1) and the IRBT. This process is also demonstrated in the case study application in Section 7.5.2 (with the result being “IRBT in Combination with Existing Metrics”).

**Geographic Equity Analysis (IRBT Alone)**

**Application Analysis**: To inform the planning of future line extensions and improvements, passengers’ experienced service reliability is to be compared across geographic areas...
of Hong Kong. Areas experiencing worse reliability will be prioritized for projects improving connectivity to Hong Kong’s CBD. Scheduled wait time variability is to be included in unreliability “cost”, because frequency changes are a potential improvement.

- **Sources of unreliability captured**: Audience (planning staff) interested in all sources of unreliability affecting passengers. Includes denied boardings, station congestion; only IRBT captures these sources.

- **Passenger Delay, Train Delay**: Because passengers’ costs of unreliability need to be quantified, measurement of passenger delay is necessary; PJOT and IRBT qualify. Train delay measurement is not needed.

- **Comparability Across Services**: Required, because reliability is to be compared across geographic areas served by different lines. All metrics qualify; because the scheduled wait time is to be included in the passenger cost, “normal” IRBT can be used.

- **Service Scope Flexibility**: Reliability results are to be aggregated by geographic area, which can encompass several lines. Furthermore, reliability is being assessed for many OD pairs requiring transfers, rather than exclusively single-line journeys. Only the IRBT can provide this service scope flexibility.

- **Ease of Identifying Unreliability Causes**: Not required, only reliability effects on passengers are of interest.

**Recommendation**: The IRBT is able to fulfill all of the requirements for this application, unlike any of the current MTR metrics. Thus, the IRBT alone is recommended for this reliability measurement application.

### 7.2 Service Scope

The service scope for IRBT implementation can be defined by the set of OD pairs included in the IRBT calculation—single OD pairs, all single-line OD pairs (line-level), all single line-direction OD pairs (line-direction level), etc. This section reviews the main decision factors in establishing an appropriate IRBT transport service scope.
Service Scope of Unreliability Sources of Interest

Different sources of unreliability affect service at different service scopes. Incidents, dispatching issues, and signal problems generally affect transit service at the line/route level, while weather-related problems are likely to have system-wide effects. Demand changes can be either system-wide or spatially concentrated (e.g., opening of a new sports venue). To assess the effects of specific sources of unreliability on passengers, the IRBT’s service scope should match that of the unreliability sources of interest. Doing so will maximize the IRBT’s sensitivity to these factors, by excluding unaffected passengers (if applicable), while ensuring the results are representative of all passengers who were affected.

Setting the spatial scope in this manner is important for evaluating the reliability impacts of service changes, which generally have a clearly-defined spatial scope (e.g., signal upgrade on a particular line, signal priority on a particular street). It can also aid the diagnosis of unreliability issues, by “isolating” certain services, or service sections, where reliability problems are suspected (e.g., passenger congestion in a particular station).

Specific Reporting Needs

For some reliability measurement applications, the spatial scope is explicitly defined; in such cases, the IRBT’s scope should be similar. A common example is the public reporting of reliability for each line/route, best done with the line-level IRBT for each line in the network.

For some potential applications, reliability comparisons may be needed at specific geographic scales. Assessing equity of service provided, for instance, could entail comparing service reliability across neighborhoods. This could be accomplished using a “neighborhood-level” IRBT, defined as the average IRBT for all OD pairs beginning and/or ending in the neighborhood (weighted by OD passenger flow). A potential advantage of such geographic groupings is their inclusion of transfer OD pairs, which account for the majority of passenger trips on MTR. The spatial flexibility of the IRBT permits the metric to be calculated for essentially any service scope that furthers reporting and analysis functions (as long as a sufficient frequent traveler group size is available).
Ease of Interpretation

In some situations, the spatial scope decision involves a trade off between better information and ease of use. For example, the line-direction IRBT better shows the effects of demand-related delays, which for certain lines can be significantly greater in one direction than the other, as exemplified in Figure 5-17. However, MTR to date has generally monitored reliability using line-level metrics; thus, the line level IRBT would likely fit better with MTR’s current procedures.

Frequent Traveler Group Size Constraints

As shown in Sections 5.4.1 and 5.4.2, the spatial extent of the IRBT significantly affects the frequent traveler group size, and thus, its ability to meet the group size requirements. As noted, it is more difficult, for a given time period, to meet the sample size requirements at the OD pair level, than at a more aggregate level (e.g., line-direction level). Thus, it is recommended that the IRBT be calculated at an aggregate level, if possible.

7.3 Calculation Schedule

The calculation schedule for an IRBT application involves two related decisions: First, is the IRBT to be calculated on a regular, continuing basis (like a stock index, for example), or for a single set of date periods? Second, if the application requires regular calculation, how frequently should it occur (weekly, monthly, quarterly, etc.)? These decisions largely depend on whether the application falls into three categories: Reliability Monitoring, Reporting, and Event Analysis, as described below.

Reliability Monitoring

Reliability monitoring (part of Reliability Management described in Section 3.1) can be described as the use of reliability metrics to detect reliability problems as they arise, so action can be taken to correct them. To consistently detect reliability problems with the IRBT, it must be calculated on a regular basis, for each operated service. This regular IRBT calculation should occur frequently enough to identify reliability problems quickly after they arise. If the IRBT is calculated infrequently, there can be a long delay between a reliability issue arising, and being detected. For example, a problem could affect passengers’ reliability
for three months before being detected by a quarterly IRBT—a much longer response time than is desirable.

However, there are “diminishing returns” with increasing the calculation frequency to improve the IRBT’s response time. Most reliability issues requiring a metric like the IRBT to detect (i.e., problems not immediately apparent to the operator) are unlikely to arise overnight. Thus, very frequent IRBT updates (e.g., daily) will produce more data, but not necessarily more meaningful information, than moderately frequent IRBT updates (e.g., weekly).

**Reporting**

Reporting applications generally require performance metrics like the IRBT to be reported on a regular basis. This is necessary in order to establish accountability of the operator to its stakeholders; if the operator is able to “pick and choose” periods to report, they could selectively choose to report only periods with good performance. The frequency of IRBT reporting should be enough to give stakeholders a general idea about system reliability. However, the frequency should not be so high as to “overwhelm” these stakeholders with more data than they have the patience or background knowledge to interpret. Where possible, the IRBT should be reported at the same frequency as the operator’s other reporting metrics.

**Event Analysis**

Event Analysis applications are those involving the analysis of reliability with respect to a specific event or series of events. Such applications do not involve IRBT calculation on a regular basis, but rather for a set of specific periods, defined with respect to the event(s). An example event analysis application is the evaluation of a service change: To determine how a service change affected reliability, the IRBT should be calculated for periods immediately before and after the service change took place. The case study presented in Section 7.5 is an example of an Event Analysis application.
7.4 Time Period Setting

This section reviews the main decision factors—some objective, others subjective—involved in setting time-of-day intervals and date periods for a given IRBT application. As will be shown, the time-of-day interval and date period decisions are closely linked; a decision regarding time-of-day intervals affects the available date period choices, and vice versa. The decision factors are summarized in Table 7.1 at the end of this section.

7.4.1 Timeframe of Reliability Factors of Interest

If the IRBT is used to shed light on a particular potential source of passenger-experienced unreliability, the IRBT’s time-of-day interval(s) and date period(s) should match the timeframe over which trends in this factor are expected to occur. If the trend is expected over a specific time-of-day period, the IRBT time-of-day interval should match. If the trend is expected to affect only certain days (e.g., weekdays), only those days should be included in the date period.

Matching the IRBT in this manner maximizes the sensitivity of the IRBT to the unreliability sources of interest, by excluding frequent travelers who traveled at times unaffected by the unreliability source(s). If a large number of such “non-affected” frequent travelers are included in the IRBT, they can “wash out” the experiences of the affected travelers in the IRBT’s calculation. The timeframe matching process is described in more detail below for several of the most important factors affecting passengers’ reliability experience:

- **Passenger Demand:** Delays related to high passenger demand, such as denied boardings, generally occur during the “peak period” of the service day. Such delays are also most likely to occur on weekdays, since demand is generally lower on weekends and holidays.

- **Incidents:** Unlike demand-related delays, train service incidents—breakdowns, signal failures, etc—occur over the whole day. Thus, measuring incident-related unreliability involves using time-of-day intervals that span the entire service day, rather than just the peak hours. If the goal is to exclude incident-related delays from reliability analysis (to enhance the IRBT’s sensitivity to other factors), a method of doing so is to exclude from the date period all days where significant incidents occur, during the time-of-day interval(s) of interest.
- **Evaluation of Service Change:** To evaluate a service change using the IRBT, the time-of-day intervals and date periods should encompass the span of service and type of service day (weekdays, weekends, holidays, etc.) affected. Ideally, two date periods should be used for service change evaluation: one entirely before the change, and one entirely after.

### 7.4.2 Rate of Variation in Reliability Factors

The rate of variation in the reliability factors of interest should be a significant factor in decisions regarding time-of-day interval length and date period length. In general, reliability factors with rapid time-of-day/date period variation should be assessed using short time-of-day intervals/date periods, in order to observe the reliability effects of this variation in the IRBT. Meanwhile, reliability factors with gradual time-of-day/date period variation should be assessed using longer time-of-day intervals/date periods, in order to minimize the IRBT's sensitivity to exogenous short-term reliability factors.

Short time-of-day intervals are important for observing the effects of time-of-day variation in passenger demand, because passenger demand often changes rapidly during the peak hours, with the worst demand-related delays at the short “peak of the peak”. This is illustrated in Figure 7-1, which shows (for Sept.-Oct. 2012) the Kwun Tong Line eastbound IRBT for 15, 30, and 60-minute intervals during the AM and PM peak hours, alongside the peak point passenger flow (in PPHPD, at 15-minute intervals). In both the AM and PM peaks, significant “spikes” in travel time variability are visible at the 30-minute and 15-minute levels, which closely follow trends in passenger demand. This close temporal correlation between demand and IRBT is obscured, however, if 60-minute intervals are used.

With respect to date period trends, on the other hand, passenger demand trends are generally long-term (for example, MTR’s ridership growth is about 4% per year). Thus, to assess the impacts of long-term passenger demand changes, long date periods should be used (e.g., quarterly). Such long date periods should minimize IRBT variation from short-term sources of unreliability, such as the occurrence of major disruptions, that would add noise to the analysis of demand trends.

Regarding the evaluation of service changes, if the change affects only a small portion of the service day (e.g., a small timetable change), the IRBT can potentially be evaluated over
a single time-of-day interval. However, a service change that affects a large portion, or all, of the service day should be evaluated for multiple time-of-day intervals. At a minimum, the peak hours should be evaluated separately from the off-peak hours, as most transit system elements are placed under much more “stress” during the peak. The date period length for service change evaluation should be long enough to “wash out” IRBT variation from shorter-term reliability factors (e.g., major incidents), but short enough not to be biased by exogenous long-term reliability trends (e.g., passenger demand trends).

The considerations in this section and Section 7.4.1 are applicable even when the aim is simply to measure “general reliability” with the IRBT. The time-of-day intervals and date periods chosen will always affect the sensitivity of the IRBT results to certain reliability factors, because any time period will be a better “match” for some reliability factors than others. Thus, to correctly interpret IRBT results for “general reliability” measurement purposes, one must consider how the time periods used affected the IRBT response to the various reliability factors of interest. Or, more proactively, the time-of-day interval and date period can be set intentionally to prioritize the IRBT’s sensitivity to the reliability factors of greatest importance to the analyst (e.g., if congestion delays are most important to stakeholders, longer date periods should be considered).

### 7.4.3 Frequency of Information Update

A decision factor affecting the setting of date periods is how frequently reliability information needs to be updated for the particular application (as set according to Section...
7.3). Generally, if frequently updated reliability information is required, shorter date period lengths are desirable since then the IRBT will respond more quickly to reliability trends. On the other hand, if an application does not need frequent updates, longer date periods are acceptable—and may even be preferred, as they should minimize the effects of short-term variation.

7.4.4 Target Audience

The target audience for the IRBT results should be a significant factor in defining time periods. If the audience is knowledgeable about transit operations, and is willing to invest time in interpreting the IRBT results, IRBT can be reported for many time-of-day intervals and date periods, to provide a detailed picture of the reliability patterns on the transit services in question. However, if the audience is less familiar with the details of transit management (e.g., the public, government leaders), or is too busy to spend time digesting large amounts of data (e.g., senior management at the transit agency), providing IRBT results for many time-of-day intervals could be “information overload”. In such situations, IRBT results should be presented for fewer, and longer, time-of-day intervals; for example, AM peak, off-peak, and PM peak.

7.4.5 Frequent Traveler Group Size Constraints

As described in Section 5.4, the need to obtain a sufficient frequent traveler group size for the IRBT, 30 at the OD-level and 200 at the line-level, imposes significant constraints on the time periods for which the IRBT can be calculated, including:

- Cannot generally be calculated for date period lengths shorter than 5 weeks
- Cannot generally be calculated for OD pairs in the off-peak (with a few exceptions)
- Long time-of-day intervals generally required at the line-level in the off-peak (1 hour or more)

These (and similar) constraints can preclude the use of the IRBT for certain reliability measurement applications. For example, the IRBT could not be effectively used to provide OD-level passenger information for off-peak times, given the second constraint above. The
IRBT is also ineffective at analyzing the passenger delays resulting from individual, single-day events (such as a major disruption), because a relatively long (greater than a month) date period is required.

The table below summarizes the time-of-day interval and date period recommendations for the IRBT, with respect to the four main decision factors described in this section:

Table 7.1: Time Period Decision Factors Summary

<table>
<thead>
<tr>
<th>Decision Factor</th>
<th>Time-of-Day Interval Considerations</th>
<th>Date Period Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability Factor Timeframe</td>
<td>Match to time-of-day interval most affected by reliability factor (e.g., pax./demand in peak)</td>
<td>Match to service days most affected by reliability factor (e.g., pax./demand on weekdays)</td>
</tr>
<tr>
<td>Reliability Factor Rate of Variation</td>
<td>High rate, magnitude of time-of-day variation: short time intervals. Little time-of-day variation: long time intervals</td>
<td>High rate, magnitude of date period variation: short date periods. Slow, long-term variation: long date periods</td>
</tr>
<tr>
<td>Frequency of Information Update</td>
<td>N/A</td>
<td>If IRBT needs to be updated frequently, shorter date periods recommended</td>
</tr>
<tr>
<td>Target Audience</td>
<td>Knowledgeable: many time intervals. Not knowledgeable (e.g., the public): few, long time intervals</td>
<td>N/A</td>
</tr>
<tr>
<td>Frequent Traveler Group Size</td>
<td>Cannot be calculated for OD pairs in off-peak. Long time intervals required for line-level in off-peak</td>
<td>Cannot be calculated for date periods shorter than 5 weeks</td>
</tr>
</tbody>
</table>

7.5 Case Study: Evaluating Effects of Political Protests

This section presents a case study of how the IRBT can be applied to “real-world” applications, using the framework developed in Sections 7.1-7.4. The application is an assessment of the effects of the Fall 2014 political protests in Hong Kong on MTR’s reliability. The occupation of major arterial streets that occurred during the demonstrations severely disrupted the city’s surface transportation network, driving MTR ridership on nearby lines to record levels. This section’s analysis aims to address two questions:

1. What impact did the protests have on MTR passengers’ experienced reliability?

2. If the reliability impact was significant, by what mechanism did it occur? (increased incidents, increased denied boardings, etc.)
Answering these questions not only helps plan for future demonstrations, but can provide insight into a more fundamental and important question: how does passenger demand affect MTR’s ability to provide reliable service?

The case study is presented in three parts. Section 7.5.1 provides background on the protests and their effects on passenger demand and train loading. Section 7.5.2 applies the process described in Sections 7.1-7.4 to select the appropriate set of metrics, service scope, calculation schedule, and time periods for this application. Section 7.5.3 presents and analyzes the IRBT results of this application, while Section 7.5.4 provides more in-depth causal analysis, including results from other reliability metrics.

7.5.1 Background on Effects on MTR Service

On September 28, 2014, political demonstrators from the “Occupy Central” movement began their several month occupation of major arterial roads in the Central/Admiralty, Causeway Bay, Mong Kok, and Tsim Sha Tsui districts in central Hong Kong. These encampments effectively made the areas shown in red in Figure 7-2 impassible to surface traffic, causing severe disruptions to Hong Kong’s surface transport network. At the protests’ height, 270 bus routes were diverted or suspended, affecting 1.5 million daily riders [31]. MTR service (lines and stations shown on the map), however, was not disrupted by the protests. The protests lasted until December 11, although the first encampment clearances began on November 18.

With MTR the only effective transport through the affected areas, system ridership increased significantly during these demonstrations. Table 7.2 shows the average weekday ridership change by line and system-wide, measured by station entries. Two periods are compared: the first six weeks of the protests, denoted the “Protest Period”, and the preceding six weeks, the “Base Period”. System-wide, weekday ridership increased by 9.6%, or 441 thousand. The large majority (73%) of this ridership increase occurred on the two lines closest to the “occupied” areas: the Island and Tsuen Wan Lines (blue and red in Figure 7-2, respectively). Not surprisingly, these lines also had the greatest relative increases in demand—particularly the Island Line, registering a striking 27.7% increase.

---

1Note: Transfer station entries are assigned to a single line, as follows. **Island Line:** North Point, Quarry Bay. **Tsuen Wan Line:** Central, Admiralty, Yau Ma Tei, Mong Kok, Prince Edward, Mei Foo. **Kwun Tong Line:** Kowloon Tong. **Tsuen Kwan O Line:** Yau Tong, Tiu Keng Leng. **East Rail Line:** Hung Hom, Tai Wai. **Tung Chung Line:** Hong Kong, Nam Cheong.
Breaking down the change in average weekday journeys by half-hour in Figure 7-3, it can be seen that the demand increase is concentrated in the peak hours, and especially in the morning peak. Sharp peaking behavior is evident, with the demand response rising from 17,158 additional passengers from 7–7:30 AM to 36,567 additional passengers (the maximum half-hour response) from 7:30–8 AM—a 214% rise, within a single hour of the service day. Interestingly, this peak demand increase occurs early in the AM peak, rather than in the “peak of the peak” of 8:30-9 AM. This is likely due to the “Early Bird Discount” promotion taking place concurrently, where passengers received at 25% discount for exiting at 29 central stations from 7:15-8:15 AM.

During the peak periods, this demand increase led to increased line passenger flow (as estimated by MTR’s algorithm described in Section 6.2). As with raw demand, the Island and Tsuen Wan Lines registered the largest and second-largest peak line flow increases, respectively, in both absolute (PPHPD) and relative (capacity utilization) terms, as shown.

\[\text{Adapted from [32]}\]
in Figures 7-4 and 7-5 for the weekday average AM and PM peak half-hour flow\(^3\). As a result, these two lines’ capacity utilization at the peak load point, already near 90% before the demonstrations, surged to over 95% for the Tsuen Wan Line, and to 98% for the Island Line in the PM peak. With such extreme capacity utilization, denied boardings and passenger-related delays are expected to have risen substantially on the Island and Tsuen Wan Lines during the protest period.

\(^3\)i.e., the half-hour time-of-day interval within the AM and PM peak periods with the highest average line flow. This can occur at different times for different lines. For example, the AM peak half-hour is 8:30-9 AM for the Island Line, but 8:15-8:45 AM for the East Rail Line.
Figure 7-4: Peak Half-Hour Line Flow (at Peak Load Point), by Line

Figure 7-5: Peak Half-Hour Capacity Utilization (at Peak Load Point), by Line
7.5.2 IRBT Implementation

With the application and reliability factors of interest defined, the IRBT can be implemented according to the procedures described in Sections 7.1-7.4:

Integration with Existing Reliability Metrics

For this application, the explicit objective is to capture the passenger’s experience of reliability, including all sources of unreliability—incidents, denied boardings, station congestion, etc. The IRBT is the only metric available that satisfies this requirement. Comparability across services is not needed, because the goal is to characterize the change in reliability due to the protests, rather than compare reliability across services. Thus, the “regular” IRBT can be used, rather than the Excess IRBT.

To effectively identify the causes of reliability changes, the “all-inclusive” IRBT needs to be accompanied by one of MTR’s two current, incident-focused metrics. The number of incidents $\geq X$ minutes is selected for this case study, because it provides a more detailed picture of the frequency and magnitude of incidents on the MTR system. It is acceptable that this metric does not measure passenger delay, because its role is only to assist the interpretation of the IRBT, rather than be a “primary” indicator of reliability change.

Service Scope

There is no explicit, a priori transport service scope for this analysis. Instead, part of assessing the impact the protests had on MTR passengers’ reliability is identifying the parts of the network significantly affected by the event. This is done by examining the ridership and train load change figures from the previous section, as well as the geographic location of the protests. The results, and the map in Figure 7-2, clearly indicate the Island and the Tsuen Wan Lines were the parts of the MTR network most strongly affected. Thus, the service scope should focus on these two lines; using a larger scope could cause the protest effects to be “diluted” by the reliability experiences of passengers who did not travel through the protest-affected parts of the MTR network.

To measure the protests’ impact effectively, it is also important to capture the substantial variation in ridership increase occurring within this service scope. Separate measures of Island and Tsuen Wan Line reliability change are needed, to determine whether the much
greater ridership jump on the Island Line caused a significantly greater change in reliability. Because demand on these lines is also known to vary significantly by direction (particularly the Island Line), it is useful to measure reliability for each direction separately. These requirements can be fulfilled by line-direction IRBTs for the Island and Tsuen Wan Lines.

Using solely the four line-direction IRBTs is insufficient, however, because they do not capture the protests’ effect on passengers transferring between the Island and Tsuen Wan Lines. Such passengers are a large portion of the traffic through the affected areas, so it is important that their experiences be taken into account. In particular, it is useful to observe if reliability effects are cumulative for travelers using both lines. To assess transfer OD pairs’ reliability, two additional aggregate IRBTs are calculated: that for all OD pairs beginning on the Island Line and ending on the Tsuen Wan Line, denoted the “Island to Tsuen Wan Line IRBT”, and the reverse, denoted the “Tsuen Wan to Island Line IRBT”.

Calculation Schedule:

For this application, continuous calculation is not needed. Two defined date periods are needed: one before the protests began on September 29, and one at the height of the protest period, September 29 to November 18.

Time Periods

Date Period: To control for normal seasonal demand variation, relatively short date period lengths are desired. However, the date period should also be long enough to obtain a robust frequent traveler group size at the line-direction level. A 6-week period is a good compromise between these two considerations. Because weekends generally have lower demand, they are excluded from the date periods. The two date periods are thus defined as follows. The “Base period”: weekdays Monday, August 18 to Friday, September 26. The “Protest period”: weekdays Monday, September 29 to Friday, November 7.

Time-of-Day Interval: Because the protests caused (or are at least coincident with) ridership increases over the entire service day, the IRBT time-of-day intervals should cover the entire day, to measure the full extent of passenger impacts. In terms of interval length, short intervals (i.e., less than an hour) are needed during the peak hours, the AM in particular, to capture how the sharp peaking in demand response affected passenger reliability. However,
the interval must also be long enough to satisfy frequent traveler group size requirements. Finally, to facilitate interpretation of results, it is desired that the time-of-day intervals be consistent across different aggregate IRBTs.

To balance these requirements, the following procedure was used to set the specific time-of-day intervals:

1. The service day was divided into six general “service periods”: AM Peak, 7–9 AM; AM Peak Shoulders, 6–7 AM and 9–10 AM; Mid-day, 10 AM–4 PM; PM Peak, 5–7 PM; PM Peak Shoulders, 4–5 PM and 7–8 PM; and Late Evening, 8 PM–12 AM.

2. For each spatial scope (e.g., Island Line westbound), and both date periods, frequent traveler group sizes were calculated over the day at 30, 60, and 120-minute intervals. (15-minute intervals were considered, but rejected as an excessive level of granularity for this application)

3. For each spatial scope and service period, the minimum time-of-day interval length was identified such that all intervals falling within the service period met the minimum frequent traveler group size (150), across both date periods.

The results are presented in Table 7.3 (Island Line abbreviated “ISL”, Tsuen Wan Line “TWL”, westbound “WB”, etc.). As can be seen, the interval lengths are mostly consistent across the different spatial scopes, with exceptions where sample sizes were insufficient to consistently achieve the short intervals desired. For example, while 30-minute intervals were desired for the PM peak hours, they were only achievable consistently for Tsuen Wan Line northbound and Island to Tsuen Wan Line trips.

7.5.3 IRBT Results

The IRBT results are presented in Figures 7-6 and 7-7. The figures show the IRBT during the Base period, the IRBT during the Protest period, and the difference between the two, labeled “Change”. To emphasize the different interval lengths, the results are plotted as stair charts. Overall, the results indicate the protests had a significant impact on passengers’ experienced service reliability—for certain journeys and times of day. Significant IRBT increases occurred in the Island Line directional IRBTs and the two transfer direction

\footnote{10PM-12AM IRBT not calculated for Island Line westbound, because the frequent traveler group size was insufficient, even for a two hour interval}
Figure 7-6: Line-direction IRBT Results, Before and During Protests
Table 7.3: Time-of-Day Intervals for IRBT Calculation

<table>
<thead>
<tr>
<th></th>
<th>ISL WB</th>
<th>ISL EB</th>
<th>TWL NB</th>
<th>TWL SB</th>
<th>ISL to TWL</th>
<th>TWL to ISL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Peak (7-9 AM)</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
</tr>
<tr>
<td>AM Peak Shoulders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6-7 AM, 9-10 AM)</td>
<td>30 min</td>
<td>60 min</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
<td>30 min</td>
</tr>
<tr>
<td>Mid-day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10 AM-4 PM)</td>
<td>120 min</td>
<td>120 min</td>
<td>120 min</td>
<td>60 min</td>
<td>120 min</td>
<td>120 min</td>
</tr>
<tr>
<td>PM Peak (5-7 PM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM Peak Shoulders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4-5 PM, 7-8 PM)</td>
<td>60 min</td>
<td>60 min</td>
<td>30 min</td>
<td>60 min</td>
<td>30 min</td>
<td>60 min</td>
</tr>
<tr>
<td>Late Evening</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8 PM-12 AM)</td>
<td>120 min</td>
<td>120 min</td>
<td>60 min</td>
<td>120 min</td>
<td>60 min</td>
<td>120 min</td>
</tr>
</tbody>
</table>

Figure 7-7: Transfer Pattern IRBT Results, Before and During Protests

(a) Island Line to Tsuen Wan Line Transfer

(b) Tsuen Wan Line to Island Line Transfer
IRBTs—especially the Tsuen Wan to Island Line Transfer IRBT, which saw the IRBT more than double in the AM and PM peaks relative to the base period IRBT. On the other hand, the IRBT changed very little on the Tsuen Wan Line, or in the off-peak hours. The maximum Tsuen Wan Line IRBT increase was only 45 seconds, an amount unlikely to be noticed by most passengers.

These results suggest a strong relationship between passenger demand increases and increases in IRBT. The much greater IRBT increase on the Island Line matches that line’s greater demand increase, compared to the Tsuen Wan Line; evidently, the 11% increase in Tsuen Wan Line demand was insufficient to cause a sustained IRBT increase. Furthermore, the sharp time-of-day peaking seen in the IRBT response matches the peaking in the system-wide demand increase seen in Figure 7-3. This correlation between the IRBT response and demand change suggests the protests mainly affected service reliability through increased demand, consistent with the sensitivity of the IRBT to demand described in Section 5.3.3.

The IRBT response for transfer passengers is strikingly different between the two transfer directions. The increases in the Island Line to Tsuen Wan Line Transfer IRBT were of similar magnitude to the Island Line directional IRBTs’ increases. This result seems reasonable, given that the Tsuen Wan Line northbound, which from Figure 7-6c appears to have been relatively unaffected by the protests, is the only feasible second leg for Island to Tsuen Wan Line transfer trips. Thus, the observed change in Island Line to Tsuen Wan Line transfer trips’ reliability likely came mostly from increased delays on the Island Line leg of these passengers’ journeys.

Tsuen Wan Line to Island Line transfer passengers, in contrast, experienced a much greater IRBT increase than seen in either Island Line direction. This behavior can potentially be explained by a combination of two factors: First, the vast majority (87%) of passengers transferring from the Tsuen Wan Line to the Island Line did so at Admiralty, and boarded the Island Line eastbound for the second leg of their journey. Second, Admiralty is close to the Island Line eastbound’s peak load point. As a result, passenger demand (measured in line PPHPD) increased much more at this location than the Island Line average for both directions4 (shown in Table 7.4). Furthermore, this larger demand increase occurred at a higher baseline capacity utilization (65% in the AM Peak, 85% in

4It should be noted that the line-direction average passenger flow is different than the line-direction peak point flow shown in Figure 7-4. The former is the average line flow over all station-to-station links in the line-direction, while the latter is only for the single busiest link in the line-direction.
the PM Peak) than the Island Line eastbound or westbound averages, shown in Table 7.5. At such high capacity utilization, increases in demand are more likely to result in increased denied boardings and congestion.

The net effect was that the great majority of Tsuen Wan to Island Line transfer passengers boarding the Island Line eastbound at Admiralty were exposed to much greater increases in denied boarding and station congestion delay than the average intra-Island Line passenger. Thus, it is expected that such passengers experienced much greater increases in travel time variability, producing greater increases in IRBT. The larger PM IRBT increase, despite the smaller peak flow increase then, is possibly due to the higher capacity utilization in the PM peak compared to the AM peak, which would be expected to magnify the effect of a given demand increase on denied boardings and station congestion.

Table 7.4: Peak Half-Hour Average Passenger Flow (PPHPD)

<table>
<thead>
<tr>
<th></th>
<th>Base Period</th>
<th>Protest Period</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Peak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISL EB, Admiralty</td>
<td>40,892</td>
<td>49,276</td>
<td>8,384</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>28,512</td>
<td>32,682</td>
<td>4,170</td>
</tr>
<tr>
<td>ISL EB, average all links</td>
<td>15,908</td>
<td>18,762</td>
<td>2,854</td>
</tr>
<tr>
<td>PM Peak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISL EB, Admiralty</td>
<td>48,478</td>
<td>53,350</td>
<td>4,872</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>17,516</td>
<td>20,654</td>
<td>3,138</td>
</tr>
<tr>
<td>ISL EB, average all links</td>
<td>26,178</td>
<td>29,272</td>
<td>3,094</td>
</tr>
</tbody>
</table>

Table 7.5: Peak Half-Hour Capacity Utilization

<table>
<thead>
<tr>
<th></th>
<th>Base Period</th>
<th>Protest Period</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM Peak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISL EB, Admiralty</td>
<td>65%</td>
<td>78%</td>
<td>13%</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>45%</td>
<td>52%</td>
<td>7%</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>25%</td>
<td>30%</td>
<td>5%</td>
</tr>
<tr>
<td>PM Peak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISL EB, Admiralty</td>
<td>85%</td>
<td>93%</td>
<td>9%</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>31%</td>
<td>36%</td>
<td>5%</td>
</tr>
<tr>
<td>ISL WB, average all links</td>
<td>46%</td>
<td>51%</td>
<td>5%</td>
</tr>
</tbody>
</table>

7.5.4 Incident Results and Causal Analysis

The Number of Incidents ≥X minutes results are presented for the Tsuen Wan and Island Lines in Figure 7-8. These charts break down the number of incidents during the Base period and Protest period into three categories: 2-5 minute incidents, 5-8 minute incidents, and ≥8 minute incidents. The first two categories can be derived with simple arithmetic
from the raw results, e.g. 2-5 minute incidents = ≥2 minute incidents - ≥5 minute incidents.
The results are grouped into five time-of-day periods: Early Morning (6-7 AM), AM Peak (7-10 AM), Mid-Day (10 AM-5 PM), PM Peak (5-8 PM), Late Evening (8 PM-12 AM).

A key feature of the incident results is the distinctly different trends seen in the peak and off-peak periods. During the off-peak, there is no consistent increase associated with the Protest period. For some lines and times of day, total incidents increase (e.g., Island Line Mid-Day), but for others, the number of incidents decreases significantly (e.g., Island Line Late Evening). Furthermore, the magnitude of these changes is relatively small, no more than ±4 incidents. During the peak periods, on the other hand, far larger changes in incident frequency are seen, with the number of Tsuen Wan Line AM Peak incidents increasing by 12 from the Base to the Protest periods.

Within the peak period results, another important distinction can be made, between the 2-5 minute incident trends and ≥5 minute incident trends. While 5-8 minute and ≥8 minute incidents show consistent increases from the Base period to the Protest period, 2-5 minute incidents show both significant increases and significant decreases. Furthermore, the magnitude of the ≥5 minute incident increases is consistently greater on the Island Line than the Tsuen Wan Line, while the reverse is true for 2-5 minute incidents.

These results suggest the protests had a much greater impact on the frequency of large incidents than the frequency of small incidents. Taken in isolation, the trends in ≥5 minute incidents follow the basic trends in passenger demand seen in Section 7.5.1: larger
in the peaks than the off-peak, and much larger response on the Island Line than the Tsuen Wan Line. This result is not surprising; as passenger demand increases, there is more likelihood that passenger-related incidents—sick passenger, bag caught in door, etc.—will occur over a given timeframe. It seems very likely, then, that a significant portion of the IRBT response seen in Figures 7-6 and 7-7 came as a result of the increases in \( \geq 5 \) minute incidents seen in Figure 7-8.

Potential Further Analysis of Hong Kong Protests

The results presented in this section represent only the preliminary level of analysis of the Fall 2014 protests’ affects on MTR’s performance. Fuller understanding of the protests’ impact, and the mechanisms by which they impacted passengers’ journeys, will require more in-depth analysis, beyond the scope of thesis. For example, it is not clear what portion of the increase in passenger-experienced unreliability (as captured by the IRBT) is directly related to the increase in incidents, and what portion is due to the general effects of increased demand (e.g., more frequent denied boardings). Furthermore, more analysis could be done to see what effect the protests had on passengers making multiple transfers, or on passengers boarding in certain geographic areas.

The IRBT, and the analysis process described in this chapter, has the potential to help answer these questions. For example, to determine the portion of unreliability change not caused by incidents, one could calculate the IRBT with “incident-affected” trips removed, defining incident-affected trips as those beginning within a certain time window (e.g., a half-hour) of a major incident on the line of interest. Meanwhile, more in-depth geographic analysis could be done with more “creative” spatial scopes for IRBT calculation—for example, all OD pairs from Kowloon to western Hong Kong Island. The temporally and spatially flexible nature of the IRBT makes the metric a potentially powerful tool for investigating these types of deep analytical questions.
Chapter 8

Conclusions

This chapter summarizes the work presented in this thesis. Section 8.1 summarizes the research work, including the main findings from each research area. Section 8.2 gives recommendations, based on the research findings, on how to effectively develop and use passenger-experienced reliability metrics for a public transport system. Section 8.3 lays out potential opportunities for future research on passenger-experienced reliability metrics.

8.1 Research Summary and Findings

This research aimed to help public transport operators more easily and effectively incorporate measurement of the passenger’s experience of reliability into their operations, communication, and planning. Towards this end, this research has developed new measures of passenger-experienced reliability, as well as a guide to implementing these new metrics.

This research also aimed to improve the conceptual understanding of passenger-experienced reliability, and the potential benefits for the public transport operator of measuring it. This work can potentially aid the building of political and management support for the investment required to implement passenger-experienced reliability measures—likely a significant challenge, within many public transport organizations. To further these objectives, the following has been achieved:

Model of Passenger’s Experience of Reliability

Chapter 2 reviewed the literature, from academia and practice, on the transit passenger’s experience of reliability. From this review, the following model of passenger-experienced
reliability was developed for high-frequency (i.e., “walk-up”) services:

Passengers generally perceive reliability as travel time variability, relative to their expected “typical” travel time, rather than schedule adherence. Passengers will react to this perceived reliability by budgeting additional “buffer time”, on top of their typical travel time, to achieve an acceptable probability of on-time arrival. The magnitude of this buffer time, then, is directly related to the passenger’s reliability experience: if unreliability increases, passengers will budget extra buffer time to compensate. This buffer time accounts not only for train delays (long headways, incidents, etc.), but also passenger demand-related delays such as denied boardings and congestion in stations.

Applications of Passenger-Experienced Reliability Measurement

Given the differences between passenger-experienced and operational reliability identified in Chapter 2, Chapter 3 identified the following public transport operator functions that could be improved through the use of passenger-experienced reliability measurement:

- **Reliability Management**: Operators’ efforts to monitor and maintain service reliability can be made more responsive to passengers’ needs through passenger-experienced reliability measurement. Specifically, such measurement would allow an operator to identify, diagnose, and take action to fix reliability problems not detectable with operational reliability metrics, such as passenger demand-related delays.

- **Passenger Information**: Reliability information provided to passengers in a manner reflective of their experience can potentially help them more effectively plan their journeys—for example, through an online journey planner.

- **Communication with Stakeholders**: Passenger-experienced reliability measures should allow an operator to more clearly and effectively communicate its reliability performance to outside stakeholders. Similarly, such measures can help stakeholders more effectively articulate their concerns to the transit operator.

- **Cost-Benefit Analysis**: Passenger costs and benefits regarding reliability in investment decisions can be better assessed using passenger-experienced reliability metrics.

- **Long-Range Planning**: Passenger-experienced reliability metrics can support long-range planning studies, such as equity analysis, regional accessibility analysis, and
Design Objectives for Passenger-Experienced Reliability Metrics

Based on the results from Chapters 2 and 3, a set of design objectives for passenger-experienced reliability metrics was derived in Chapter 4:

- Metrics should be representative of the passenger’s experience, reflecting the model developed in Chapter 2. To achieve this, metrics should, in particular: distinguish between service variability and schedule adherence, include all sources of unreliability, be calculated at the OD-pair level, and control for variation in passenger behavior (e.g., slow vs. fast walking speed).

- Metrics should be meaningful for passengers and stakeholders—the metrics should ideally be intuitively understandable to passengers and non-experts, objective, relate to passengers’ reliability experience, and be useful for planning journeys.

- Metrics should be easily comparable across services and times of day—different lines, peak vs. off-peak, etc. Towards this end, the metric should be independent of the service schedule, and be an absolute, not relative, measure of reliability.

- Metrics should be flexible in the time periods they are calculated for. First, they should be calculable for relatively short portions of the service day (i.e., an hour or less) to capture time of day variation in reliability. Second, they should be calculable for flexible date periods (e.g., weekdays in April and June 2014), to easily separate types of service days (weekday, weekend), or control for specific events.

- Metrics should also be flexible in service scope, calculable at more than just the line-level; they should be calculable at the OD pair level, line-direction level, transfer pattern level, geographic area level, etc.

Development of the IRBT and PPRBT

In the first parts of Chapters 5 and 6, two new metrics were developed to measure the passenger’s experience of reliability, termed the Individual-based Reliability Buffer Time (IRBT) and the Platform-to-Platform Reliability Buffer Time (PPRBT), respectively. These two metrics are conceptually similar; both aim to estimate, for a given time period and service,
the typical passenger’s 95th percentile buffer time, the difference between their 95th and 50th percentile travel times. This represents the typical buffer time needed for a 1-in-20 probability of late arrival. The main methodological contribution of these metrics are their different ways of calculating the buffer time estimate, described below.

The IRBT is calculated from passengers’ travel times obtained from AFC data, for a system with entry and exit transactions, as follows: First, for a given OD pair and time period, travelers taking more than 20 trips are identified. Then, each frequent traveler’s 95th and 50th percentile travel times are calculated. The difference between the two is then set as their “individual buffer time”. Finally, the IRBT is calculated as the median “individual buffer time”, over all frequent travelers. This method addresses an issue with a previous buffer time metric, the AFC-based RBT developed by Chan [4] and Uniman [2]. That metric, by calculating a single travel time distribution over all passengers, did not distinguish between travel time variability due to service unreliability, and travel time variation between customers with different behavior. The IRBT, by calculating travel time variance at the individual level, removes this upward bias from cross-passenger variation.

The PPRBT estimates the typical travel time distribution for a given OD pair and time period by simulating the journeys of a large number of “virtual passengers”, and finding the difference between the 95th and 50th percentiles of the output travel time distribution. The virtual passengers board and alight trains according to the actual train arrival and departure times recorded in AVL data. Thus, they are subject to the vehicle delays and capacity restrictions experienced by actual passengers. For transfer OD pairs, this process is repeated for each leg of the most common route choice. The PPRBT does not include platform access and egress time variability—hence its “Platform-to-Platform” designation.

A major component of this simulation is a method of estimated denied boarding delays. For each arriving train, the available passenger capacity and number of passengers waiting to board are estimated; the difference is the number of denied boardings. This process is modeled as a first-in, first-out queueing process. Passenger demand is estimated from input OD matrices, while available capacity is estimated from train load estimates and assumed vehicle capacity. This improves upon an earlier AVL-based buffer time metric, the AVL-based RBT (developed by Ehrlich [5]), which did not model denied boardings.

Both the IRBT and PPRBT can be aggregated across many OD pairs by taking the weighted average with respect to the passenger flow over each OD pair. This method can
be used to calculate, for example, a line-level IRBT or line-level PPRBT.

**Testing and Evaluation of the IRBT and PPRBT**

In the second parts of Chapters 5 and 6, the IRBT and PPRBT are demonstrated using data from MTR, and then evaluated according to the design objectives developed in Chapter 4. The IRBT was compared to corresponding AFC-based RBT results; the IRBT was consistently much lower, indicating a significant amount of passenger behavior “noise” was removed by the IRBT calculation process. The IRBT’s response was also found to be strongly related to incidents, passenger demand, and scheduled headway, as expected and desired. A potentially significant limitation of the IRBT was also discovered: the need to achieve a sufficient number of frequent travelers severely restricts the use of the IRBT at the individual OD pair level, thus precluding its use for journey planner applications, and places some restrictions on time periods the aggregate IRBT is calculable for, especially in the off-peak. However, despite these concerns, the IRBT was determined to be a good general-purpose passenger-experienced reliability metric.

The PPRBT was assessed in two ways. First, the PPRBT’s response was tested with respect to headway, demand, and capacity. The results were found to be consistent with the model’s assumptions, and intuitive expectations. When passenger demand is high, decreasing the train capacity parameter, or increasing passenger demand further, causes the PPRBT to increase, due to more frequent denied boarding predictions. Meanwhile, when demand is low (i.e., small likelihood of denied boardings), the PPRBT is positively correlated with scheduled headway, as expected.

Second, and probably more importantly, PPRBT results were compared to corresponding IRBT results, allowing the simulation output to be compared against actual passengers’ reliability experiences. The results were mixed. In periods of low demand, with no denied boardings, the PPRBT and IRBT results were generally very close. However, during periods of high demand, the PPRBT was a poor estimator of the IRBT. This is a serious problem, because the peak hours are generally the times when passenger-experienced reliability measurement is most needed, and when the IRBT is most representative of the passenger’s reliability experience. Due to this issue, the PPRBT was not recommended for general use as a reliability metric.
Implementation of the IRBT

An implementation guide for the IRBT was developed in the first part of Chapter 7, proposing strategies for making the important decisions pertaining to setting up the IRBT for a specific application. These decisions fall into following four main categories:

- **Integration with Other Metrics:** Should the IRBT be used alone, or with one or more of the operator’s currently-used metrics? If so, which one(s)? These questions can be answered by the following process: (1) Assess the application’s metric design needs. (2) Assess the design characteristics of the current metrics, and compare them to the IRBT’s. (3) Find the set of metric(s), among the IRBT and current metrics, whose design attributes best match the application’s design needs.

- **Service Scope Setting:** What level of spatial aggregation should be used for IRBT calculation? (i.e., line level, line-direction level, etc.) Key considerations for this include specific reporting needs, ease of interpretation, frequent traveler group size constraints (larger spatial scopes achieve greater group sizes), and the service scope of reliability issues.

- **Calculation Schedule:** Does the IRBT need to be calculated on a regular basis, or for a set of specific periods? If regularly, how frequently should the IRBT be calculated? These considerations depend largely on the general type of application.

- **Time Period Setting:** For what time-of-day intervals and date periods should the IRBT be calculated? Considerations for this include the timeframe and trends in reliability factors of interest, the target audience, calculation schedule, and frequent traveler group size constraints.

**Implementation Case Study: Investigating Effects of Hong Kong Protests**

The implementation of the IRBT was demonstrated in Chapter 7 through a case study application. The IRBT was applied, using the steps outlined in the implementation guide, to assess the effects of the Fall 2014 political demonstrations in Hong Kong on MTR’s service reliability. The IRBT results demonstrated that a significant demand surge on two MTR lines produced a significant deterioration in passenger-experienced reliability on those lines.
8.2 Recommendations

This section summarizes this research’s main recommendations for public transport operators on passenger-experienced reliability measurement.

Need for Passenger-Experienced Reliability Measures

Investment in measuring the passenger’s experience of reliability is justified for most operators, because such measurement can improve the operator’s responsiveness to passenger needs in a number of ways, including:

- More efficient detection of reliability problems related to passenger demand and transfers, allowing action to be taken more quickly to address them.
- Ability to evaluate efficacy of reliability improvements from passenger’s perspective.
- Better communication of reliability performance to stakeholders.
- Provide reliability information to passengers useful for planning journeys.

Effective measurement of the passenger’s reliability experience can only be done using specifically passenger-focused metrics, as traditional “operational” metrics do not effectively capture the passenger’s experience. The most important requirements for such metrics are:

- Measure travel time variability, not schedule adherence
- Capture all parts of the passenger’s journey (access/egress time, waiting time, in-vehicle time, etc.)
- Capture all sources of delay (train delay, long headways, denied boardings, etc.)
- Be measured at the OD-level, then aggregated to higher spatial levels if needed

Potential Applications for IRBT, PPRBT, AFC-based RBT, AVL-based RBT

Based on the description and evaluation of the IRBT, PPRBT, and RBT (both AFC-based and AVL-based) presented in this thesis, conclusions can be made on the relative strengths and limitations of these four metrics—two new, two developed in prior research—for practical application. These are summarized in Table 8.1.
<table>
<thead>
<tr>
<th></th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRBT</td>
<td>• Captures all sources of unreliability and journey components</td>
<td>• Frequent traveler group size minimums severely limit calculation for OD pairs, especially in off-peak</td>
</tr>
<tr>
<td></td>
<td>• Not influenced by passenger behavior variation—represents operator’s performance</td>
<td>• Time of day, interval length, date period length highly restricted</td>
</tr>
<tr>
<td></td>
<td>• “Absolute” measure of unreliability, in Excess IRBT form</td>
<td>• Needs entry and exit AFC data</td>
</tr>
<tr>
<td>PPRBT</td>
<td>• Not influenced by passenger behavior variation—represents operator’s performance</td>
<td>• Does not include access/egress time</td>
</tr>
<tr>
<td></td>
<td>• Calculable for any time period</td>
<td>• Does not model denied boardings effectively</td>
</tr>
<tr>
<td></td>
<td>• Does not require exit AFC data</td>
<td>• Complicated to set up, calculate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Requires route choice, train load data</td>
</tr>
<tr>
<td>AFC-based RBT</td>
<td>• Captures all sources of unreliability, journey components</td>
<td>• Cross-passenger variation biases results. Results do not reflect operator’s performance.</td>
</tr>
<tr>
<td></td>
<td>• Can be expressed in terms of absolute travel time range. Easy to understand.</td>
<td>• Needs entry and exit AFC data</td>
</tr>
<tr>
<td></td>
<td>• Calculable for any time period</td>
<td></td>
</tr>
<tr>
<td>AVL-based RBT</td>
<td>• Not influenced by passenger behavior variation—represents operator’s performance</td>
<td>• Does not capture demand-related delays</td>
</tr>
<tr>
<td></td>
<td>• Only requires AVL data, OD matrix</td>
<td>• Does not include access/egress time</td>
</tr>
<tr>
<td></td>
<td>• Not computationally difficult to calculate</td>
<td>• Cannot be calculated for transfer OD pairs</td>
</tr>
</tbody>
</table>

Based on these metric strengths and limitations, some tentative recommendations can be made regarding the use of these four passenger-experienced reliability metrics for the four reliability measurement applications discussed in Chapter 3. These are presented in Table 8.2. In general, the IRBT is suitable for most applications, the PPRBT could become suitable with an improved denied boarding delay model, and both RBTs have limited applicability. An exception is for passenger information, where the AFC-based RBT is likely the best metric, due to its ability to be presented as an absolute travel time range as in
Table 8.2: Metric Application Recommendations

<table>
<thead>
<tr>
<th>Metric Implementation Area</th>
<th>IRBT</th>
<th>PPRBT</th>
<th>AFC-based RBT</th>
<th>AVL-based RBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability Management Analysis</td>
<td>Good for analyzing all sources of unreliability, including demand-related. Not best for analyzing single incident effects.</td>
<td>Good for operational reliability analysis. If denied boarding model is improved, could be good for all sources of unreliability.</td>
<td>Not recommended, because potentially biased by exogenous cross-passenger variation</td>
<td>Good for operational reliability analysis (running times and headways)</td>
</tr>
<tr>
<td>Passenger Information</td>
<td>Not recommended, because it cannot reliably be calculated at OD-pair level, especially in off-peak.</td>
<td>Not recommended, because it leaves out access and egress times, and has problems estimated denied boarding delays</td>
<td>Can be effective if given as absolute travel time range in journey planner</td>
<td>Not recommended, because it leaves out denied boarding delays and access/egress times</td>
</tr>
<tr>
<td>Reporting for Stakeholders</td>
<td>Recommended for line-level and line direction-level reporting, as it captures all sources of unreliability</td>
<td>Could be effective if denied boarding model improved</td>
<td>Not recommended, because it is biased by cross-passenger variation, and thus does not reflect operator’s performance</td>
<td>Can be used for reporting operational reliability, but not general passenger-experienced reliability</td>
</tr>
<tr>
<td>Cost-Benefit Analysis</td>
<td>Recommended, because it captures all sources of unreliability</td>
<td>Could be effective if denied boarding model improved</td>
<td>Not recommended, because it is biased by cross-passenger variation, and thus does not reflect operator’s performance</td>
<td>Not recommended, because it doesn’t include denied boarding delays</td>
</tr>
<tr>
<td>Long-Range and Regional Planning</td>
<td>Good for analyzing all sources of unreliability, including demand-related.</td>
<td>Good for operational reliability analysis. If denied boarding model is improved, could be good for all sources of unreliability.</td>
<td>Not recommended, because potentially biased by exogenous cross-passenger variation</td>
<td>Good for operational reliability analysis (running times and headways)</td>
</tr>
</tbody>
</table>

**Metric Implementation**

The implementation of passenger-experienced reliability metrics such as the IRBT should follow a clear, organized process that addresses four key implementation decisions:

- How should the new metric be integrated with an operator’s existing metrics?
- What spatial scope (lines, OD pairs, etc.) should be used for calculation?
• How frequently should the metric be calculated?

• What time of day intervals and date periods should be used for calculation?

This process should consider the particular needs for each reliability measurement application, including the sources of unreliability of most interest, the metrics’ target audience, the frequency of updates desired, and frequent traveler group size limitations. It must also consider the design characteristics of both the IRBT (or other new metric) and the operator’s current reliability metrics.

Effective integration with existing metrics is particularly important for successful implementation of passenger-experienced reliability metrics. Such metrics should not be considered a replacement for traditional operational reliability measures (e.g., OTP), but rather a complement to them, offering a different perspective on reliability. For some applications, knowing the passenger’s reliability experience is most valuable, while for others, operational reliability measurement is better. And in some situations the best results can be achieved using both types of metrics, where comparing the metrics’ results can provide insight unobtainable from either type of metric in isolation. Passenger-experienced reliability metrics can also be used effectively with non-reliability measures of service quality, such as average crowding or average headway, to provide a more complete assessment of the passenger’s experience of transit service.

8.3 Future Research

This work suggests a number of directions for future research which could further improve the methods available for measuring public transport passengers’ reliability experience, and facilitate their practical implementation by public transport operators.

Improved PPRBT Algorithm for Passenger Delays

The largest limitation to the practical use of the PPRBT is its inability to consistently assess the delays associated with high passenger demand. If this issue could be resolved through an improved algorithm, the PPRBT could be an effective tool for assessing reliability of large public transport systems without exit AFC data.

A good starting point would be to develop a more sophisticated model of denied boardings, and then calibrate the vehicle capacity parameter(s) using in-person observations of
denied boardings at stations. This may entail station-specific capacity values, to account for factors like uneven train loading and uneven platform queueing. It could also involve modeling vehicle capacity as a function of passenger demand, rather than a static value as done in this thesis.

A related issue to be resolved is the passenger delay model’s reliance on detailed route choice data, OD matrices, and train load estimates. This data is available from MTR, but is probably not for most other large public transport systems, especially those without exit AFC data. Route choice is particularly important, as many large metro systems have significantly more path choice than MTR’s largely radial system (e.g., London, Paris).

Another issue meriting further research is whether access and egress time variation needs to be included into the PPRBT. It is known that access and egress times increase when passenger congestion increases, but it is not known whether this adds significantly to overall travel time variation. Empirical study needs to be done to determine whether this effect is significant enough to necessitate inclusion in the PPRBT travel time estimation algorithm.

**IRBT and PPRBT in Combination**

It is worth investigating potential procedures for the combined use of the IRBT and PPRBT for passenger-experienced reliability measurement. In terms of their accuracy with respect to passenger demand, these two metrics’ strengths are complementary: the IRBT is most effective when passenger demand is high (because large frequent traveler group sizes are obtained), while the PPRBT is most accurate when passenger demand is relatively low (when no denied boardings occur). Thus, by switching between the two, one could theoretically have good passenger reliability estimates throughout the entire service day, in both the peak and off-peak. However, some practical implementation questions have to be answered; for example, what criteria should decide when to use each metric?

**Frequent Traveler Group Size Standard**

The method of determining the minimum frequent traveler group size standards for the IRBT in Section 5.4.1 was based on fairly arbitrary correlation and error thresholds for OD pair-level and line-level IRBT results for different-sized random samples of frequent travelers. This method is not as statistically rigorous as may be desired. More research is merited to determine a more statistically sound way of setting the group sample size.
standards, that is less subject to subjective parameters. It is also worth investigating whether minimum frequent traveler group sizes are universal, or vary across systems, across lines within a system, and over different times of day.

**Regression Model of IRBT**

It was shown in Section 5.3 that the IRBT is sensitive to incidents, scheduled headway, and passenger demand. A next step would be to estimate a regression model for the IRBT with respect to these three factors. (Incidents could be represented a number of ways—number of incidents, number of incidents ≥X minutes, total delay, etc.) If the model was found to explain a large majority of IRBT variation, it could be possible to use a “model IRBT” as a reliability metric in place of the actual IRBT. Such a “model IRBT” would have less intensive data requirements than the actual IRBT (incident records and AVL data are much smaller than bulk AFC data), and would not be bound by minimum frequent traveler group size constraints, making it an easier-to-use metric than the actual IRBT.
Bibliography


