Quantifying Impacts of Transit Reliability on User Costs

Jeffrey M. Casello, Akram Nour, and Bruce Hellinga

Transportation modeling frameworks assume that travelers are economically rational; that is, they choose the lowest-cost alternative to complete a desired trip. The reliability of travel time is of critical importance to travelers. The ability to quantify reliability allows planners to estimate more accurately how system performance influences local travel behavior and to evaluate more appropriately potential investments in the transportation system infrastructure. This paper presents a methodology that makes use of automatic vehicle location data from the regional municipality of Waterloo, Ontario, Canada, to estimate the reliability of transit service. On the basis of these data, the impacts of unreliable service on generalized transit user costs are quantified by use of a simulation model of bus arrivals and passengers' desired arrival times. It is shown that the increasing reliability of arrivals at a station can decrease transit users' generalized costs significantly and by as much as 15% in a reasonably reliable network. It is further posited that the inclusion of uncertainty in the calculation of generalized costs may provide better estimates of mode splits in travel forecasting models. A description of future applications of the model concludes the paper.

From a user's perspective, reliability in the transit network involves departing from the origin station on time, having reasonable limits on in-vehicle time, and most importantly, arriving at the destination station within a time frame that allows the traveler to reach his or her final destination without being late. The ability to quantify the degree of unreliability of a service allows transit planners to better estimate mode splits by the use of travel forecasting models. Furthermore, quantitative assessments of reliability provide estimates of tangible user benefits (through travel savings) that can be compared with the costs of investment in the infrastructure to upgrade reliability, such as queue jumpers, transit signal priority, or other means.

This paper presents a methodology that makes use of automatic vehicle location (AVL) data to estimate transit reliability at all stops along an express bus line service. These data were then used with various assumptions of passenger behavior to estimate the impacts of unreliable service on generalized transit user costs through a simulation model. The results obtained with the model suggest that contemporary planning techniques may underestimate the generalized cost of transit on unreliable networks. An estimate of user benefits as a result of improved reliability in the network is also provided.

J. M. Casello, School of Planning and Department of Civil and Environmental Engineering, and A. Nour and B. Hellinga, Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. Corresponding author: J. M. Casello, jcasello@fes.uwaterloo.ca.

Transportation Research Record: Journal of the Transportation Research Board, No. 2112, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 136–141.
DOI: 10.3141/2112-17

LITERATURE REVIEW

The issue of reliability in transit networks has been modeled for some time. Some of the earliest work was done by Osuna and Newell (*1*) as well as Wilson et al. (*2*). With the introduction of the AVL technology, the opportunity to capture vast quantities of reliability data arose (*3*). Maximizing the benefit of this data collection is a major effort for many transit agencies.

The work presented here draws heavily from the formulation presented by Furth and Muller (*4*). Those investigators quantified the expected and the excess waiting times as a stochastic function of possible headways. They assumed that passengers choose lines independently of headways and that passenger arrivals were not dependent on headways (i.e., uniform arrivals). They quantified the waiting times on the basis of extreme cases of reliability, for example, the 95th percentile wait time. They suggest, and the authors of the present report concur, that the mean waiting time is a poor indicator of waiting time penalties in an unreliable network.

A similar approach is taken here, but the concentration is on arrival times at the traveler's destination and the likelihood that an arrival will satisfy the traveler's trip objectives. Furthermore, the analysis is extended to include quantification of the generalized cost by using a linear weighting proposed by Kittelson and Associates et al. (*5*) and used in most travel forecasting models. In the generalized cost model used in the present study, the impacts of early arrivals, late arrivals, and departure time are explicitly treated and shifted as required by unreliability. For the impacts of late and early arrivals, the model is based on the work of Small (*6*), as used by Bates et al. (*7*).

METHODOLOGY

The quality of service experienced at a given station is defined first. The regional municipality of Waterloo, Ontario, Canada, operates the iXpress service, which is a limited-stop, express service that travels between Waterloo, Kitchener, and Cambridge. The alignment, shown in Figure 1, is approximately 33 km in length and consists of 13 stops. Along the route there are four downtowns (Cambridge has two downtowns), two universities, office complexes, major hospitals, and regional shopping centers. iXpress operates throughout the day with 15-min headways.

Each of the iXpress vehicles is equipped with AVL technology. Real-time arrival information was collected for every stop during the morning and the afternoon peak periods for a week. In total, approximately 95 observations were gathered at each station. From these, service reliability was defined as the difference between the actual arrival time (AAT) and the scheduled arrival time (SAT). For each station, histograms of service reliability, shown in Figure 2, were

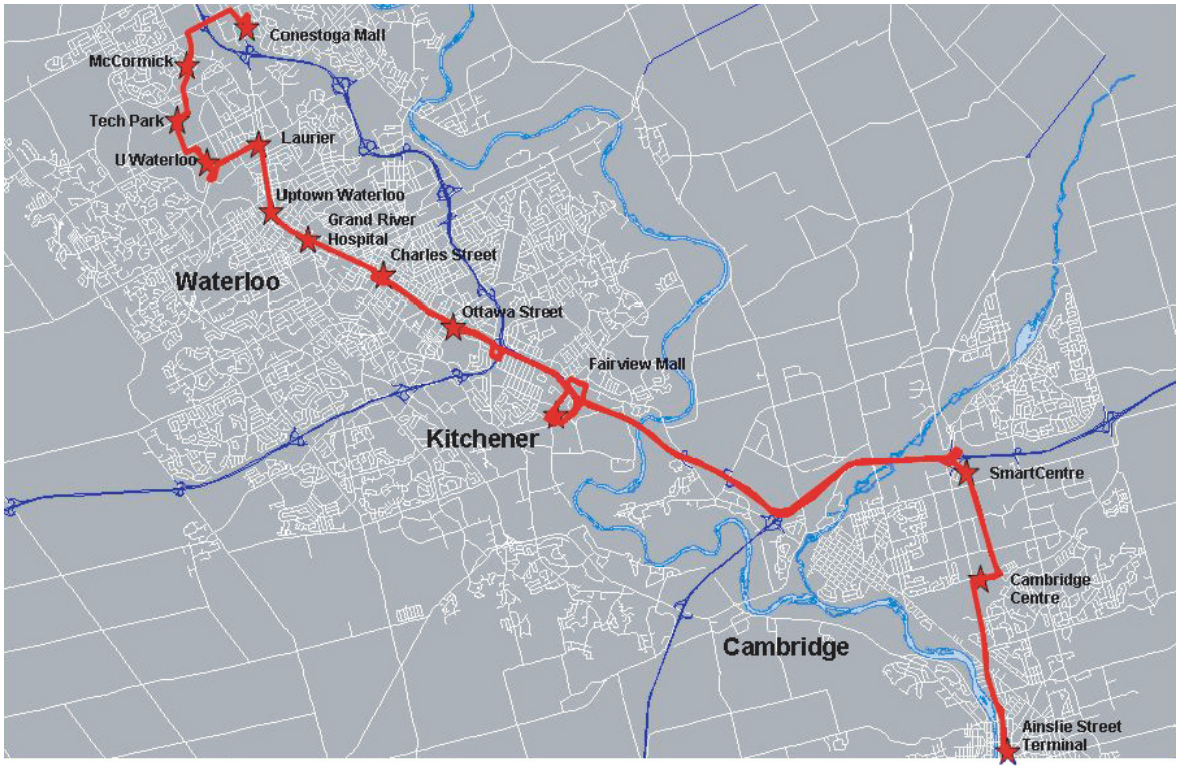

FIGURE 1 iXpress route serving Waterloo, Kitchener, and Cambridge.

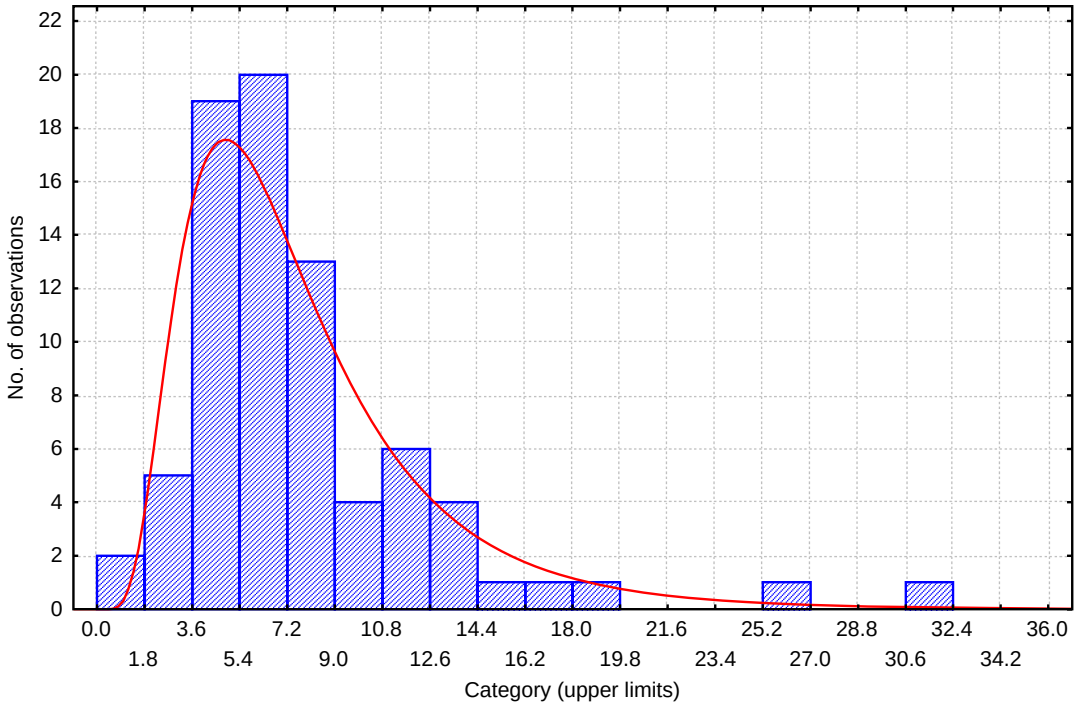

FIGURE 2 Observed frequency of service reliability, Ainslee terminal to Cambridge Centre (distribution: log normal; chi-square test = 4.92718; degrees of freedom = 4 (adjusted); p = .29485).

generated, and on the basis of those histograms, goodness-of-fit tests were completed to suggest an appropriate distribution. In each case, the probability density function (PDF) is a log-normal distribution that meets a priori expectations: there are a few early arrivals, many arrivals at about the scheduled time, and a longer range of arrivals at times much later than the scheduled time. The log-normal distribution requires that all observations be greater than zero. To accommodate early arrivals, a constant approximately equal to the minimum observation was added to the test statistic.

For the 13 stations, a wide range of standard deviations (a measure of variance in service reliability) was observed. The range of standard deviations was from 0.12 to 1.06 min, and the average deviation was 0.44 min. In all, iXpress is a relatively reliable system. From these empirical data, three station types were defined: high reliability, medium reliability, and low reliability. Each type has a similar mean but various deviations.

Three groups of travelers with various risk tolerances that may represent the travelers' personalities or trip purpose were next defined. One subset of travelers is very risk averse (RA) and chooses a transit departure only if the likelihood of arriving late with that departure was less than 10%. This group may represent those commuters for whom work start times are fixed and highly inflexible. A second group of travelers is moderately risk averse (MRA) and selects a transit departure if the risk of arriving at the destination late is less than 30%. Finally, a risk-neutral (RN) group was defined. These individuals select a transit departure if the probability of a late arrival is less than 50%. The RN group may be considered recreational travelers for whom arrival times have some flexibility.

With these definitions in place, travel behavior rules were created for each of the travelers to each of the stations. Assume that the necessary arrival time (NAT), which is the latest time that a traveler can arrive without being late, is a random variable that is uniformly distributed between two subsequent bus arrivals. Δ* can be defined as the difference of the bus's SAT and AAT. On the basis of the cumulative distribution function of the service reliability statistic and the traveler, there exists some Δ* for which the probability of an arrival before SAT + Δ* is equal to the traveler's risk threshold. If the traveler's NAT is later than the station's Δ*, then that traveler will choose the first transit arrival before his or her NAT. The relationship is shown graphically in Figure 3. This relationship can be expressed mathematically as follows:

Travelers choose A_1 if

$$\text{NAT} \geq \Delta^* \tag{1}$$

where Δ* is given by Pr(AAT – SAT ≥ Δ*) ≤ risk tolerance (where Pr is probability), and A_1 is the bus arrival with the first SAT before the NAT.

Suppose a traveler must reach his or her destination stop at 14:00 h. The closest scheduled bus arrival time is 13:55 h. If arrivals at that stop are sufficiently reliable that the bus scheduled to arrive at 13:55 h actually arrives before 14:00 h 90% of the time, then even the most risk-averse traveler in the model will choose that departure time. If, alternatively, the bus scheduled to arrive at 13:55 h actually arrives before 14:00 h only 40% of the time, then none of the travelers has a sufficient risk tolerance to choose the bus scheduled to arrive at 13:55 h.

Because the PDF of arrival times at each of the stations is available, the cumulative distribution function can be computed, and from that the number of minutes after the SAT that satisfies the risk aversion threshold for each of the travelers in the model can be determined. This is shown in Table 1.

One can interpret Table 1 as follows. A RA traveler traveling to a station with known, low-reliability service will choose the scheduled bus arrival immediately before his or her appointment only if the appointment occurs later than 7 min after the SAT. This is because the data on actual arrivals at this station suggest that there is a chance of only 10%, the RA traveler's threshold, of an arrival later than 7 min after the SAT.

Because it was assumed that the NATs are uniformly distributed, the probability that NAT will occur after Δ* can also be estimated. Mathematically, this is given by

$$\Pr\left(\text{NAT} \geq \Delta^*\right) = 1 - \frac{\Delta^*}{h} \tag{2}$$

where h is the line's headway, which in this case is 15 min.

An appropriate question to be asked here is what happens to those trips for which the NAT falls too near the SAT to allow the traveler to choose that arrival. It was assumed that the traveler then elects to travel on an earlier bus. The traveler then arrives at the destination one headway earlier than the original scheduled arrival time plus any unreliability that the traveler may experience on this bus. Mathematically, the SAT on the previous bus, SAT_{-1}, becomes

$$\text{SAT}_0 - \text{headway} + \text{unreliability}$$

The impacts of this unreliability are explored in the next section.

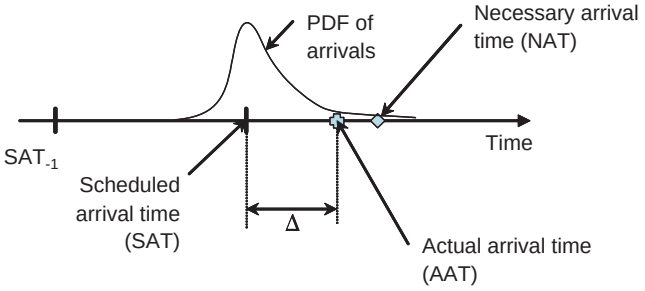

FIGURE 3 Graphical representation of bus and passenger arrival times.

TABLE 1 Station Arrival Distributions and Critical Arrival Times

			Δ* for Each Station and Traveler Type		
Station Type	μ	σ	RA	Moderately RA	RN
High reliability	1.245	0.2621	1.9	1.0	0.5
Medium reliability	1.228	0.4416	4.0	2.3	1.4
Low reliability	1.062	0.7792	7.0	3.4	1.9

Perceptions of Travel Time

Typically, when travel time is measured, modelers use a generalized cost formulation that quantifies a linearly weighted sum of travel time components. A common example is as follows:

$$GC_T = (\alpha_0 AT + \alpha_1 WT + \alpha_2 IVT)VOT + fare \quad (3)$$

where

GC_T = generalized cost of a trip by transit (dollars),
AT = access time to the line (min),
WT = waiting time, modeled as half the headway for short headways (min),
IVT = in-vehicle time (min),
VOT = value of time (dollars per minute),
fare = transit fare (dollars), and
α_i = relative importance of the component.

Although this formulation adequately measures the actual time and out-of-pocket costs from the time of departure from an origin (other than the transit stop) to the time of departure at the transit stop, it fails to account for two additional costs borne by transit travelers. First, because transit has discrete departure and arrival times, there is an inherent penalty for early arrival that is not typically counted. Second, in light of reliability, travelers may experience a penalty for a late arrival or may make travel choices to avoid being late (as described above) and therefore incur greater early arrival penalties.

Now, returning to the example, if a traveler's NAT is after Δ*, then the traveler will choose the SAT nearest his or her NAT. The bus's AAT, however, is stochastic, which means that the traveler may arrive very early (relative to the NAT) or may arrive after the NAT. If a traveler's NAT is before Δ*, then the traveler chooses an earlier bus to minimize the potential for being late. In so doing, the traveler increases the cost associated with leaving earlier and arriving well before the NAT. This range of possibilities carries with it an inconvenience and, as such, an additional, quantifiable generalized cost. An attempt is made to model these costs by following the example of Bates et al. (*7*).

Bates et al. suggest that the following costs are associated with early and late arrivals (*7*). For early arrivals, the cost decreases as the AAT moves toward the NAT. If the AAT equals the NAT, then zero cost is experienced. If the AAT is later than the NAT by any amount, then the traveler experiences a fixed cost, representative of the failure to be on time. The late penalty also increases with increasingly late AATs. These cost functions are shown in Figure 4.

In the case in which NAT is after Δ*, three separate responses to early and late arrivals that are representative of the example traveler's characteristics are defined. For the RA traveler, it is assumed that the cost structure of the model of Bates et al. has very low penalties for early arrival (because such risk aversion likely produces frequent early arrivals) (*7*). For the moderately RA traveler, a slightly higher penalty function for early arrivals but a slightly lower penalty function for late arrivals is assumed. Finally, for the RN traveler, early and late arrival penalties are assumed to be equal. These multiclass cost functions are shown graphically in Figure 5 and quantitatively below:

Traveler Type	*Early Arrival Penalty (min)*	*Late Arrival Penalty (min)*
RA	0.25 ∗ (NAT – AAT)	0.5*h* + (AAT – NAT)
Moderately RA	0.5 ∗ (NAT – AAT)	0.25*h* + 0.5(AAT – NAT)
RN	0.6 ∗ (NAT – AAT)	0.6(AAT – NAT)

For the case in which NAT is before Δ*, the penalty for early departure is quantified as one headway.

There are now four cases, as follows:

1. NAT is after Δ* and the bus's AAT is before NAT; a penalty for early arrival is incurred.
2. NAT is after Δ* but the bus's AAT is after NAT; a penalty for late arrival is incurred.
3. NAT is before Δ* and the bus's AAT is before NAT; penalties for early departure and early arrival are incurred.
4. NAT is before Δ* and the bus's AAT is after NAT; penalties for early departure and late arrival are incurred.

The generalized cost equation can be rewritten for each case. For simplicity, it is assumed that AT is negligible (equal to zero) and that WT is equal to one-half of the headway (0.5 h), or 7.5 min. Furth and Muller treat the impacts of reliability on WT (*4*). It is also assumed that VOT and fare are equal in all cases and can therefore be eliminated. This results in the following four generalized cost equations.

Case 1:

$$GC_T = (2.5 \cdot WT + SIVT + 1.25 \cdot late + EAP)$$

Case 2:

$$GC_T = (2.5 \cdot WT + SIVT + 2.0 \cdot late + LAP)$$

Case 3:

$$GC_T = (EDP + 2.5 \cdot WT + SIVT + 1.25 \cdot late + EAP)$$

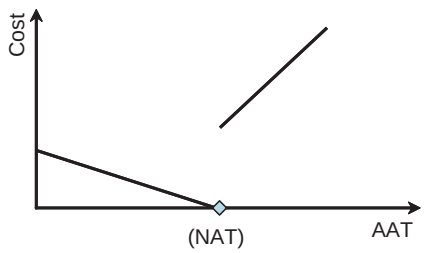

FIGURE 4 Penalties for early and late arrivals.

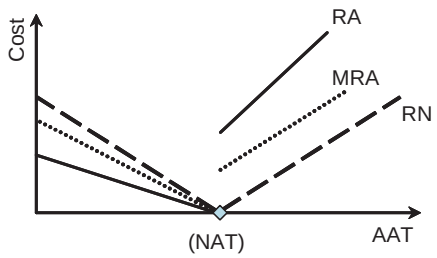

FIGURE 5 Graphical representation of multiclass penalty functions.

Case 4:

$$GC_T = (EDP + 2.5 \cdot WT + SIVT + 2.0 \cdot late + LAP)$$

where

SIVT = scheduled in-vehicle travel time (min),
late = AAT – SAT (min),
EAP = early arrival penalty,
LAP = late arrival penalty, and
EDP = early departure penalty.

These generalized cost functions disaggregate the travel time components with weighting for each one on the basis of various sources. It is standard practice to assign in-vehicle travel time a value of 1.0 and rank all other time components as more or less important. In this case, however, the in-vehicle travel time was further disaggregated into two components: SIVT and the duration of the trip that is longer than expected, after the general form proposed by Noland and Small (*8*). The SIVT component is given the standard weight of 1.0, whereas the longer than expected portion of the trip is given a higher weight that varies depending on whether the bus's AAT is later than the NAT. The weighting of "late" is lower in Cases 1 and 3 to represent a passenger's tolerance of behind-schedule operation that still results in an early arrival. The weighting of "late" is much higher in Cases 2 and 4 because the extra travel time causes the passenger to arrive after the NAT. The waiting time weighting of 2.5 is derived from the average perception of WT of Kittelson and Associates et al. (*5*), and the penalties for late and early arrivals are derived from the previous equations.

Modeling NAT and AAT

To account for both the discrete arrivals and the reliability factors, it is necessary to predict the difference of AAT and NAT. An analytic solution to this problem requires the convolution of the log-normal PDF of arrival times and the uniform PDF of NAT (*9*). Mathematically, this is quite complex. Instead, the authors elected to simulate the results with appropriately distributed arrival and NAT events.

RESULTS

In this study, 10,000 travelers who were equally likely to have each risk-aversion characteristic and who were equally likely to have a destination of each reliability category were created. A scheduled in-vehicle time of 20 min was assumed. This results in a traditional generalized cost of 38.75 min. The AVL-derived arrival distributions presented above for the stations were used. The model predicts frequencies of 78.4%, 4.5%, 17.1%, and 0.1% for Cases 1 through 4, defined above, respectively. That is, 78.4% of the time a passenger will choose the bus that is scheduled to arrive nearest the passenger's NAT and actually arrive before the NAT. Only 4.5% of the time will a passenger choose the bus that is scheduled to arrive nearest the NAT and arrive late. Just over 17% of travelers will choose an earlier bus to avoid the possibility of being late, with nearly all of them arriving on time. The model predicts a probability of 0.1% that a passenger will elect to take an earlier bus and still arrive late.

To investigate the impacts of discrete arrival times and unreliability on generalized costs, the model's generalized cost results are shown in Figure 6. The dashed line represents the traditional generalized cost.

The range of costs ranges from 43.9 min for RA travelers traveling to a high-reliability stop to 50.8 min for moderately risk averse travelers traveling to a low-reliability stop. For each class of traveler the generalized cost increases with decreased reliability. The least-reliable stop has generalized costs that exceed those for the most-reliable stop by about 7 min, or more than 15%. The moderately risk averse traveler also has the highest costs in each case. This traveler has a slightly higher threshold for choosing the first arriving bus, which results in the traveler experiencing late penalties more frequently than the RA traveler. So, although the most-RA traveler is more likely to experience a penalty for an early departure, that traveler is much less likely to experience a penalty for a late arrival.

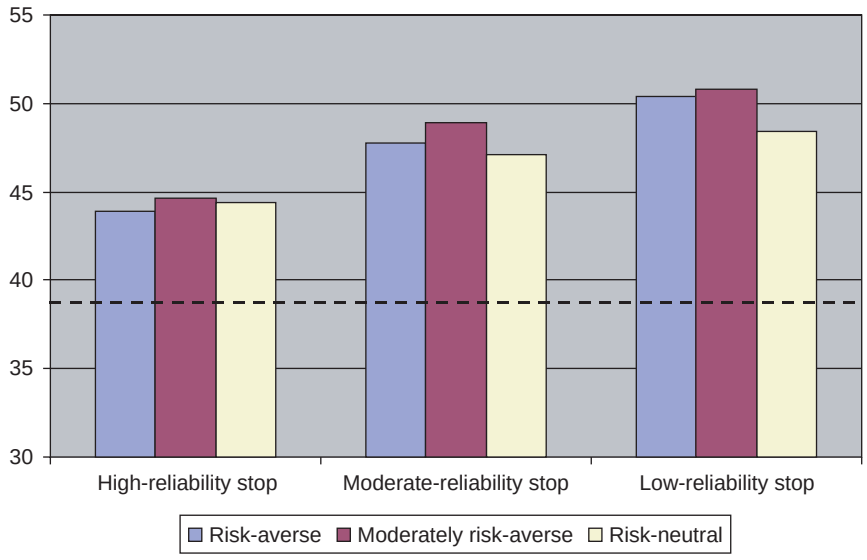

FIGURE 6 Generalized cost (minutes) results from the model.

The model also provides intuitive results for the RN traveler. As reliability decreases for the destination stop, the RN traveler is least affected of all travelers. This is indicative of the overall risk tolerance of that type of traveler and the equal (and generally smaller) perception of the penalty associated with either result.

Two conclusions may be made from the information in Figure 6. First, modelers who are predicting mode split on the basis of traditional generalized costs are likely to be underestimating the generalized cost of transit. In the example presented here, the underestimation is approximately 20% to 30%. This systematic underestimation may partially help to explain the need for the so-called transit bias coefficient that is often used to calibrate predicted mode splits to observed values. Second, this formulation can be used to quantify the actual costs of unreliable transit service. In this example, the average costs are 44.3, 48.0, and 49.9 min for high-, moderate-, and low-reliability stops, respectively. If all stops were upgraded to high reliability, the model suggests that savings of 3.1 min (approximately 7%) per passenger are possible. Multiplying this time savings per capita times ridership and the value of time provides a financial estimate of the benefits accrued. This value can be directly compared with potential infrastructure investments, such as investments in queue jumpers and transit signal priority.

CONCLUSIONS AND FUTURE WORK

This model described here is based on data from the region of Waterloo, Ontario, Canada. It demonstrates a clear methodology that can be used to assess the impacts on reliability in a reasonably reliable network. The study has shown that increasing the reliability of arrivals at a station can decrease the generalized costs of transit users. The authors further posit that the inclusion of uncertainty in the calculation of generalized costs may provide better estimates of mode splits in travel forecasting models.

Given that this model has been created, it is a relatively straightforward exercise to test different transit systems for which AVL data exist. This allows comparison of the impacts of overall reliability on users from across networks. The model formulation also allows assessment of the impacts of longer headways on penalties for early arrivals. Perhaps most importantly, the formulation presented here provides an opportunity to calibrate a model of user perceptions for disaggregate travel times. A generalized cost model that includes separate weightings for deviations from expected travel times, as well as early and late arrivals, may be able to be calibrated.

ACKNOWLEDGMENTS

This research was sponsored by the Natural Sciences and Engineering Research Council of Canada and the regional municipality of Waterloo.

REFERENCES

1. Osuna, E. E., and G. F. Newell. Control Strategies for an Idealized Public Transportation System. *Transportation Science*, Vol. 6, 1972, pp. 52–72.
2. Wilson, N. H. M., D. Nelson, A. Palmere, T. H. Grayson, and C. Cederquist. Service-Quality Monitoring for High-Frequency Transit Lines. In *Transportation Research Record 1349*, TRB, National Research Council, Washington, D.C., 1992, pp. 3–11.
3. Furth, P. G., B. Hemily, T. H. J. Muller, and J. G. Strathman. Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Review and Potential. In *TCRP Web Document 23 (Project H-28)*, Transportation Research Board of the National Academies, Washington, D.C., 2003.
4. Furth, P., and T. H. J. Muller. Service Reliability and Hidden Waiting Time: Insights from Automatic Vehicle Location Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1955*, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 79–87.
5. Kittelson and Associates, KFH Group, Inc., Parsons Brinckerhoff Quade and Douglass, Inc., and K. Hunter-Zaworski. *TCRP Report 100: Transit Capacity and Quality of Service Manual*, 2nd ed. Transportation Research Board of the National Academies, Washington, D.C., 2003.
6. Small, K. A. The Scheduling of Consumer Activities: Work Trip. *American Economic Review*, Vol. 72, 1982, pp. 172–181.
7. Bates, J., J. Polak, P. Jones, and A. Cook. The Valuation of Reliability for Personal Travel. *Transportation Research E*, Vol. 37, 2001, pp. 191–229.
8. Noland, R., and K. A. Small. Travel Time Uncertainty, Departure Time and the Cost of the Morning Commute. Presented at 74th Annual Meeting of the Transportation Research Board, Washington, D.C., 1995.
9. Meyer, P. L. *Introductory Probability and Statistical Applications.* Addison-Wesley Publishing Company, Reading, Mass., 1965.

The Transit Capacity and Quality of Service Committee sponsored publication of this paper.