EJTIR

Issue 17(2), 2017
pp. 263-278
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Modelling travel time reliability in public transport route choice behaviour[1]

Andele B. Swierstra[2]
Panteia, Netherlands

Rob van Nes[3]
Section of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Netherlands.

Eric J.E. Molin[4]
Section of Transport and Logistics, Faculty of Technology, Policy and Management, Delft University of Technology, Netherlands.

The implementation of travel time reliability (TTR) in route choice behaviour is still not very common in transport models, especially not in a public transport context. The reasons probably are that it is difficult to measure and that there is no agreement how it best can be represented in utility functions. Typically, it is represented by a standard deviation, however, particularly in public transport choices it is more likely that travellers think about the consequences of unreliability in travel times in terms of buffer times. This paper contributes to the literature by comparing five different model specifications of TTR in public transport route choices that are either based on standard deviations or on buffer time indicators. The models are estimated from choices observed in a stated choice experiment. To address heterogeneity, a latent class model is estimated. The results suggest that the reliability buffer time indicator outperforms the standard deviation indicator. Furthermore, the reliability buffer time parameter is only statistically significant in two of the four classes. The other two classes are particularly sensitive to making transfers and to low frequencies of public transport services, suggesting different strategies to deal with TTR.

Keywords: *Public transport, Travel time reliability, Route choice behaviour.*

1. Introduction

Policy makers in many countries aim to enhance sustainable mobility (Van Oort *et al.* 2013) by improving public transport (PT) service quality in order to increase the overall attractiveness of public transport. One of the aspects of PT service quality is travel time reliability (TTR). However, reliability is usually not included in transport models and consequently these models are not able to quantify the effects of enhanced service reliability. Therefore, these reliability effects cannot be taken into account in societal cost benefit analyses of, for example, infrastructural projects.

[1] An earlier version of this paper appeared in the compendium of papers of the 95th annual meeting of the Transportation Research Board, Washington DC, January 10-14, 2016.
[2] E: A.Swierstra@panteia.nl
[3] E: R.vanNes@tudelft.nl
[4] E: E.J.E.Molin@tudelft.nl

In transport models, typical attributes of specific route alternatives, such as travel cost and travel time, are valued and transformed to the same unit, usually cost in euros or dollars (or in any other currency). Incorporating reliability in a generalized cost function is however more complicated, because there is no widely accepted agreement on its unit of measurement as is the case for travel cost (measured in the nation's currency) and travel time (typically measured in minutes or hours). Because the measurement unit of reliability is not trivial (Van Loon *et al.* 2011), several distinct modelling approaches can be found in the literature. The two most common approaches are the so-called scheduling approach and the centrality-dispersion approach. Because this study was carried out in the context of improving a four-step static model (that is the VRU-model, the regional transportation model of Utrecht, the Netherlands), this paper focuses on the centrality-dispersion approach since this approach better matches the scope of static 4-step models (Paulley *et al.*, 2006). Also Carrion & Levinson (2012) concluded that the centrality-dispersion approach is preferred on practical grounds, while the scheduling approach is favoured theoretically. Fosgerau & Karlström (2010) also proved that, under some assumptions, the scheduling approach can be converted to a centrality-dispersion form. However, it should be noted that this relation becomes increasingly complex and non-linear when scheduled services are considered.

The centrality-dispersion approach, also known as the mean-variance approach, was first proposed by Jackson & Jucker (1982). It is based on the assumption that travellers place a disutility on travel time variability (TTV) itself, and the uncertainty that is associated with that. This approach assumes that the traveller makes a trade-off between expected travel time and TTV. Different measures exist for TTV of which the standard deviation is most often used (Significance, 2013, Kouwenhoven et al., 2014). Mean and median travel time are predominantly used as a centrality measure. Note that TTV and TTR are reversely related concepts: if travel time variability is high, travel time reliability is low and vice versa.

However, many different dispersion measures for both car and public transport travel can be found in the literature, because there is no consensus about how travellers perceive reliability and make decisions accordingly. The following two dispersion measure types are identified in Lomax *et al.* (2003) and are proposed in a public transport context in Van Oort (2011). The first are *Statistical range measures*, which include a Standard Deviation (SD). These are the predominantly used measures in the literature (Significance, 2013; Tseng, 2008; Hollander, 2006, Kouwenhoven et al., 2014). The second are *Percentile difference measures*, which are based on the assumption that travellers incorporate a buffer time in their trip to account for unreliability of their travel time in order to arrive on time for their planned activity. This measure was first proposed in Furth & Muller (2006) as the so-called *Reliability Buffer Time (RBT).* It is usually expressed as the difference between the 80^{th}, 90^{th} or 95^{th} percentile and the median travel time. Other measure types include: *(i) Tardy-trip measures*, which use the amount of trips that result in late arrivals, and *(ii) Probabilistic measures*, which express reliability in, for example, the probability that a trip can be made within a specified interval of time.

Because underlying aim of this study was to improve a static 4-step model, we focus however on Statistical range and Percentile difference measures. Table 1 presents the SP/RP studies found in the literature that used a centrality-dispersion approach. These studies are categorized in the use of their TTV indicator (SD or RBT), in the inclusion of other attributes in addition to TTV, that is travel cost, travel time, transfers and/or frequency, and mode type examined, either car or PT. The table makes clear that TTR has mainly been studied in car choice contexts. However, at least three important differences exist with respect to studying TTR in a public transport (PT) context. First, due to scheduled arrival times, a PT passenger is probably more aware of the precise arrival time than a car driver. This allows the passenger to compare the actual arrival time with the scheduled one. Any differences might be interpreted as unreliability, whereas in case a car route always results in a 5 minutes delay, actually the car driver may not perceive travel time unreliability. Thus, the incorporation of scheduled travel time in the generalized cost function

might be more appropriate in a PT context than in a car context (Bates *et al.*, 2001). Second, a trip made by public transport may consist of multiple legs that requires one or more transfers. When a vehicle arrives late and the transfer is missed, this immediately results in added travel time with a full headway. A missed transfer therefore always depends on the interaction between the two vehicles (Lee, 2013; Mai *et al.* 2012), and a few minutes late can have large additional travel time as a result. Third, the frequency of public transit services limits the departure times for the traveller to several discrete times per time unit (hour, day, week). Therefore, the PT arrival time has a more discrete nature than the car arrival time which may not always be ideal for travellers regarding their planned activities (Van Oort, 2011; Ma *et al.* 2014).

Table 1. Attributes used in the literature on travel time reliability

Other considered attributes	**TTV indicator**	
	SD	**RBT**
Travel cost Travel time	Car: Asensio & Matas (2008) PT: Hollander (2006). Both: Significance (2013), Tseng (2008) and (Li *et al.* (2010), Kouwenhoven et al., 2014.	Car: Ghosh (2001), Liu *et al.* (2004) and Small *et al.* (2005). PT: none
Travel cost Travel time Transfer Frequency	PT: none	PT: none

Thus, modelling TTR in a PT context induces more complexity than modelling TTR in a car choice context, hence, the best TTR modelling approach in car context is not necessarily the best modelling approach in a PT context. Furthermore, Table 1 makes clear that to the best of our knowledge, no TTR studies exist that applied RBT in a public transport context. Finally, no study to date took PT-specific attributes into account, such as the number of transfers and service frequency, whereas these play an important role in the TTR impact in a PT context, as just argued. This paper therefore intends to make the following contributions to the literature. First, to find which representation is better among two centrality-dispersion type alternatives in the context of modelling public transport choices. To that effect, various model specifications based on SD and RBT are developed and estimated from choices observed in a stated choice experiment. Results are compared in terms of interpretability and model fit. Second, to examine the relative importance of TTV attributes in a context that also describes the PT-specific attributes making transfers and service frequency. Therefore, this research will focus on pre-trip route choice to catch actual behaviour concerning travel time reliability not affected by mode preference and other long-term behaviour found in Peer *et al.* (2015). Third, to test whether different strategies exist among travellers with respect to dealing with TTR. To that effect, a Latent Class Model (LCM) is estimated to identify latent segments in the population with different valuations of TTR.

The remainder of this paper is structured as follows. In the next section, the model alternatives are developed. Subsequently, the stated choice experiment and the data collection procedure are discussed. This is followed by a presentation and discussion of the results. Finally, some conclusions are drawn.

2. Development of model alternatives

In this section, we develop alternative model specifications, which are based on the two main conceptual models for dispersion measures discussed in the Introduction, either Standard Deviation or Reliability Buffer Time. As earlier discussed, the standard deviation is the most widely used indicator for TTV. A study by Benwell & Black (1984) suggests, however, that this

indicator may not be the best one. In the latter study travellers were asked to choose between three alternative patterns of lateness with the same mean delay value as shown in Table 2. Surprisingly, the alternative with the highest standard deviation was most preferred by the respondents, while the alternative with the smallest standard deviation was the least preferred. The 80th –percentile (presented in the fourth column), however, seems to explain the results better: The alternative with the lowest 80th-percentile is most preferred, while the alternative with the highest value is the least preferred. The effect that travellers prefer an alternative with a larger travel time standard deviation will be further referred to as the Benwell & Black-effect. These results contradict the hypothesis that a larger travel time standard deviation will always result in less attractive route alternatives. Since the 80th-percentile is directly related to the concept of RBT, we expect that RBT is a better indicator of TTV than SD. However, Table 2 also suggests that much heterogeneity in preferences with respect to TTV among the travellers exists, which suggests that this should be taken into account in the modelling.

Table 2. Results of Benwell & Black (1984) (modified: 80th-percentile added)

Series of delay patterns	Mean delay	S.D.	80th-percentile	Ranked first	Ranked last
0,0,5,6,8,7,6,4,5,9	5	2.86	7	38%	47%
0,0,0,0,0,0,25,5,10,10	5	7.75	10	6%	29%
0,0,0,0,0,0,0,0,20,30	5	10.25	0	56%	24%

The following five model specifications either based on the standard deviation (SD, used in variants 1a and 1b), or Reliability Buffer Time (RBT, used in variants 2a, 2b and 2c) will be compared in this paper. These are summarized in Table 3 and graphically depicted in Figure 1.

1a The first model alternative (1a), Mean-SD (M-SD), is used in various studies (Significance, 2013; Tseng, 2008; Hollander, 2006; Asensio & Matas, 2008, Kouwenhoven et al., 2014), and it only uses the mean travel time T_{mean} and its standard deviation σ_T.

1b The second model alternative (1b), Scheduled-Mean-SD (SM-SD) is a more elaborate variant of the previous model alternative. It is based on Van Oort *et al*. 2014 and assumes that travellers experience two effects if confronted with variability in travel time duration for the same service: first, additional travel time, which in this alternative is represented by $T_{add,mean}$, calculated as the mean travel time (T_{mean}) minus the scheduled travel time ($T_{scheduled}$). Second, uncertainty around the mean travel time, which results in uncertain arrival times, which in this alternative is represented by the standard deviation of travel times σ_T as a dispersion measure.

2a Model alternative 2a, Median-RBT (M-RBT), is extensively used in Uniman (2010). It is, as the following two alternatives, based on the RBT. This alternative, as well as model 1a, does not take $T_{scheduled}$ into account, but assumes that only T_{median} and RBT play a role.

2b Alternative 2b, Scheduled-Median-RBT (SM-RBT), was originally proposed in Van Oort (2011). It uses the scheduled travel time $T_{scheduled}$, the added median travel time $T_{add,median}$ and the RBT. For reasons explained in Section 3.1, the 80th-percentile travel time will be used for RBT in this study.

2c Finally, model 2c, Scheduled-RBT (S-RBT), uses a relatively new definition of RBT and was first proposed by Ma *et al.* (2014). It only uses $T_{scheduled}$ and $RBT_{scheduled}$, which is defined as the difference between $T_{80^{th}\text{-percentile}}$ and $T_{scheduled}$ (see Figure 1). Ma *et al.* (2014) originally defined it as the difference between the Mth -percentile travel time and the typical travel time $T_{typical}$. However, since the direct expectation of the travel time of a trip comes from the timetable published by the operator, Ma *et al.* (2014) also stated that the scheduled travel time should be chosen as $T_{typical}$. Note that model 2c does not

take any statistically based centrality measure such as mean, median or standard deviation into account.

Table 3. Considered model alternatives

		Centrality measures		Dispersion measures		
Model alternative	**$T_{scheduled}$**	**T_{mean}**	**T_{median}**	**SD**	**RBT**	**Proposed/used in**
1a) M-SD		x		x		Significance (2013), Asensio & Matas (2008), Hollander (2006) and Tseng (2008) , Kouwenhoven et al., 2014
1b) SM-SD	x	x		x		Van Oort *et al.* (2014)
2a) M-RBT			x		x	Uniman (2010)
2b) SM-RBT	x		x		x	Van Oort (2011)
2c) S-RBT	x				x	Ma *et al.* (2014)

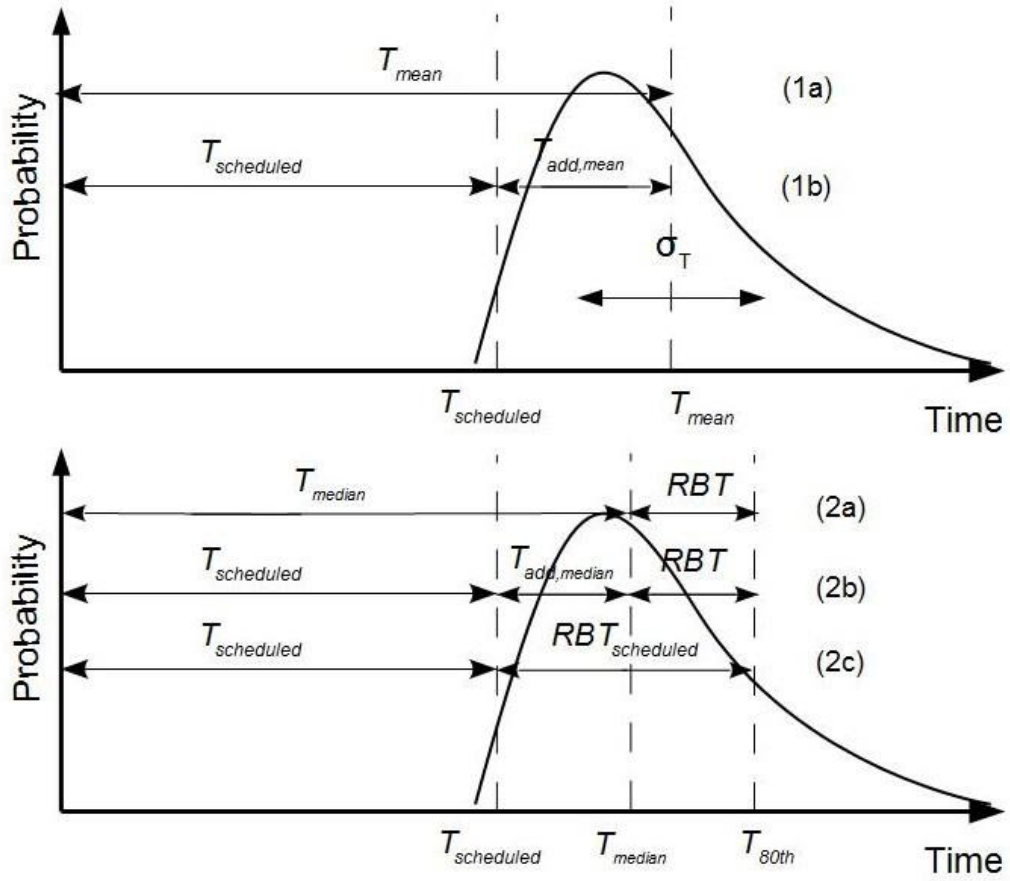

Figure 1. Illustration of SD- and RBT-models

3. Methodology

This section describes the methodology applied in this research. First, the construction of the conducted stated choice experiment and the data collection procedure are described. Subsequently, the Latent Class Model (LCM) is briefly explained.

3.1 Stated choice experiment

The collection of revealed preference (RP) data tends to be quite challenging and expensive (Peer et al., 2015), since it has been proven to be difficult to find real choice situations with sufficient variation in TTV, together with other relevant attributes, in order to obtain statistically reliable

estimates (Bates *et al.*, 2001). Regarding the timeframe of this research the stated preference (SP) method is preferred above the RP method. Therefore a stated choice experiment is constructed to observe choices between transit route alternatives. Each choice set describes two unlabelled PT routes, of which the attribute levels vary across the different choice sets. Following the format suggested by Tseng (2008) as the best one understood by respondents, travel time variability is operationalized by presenting a reference travel time, in this study the scheduled travel time, plus 5 travel times of which respondents have to assume that these all have equal probability of occurring for that particular route. Because the 5 travel times are shown in increasing order, the fourth presented travel time travel is the 80^{th}-percentile travel time; therefore this percentile will be used to define the RBT in the analysis. In addition to TTV, three other attributes are varied in the choices sets: travel cost, presence of a transfer and frequency (Tuinenga, 2014) (Hoogendoorn-Lanser *et al.* 2005). The frequency attribute is processed as the average waiting time defined as the headway divided by two, under the assumption that travellers arrive randomly at a stop. Hence, the alternatives are defined by 5 attributes.

It is widely believed that the validity of the SP results increase if respondents are presented choice situations that are familiar to them, since "it creates more realism in the SP experiments by assuring that the alternatives are similar to that which the respondent has experienced in an RP setting" (Train & Wilson, 2008). To achieve this, respondents are presented choice alternatives that aresimiliar in trip length and trip purpose as those they encounter in their daily lives. Therefore, respondents were first asked for which trip purpose they use public transport most often, either for commuting/business, for educational, for leisure/shopping or other purposes. Thereafter, the length of this trip is administered. Because the trip purpose 'other' is often a non-recurrent trip, for this 'purpose' the length of the last trip is administered. Subsequently, the respondent is assigned to the corresponding distance class, either 0-10 km, 10-30 km or >30 km. The respondents are then presented with 12 choice tasks with attribute ranges typical for trip lengths of either 5 km, 20 km and 65 km respectively (see Table 4). The corresponding attribute levels from the distance classes are adopted from Arentze & Molin (2013). An example of a choice set is presented in Figure 2.

Table 4. Attribute levels

Parameters	**Distance class** **< 10 km** (Trip length 5 km)	**Distance class** **10 - 30 km** (Trip length 20 km)	**Distance class** **> 30 km** (Trip length 65 km)
$T_{scheduled}$ (min)	15/20/25	30/40/50	70/80/90
RBT (min)	0/3/6	0/5/10	0/5/10[5]
$T_{add,median}$ (min)	0/3/6	0/5/10	0/5/10
Travel cost (€)	0/1/2	0/2/4	0/7.5/15
Transfers	0/1	0/1	0/1
Frequency (per hour)	2/8	2/8	1/4

[5] Note that these values are similar to those of the second distance class. However, larger values were considered to be unrealistic by the participants of the pilot survey.

PT-alternative A	**PT-alternative B**
Scheduled travel time **25 min**	Scheduled travel time **15 min**
There is an equal probability on the following travel times (potentially missed transfers included) **25 min** **28 min** **28 min** **31 min** **36 min**	There is an equal probability on the following travel times (potentially missed transfers included) **15 min** **18 min** **18 min** **21 min** **26 min**
Travel cost **€ 1**	Travel cost **€ 2**
Number of transfers **0**	Number of transfers **1**
Lowest frequency **8 times/hour**	Lowest frequency **2 times/hour**

Figure 2. Example of a choice set

In order to obtain the most reliable estimates, hence those with the smallest possible standard errors, and to avoid dominance among the alternatives, the D-efficient design method is applied (Bliemer *et al.*, 2009) to construct the choice alternatives. In order to construct such a design, the best available estimates of the true parameter values, so-called priors, are required. For this, estimates found in the literature (Significance, 2013, Kouwenhoven et al., 2014) were first used as priors in a small pilot choice experiment, which was conducted to test the choice experiment and obtain better priors. The experimental design is arbitrarily optimized for model 2b SM-RBT. The NGENE software package (ChoiceMetrics, n.d.) was applied to generate the experimental design.

An Internet panel was used for recruiting respondents, resulting in 525 respondents. Respondents were mostly recruited from urban areas in the Netherlands in order to have a higher probability of recruiting regular transit users. This resulted in 75% of the respondents that stated they use public transport more than once per month, which suggests that most respondents are indeed regular transport users. Because respondents were recruited from a commercial panel that rewards respondents for filling out questionnaires, there is a certain risk that part of the respondents only participate for financial gain and are not fully engaged. After we could not meaningfully interpret estimated MNL and Mixed Logit Models from the observed choices in the experiment (see section 4), we tried to identify respondents that were not fully engaged in the choice task by applying the following selection criteria: Respondents that either failed to correctly answer the dominant stated choice question, gave repetitive answers (only alternative A or only alternative B) or completed the survey faster than a certain time restriction (200 seconds: which involves that respondents need on average at least 10 seconds for each choice task and 80 seconds to read and understand the explanation of the experiment), were removed from the dataset. This led to the exclusion of 101 respondents, resulting in 424 respondents. From this reduced sample we still were not able to estimate interpretable MNL and ML models, however, we were able to meaningfully interpret a latent class model (see next subsection).

Because this result came at the expense of removing a considerable group of respondents, we tested whether the criteria for removing a respondent could be relaxed. We consecutively estimated a latent class model from three different response groups, which each were constructed

by applying only one of the selection criteria. The model most akin to the final model presented in this paper, was estimated from the group that excluded the respondents that did not meet the time criterion. The models estimated from excluding repetitive answers or opting for dominated options, the response groups that still included those who did not meet the time criterion, resulted in uninterpretable models. Applying the time criterion led to excluding 67 respondents and was the criterion the led to largest number of excluded respondents. The model estimated from this group (N=458) led to the same number of classes, but the structure was somewhat less clear than the model estimated from applying all three criteria (N=425): the SRBT is only statistically significant at the 0.05 level in one of the classes and additionally statistically significant at the 0.10 level (two-sided test) in another class. Moreover, a model without a membership function fits better. This suggests that not only the respondents that did not meet the time criterion were not fully engaged in the choice task, but also the respondents that did not meet the other two criteria. Based on these test results, we still have most confidence in the model we estimated from the response group we obtained by applying all three selection criteria.

3.2 Latent class model

A latent class model is estimated to address heterogeneity in preferences among the travellers. It is assumed that segments exist in the population that have different preferences, but which are internally relatively homogenous. These classes or segments cannot be observed and are therefore latent and emerge in the estimation process (Nylund *et al.*, 2007). The latent class model estimates a set of parameters for each class, where each set describes the preference tastes of that class. Simultaneously, a class membership model is estimated that predicts the probability of each individual belonging to each class, in which observed individual characteristics such a socio-demographic variables may be included as predictors.

The latent class model does not assume any particular shape for the distribution of tastes, since the classes are discrete (Significance, 2013). Each set of class specific parameters may be regarded as stemming from an MNL model. However, due to the fact that a set of parameters is estimated for each class, the latent class model does not have the Independence of Irrelevant Alternatives property, which allows for more valid predictions than the MNL model. The LC models will be estimated using the software package Nlogit 4.0.

In a latent class model, the number of segments is unknown and, therefore, the optimal number of segments needs to be determined. The likelihood ratio test cannot be applied in this case as the different models are not nested. Instead, the optimal number of classes is determined by using the Akaike Information Criterion and the Bayesian Information Criterion (Gupta & Chintagunta, 1994). The goal is to find the model with the minimal AIC or BIC, where the BIC penalizes extra classes more heavily and shows the better performance of the two (Nylund *et al.*, 2007).

4. Results

In order to have basic reference models, MNL models were first estimated for all five model alternatives. However, the results were not satisfactory: not in any of the model alternatives a TTR related parameter was found to be statistically significant or had the expected sign. Also more advanced models did not result in more satisfying results: we tried segmentation on observable characteristics, estimation of interaction effects with personal characteristics, and mixed logit models involving randomly distributed parameters to address taste heterogeneity. This suggests that taste heterogeneity exists that cannot be captured by segmentation based on observable characteristics or by assuming distributions around mean parameters. To examine whether different strategies exist to deal with TTV in the context of PT travel, latent class models were estimated. The results of the best fitting LC-models for each of the five model alternatives are presented in Table 5. The optimal number of classes for each model alternative is shown in the second column of Table 5. The table shows that the minimum adjusted Rho square value of

any of the latent class models is 0.294 which is a considerable improvement in goodness of fit compared to the estimated MNL and mixed logit models, which had adjusted Rho-square values in the range of 0.180-0.200 and 0.230-0.250 respectively.

Table 5. Results of latent class model estimations for all model alternatives

Model	No. of classes	No. of parameters	Adj. Rho²	AIC	BIC	Correct signs for all TTR parameters?	TTR parameters significant?
1a M-SD	5	29	0.318	0.952	0.989	No	Yes
1b SM-SD	4	27	0.316	0.954	0.989	Yes	No
2a M-RBT	4	23	0.294	0.982	1.012	No	Yes
2b SM-RBT	4	27	0.301	0.974	1.009	No	Yes
2c S-RBT	4	23	0.316	0.953	0.983	Yes	Yes

The last two columns of Table 5 indicate whether all TTR related parameters have the correct sign and whether any of the TTR parameters are statistically significant in all classes. Only in models 1b and 2c all TTR parameter estimates have the correct sign. Of those two models, only model 2c contains statistically significant TTR related parameters. Moreover, although model 2c has the same adjusted ρ^2 value as model 1b[6], it has a smaller number of parameters and is thus the more parsimonious model. Hence, it is the only model in which all TTR related variables have the correct sign, one of its parameters is statistically significant in at least one class, it is parsimonious and has a good model fit. Hence, we conclude that model 2c is the preferred model. This result suggests that, as expected, RBT is a better indicator of TTR than SD in the context of PT route choices.

Table 6 shows the estimates of model 2c. Each class is highly sensitive to a single attribute, which is consequently used to label the segment. The segments are briefly discussed in the following:

- In segment 1 the parameters of $RBT_{scheduled}$ and T_{wait} are not statistically significant, and compared to most other segments, the $T_{scheduled}$ parameter has a somewhat lower value. Most remarkable is the very large parameter value estimated for Transfer. Apparently the main strategy of this segment in dealing with travel time uncertainty is to avoid transfers, while they don't mind waiting and traveling somewhat longer. Hence, it is labeled as 'transfer sensitive'.
- Segment 2 does not have a significant $RBT_{scheduled}$ parameter. Of all segments, this segment is most sensitive to waiting time; its parameter value is almost three times as high as its scheduled travel time value. The main strategy of this 'waiting time sensitive' group is choosing route options with a high frequency of service.
- Segment 3 is the only segment with significant parameter estimates for all attributes. Most distinctive of this class is the very high value of the cost parameter and it is therefore labeled as 'cost sensitive'.
- Also in segment 4 the estimate of $RBT_{scheduled}$ is statistically significant, while its transfer parameter is not statistically significant. Compared to the previous segment, all parameters have somewhat lower values, which is illustrated by the fact that the ratio between its $RBT_{scheduled}$ and $T_{scheduled}$ parameter is almost the same as in segment 3 (0.168 and 0.163 respectively). Although its Tscheduled parameter is somewhat lower than that of segment 3, within this segment this parameter has the highest t-value and therefore

[6] Note that model 1a has an even better model fit than model 1b, however the SD-parameter is estimated with the incorrect sign for all segments, which might be explained by the Benwell & Black-effect.

this segment is labeled 'time sensitive'. However, it should be noted that this group has less distinct features than the other segments.

All class probabilities are between 20-30%, hence, the four segments have fairly equal class sizes. The results make clear that segments 3 and 4 have statistically significant $RBT_{scheduled}$ parameters. As those two segments represent 56.3% of the travellers, this suggests that the majority of the travellers derives a direct disutility from the amount of buffer time. However, this does not mean that only those two groups are sensitive to TTV: the main strategies applied by the first two groups, that is avoiding transfers and choosing for high frequency services, may be regarded as a general strategy to deal with uncertainty in PT travel times. In sum, the results suggest the existence of four different decision strategies in making PT route choices in the context of travel time variability.

Table 6. The estimated parameters of the segments of model 2c ($RBT_{SCHEDULED}$)

	1. "Transfer sensitive"		2. "Waiting time sensitive"		3. "Cost sensitive"		4. "Time sensitive"	
	Par.	*t-ratio*	Par.	*t-ratio*	Par.	*t-ratio*	Par.	*t-ratio*
Attributes								
$T_{scheduled}$	-0.081	*-3.076*	-0.038	*-5.917*	-0.127	*-9.011*	-0.103	*-15.737*
$RBT_{scheduled}$	0.033	*1.878*	0.010	*1.085*	-0.021	*-2.334*	-0.016	*-3.178*
Cost	-0.374	*-3.607*	-0.547	*-17.522*	-2.058	*-21.123*	-0.246	*-9.537*
Transfer	-2.185	*-11.077*	-0.559	*-5.588*	-0.937	*-9.688*	-0.023	*-0.457*
T_{wait}	0.002	*0.299*	-0.106	*-19.875*	-0.032	*-7.565*	-0.009	*-4.979*
Class probability	23.3%		20.4%		27.0%		29.3%	
Class membership								
Constant	-0.842	*-3.543*	-0.875	*-3.631*	-0.022	*-0.108*	0.000	(fixed)
Distance	0.018	*3.813*	0.016	*3.282*	-0.003	*-0.367*	0.000	(fixed)

To explain class membership, we explored different variables as predictor in the MNL class membership function. Age and travel distance appeared as the only two serious candidates for inclusion. However, inclusion of only distance resulted in a higher adjusted Rho-square value than inclusion of distance and age together or inclusion of only age. Thus, only distance was included in the membership function of the final model. The parameters of the function that predicts for each respondent the probability of belonging to each class are presented at the bottom of Table 6. The estimated constants of segments 1 and 2 are both statistically significant and negative, and indicate an overall smaller probability of belonging to these groups compared to the fourth segment (the reference group of which parameters are fixed to zero), which is reflected in the lower class probability values that are also presented in the Table. Furthermore, the distance parameters of the first two groups are statistically significant and positive, meaning that with increasing distance of the reference trip, the probability of belonging to these two groups increases. Among others, this suggests that travellers with longer reference trips are more sensitive to making transfers (segment 1) or to waiting time (segment 2), while they are less sensitive to the amount of buffer time.

5. Discussion of results

The results of this research are compared with those of other studies in Table 7. For this we calculate several ratios that allow comparison with other studies. The Reliability Ratio (RR) is here defined as the Value of Reliability (VOR) divided by the Value of Time (VOT). The waiting time factor gives the ratio between the waiting time parameter and the scheduled travel time parameter. The transfer penalty gives the ratio between the transfer parameter and the scheduled travel time parameter. No studies were found which used $RBT_{scheduled}$ as an operationalization of

reliability. Also studies that directly value RBT are scarce. In the beginning 2000's some RP studies were conducted on a motorway stretch in California. The RR values found in these studies are compared with the average value of the RR of all classes found in the latent class analysis. Also the RR found in the national VOT study of the Netherlands, for BTM and train, is presented, as well as some other SP studies. These studies used the standard deviation of travel times as a TTR indicator. From Tseng (2008) we know that this TTR definition is roughly equal to $T_{85th\text{-}percentile} - T_{50th\text{-}percentile}$, when a normal distribution for travel time is assumed (which is not true in most cases). Table 7 also compares the waiting time factor and transfer penalty found in this research with a few other studies. Hoogendoorn-Lanser *et al.* (2005) conducted a RP survey for the Rotterdam-Dordrecht region, while Tuinenga (2014) (RP study) is applicable on the Paris metropolitan region.

Table 7. Comparison with results of other studies

Research	RR	TTR indicator	Average VOT	Range VOT	Waiting time factor (min)	Transfer penalty (min)
Small *et al.* (2005) (car)	0.91	80th - 50th percentile	-	-	-	-
Liu *et al.* (2004) (car)	1.61	90th - 50th percentile	-	-	-	-
Ghosh (2001) (car)	1.17	90th - 50th percentile	-	-	-	-
Hollander (2006) (PT)	0.10	Standard deviation	-	-	-	-
Tseng (2008) (car/PT)	0.50	Standard deviation	-	-	-	-
Significance (2013), Kouwenhoven et al., 2014 (car/PT)	0.60	Standard deviation	€8.25	€6.00 - €19.75	-	-
Shires & De Jong (2009) (car/PT)	-	-	€15.17	€6.21 - €24.83	-	-
Börjesson & Eliassion (2011) (car/PT)	-	-	€5.08	€2.80 - €7.20	-	-
Hoogendoorn-Lanser *et al.* (2005) (PT)	-	-	-	-	2.2	5.1-11.4
Tuinenga (2014) (PT)	-	-	-	-	1.34	1.06-9.70
Current research	0.00-0.17	80th perc. - $T_{scheduled}$	€12.26	€3.37 - €24.64	0.63	10.67

The results indicate that the RR found in this research is much smaller than in the other studies, however, as indicated before, those studies use other TTR indicators and are more or entirely focused on private cars. Furthermore, the (average) transfer penalty found in this research is on the high end of the range found in the other studies. A possible explanation is that presenting variable travel times makes travellers more aware of the risk of a transfer, that is that it immediately increases travel time by a full headway. Despite the fact that in the survey it was specifically stated that potentially missed transfers were included in the presented variable travel times, it seems that respondents tend to be risk avoiding when variable travel times are shown in combination with a transfer. The fact that the transfer penalty is especially high in the segment with an insignificant $RBT_{scheduled}$ parameter supports this explanation.

Moreover, the waiting time parameter is relatively low for the classes with a statistically significant $RBT_{scheduled}$ parameter. A possible explanation for this is that some of the waiting time

related disutility is allocated to $RBT_{scheduled}$ (or transfer) parameter, since the magnitude of the consequence of a missed connection is indicated by the frequency of the service (low frequencies result in long waiting times). Thus, the results might show that the respondents' perception of TTR, frequency and transfer, and its valuation, are intertwined. The interaction between TTR and transfer was probably also encountered in Significance (2013) (SP study), where the estimation of a transfer parameter in the model lead to 'wrong signs and/or insignificant estimates for this variable, and inconsistencies between purposes' (Significance, 2013). The identification of the four latent classes make clear that different strategies exist to deal with this complex interplay of factors related with travel time variability in public transport.

A remarkable finding is that in most of the classes a single parameter is by far the most prominent parameters as suggested by very high t-values. This suggests that instead of making trade-offs among all attributes, a considerable part of the respondents applied simplifying heuristics to complete the choice task, such as for example a lexicographic choice rule, that is that one considers only a single attribute and one always chooses the alternative that has the best values on that attribute. To examine to what extent this is the case in this study, we conducted the following analysis. We first assigned each respondent to only one class that is the class for which they have the highest estimated probability of belonging to. Then we examined which share of the class members always chooses for the alternative that scored best on the prominent attribute in that class. Only the choice sets that have different values on the prominent attribute were considered. The results are presented in Table 8.

Table 8. Percentage of respondents that based their decision on the segment-specific prominent attribute

Class	Decision based on the prominent attribute
Cost sensitive	60%
Time sensitive	17%
Transfer sensitive	30%
Waiting time sensitive	25%

The highest percentage (60%) of always choosing for the prominent attribute is observed in the cost sensitive class. However, this is the only class in which all parameters are statistically significant and what is more, they all have the expected sign. This suggests that although members of this class consider travel costs most important, they actually also considered the other attributes and made trade-offs among them. For the three other classes much smaller percentages of always choosing the alternative that scores best on the most prominent attribute are observed. From these results we may conclude that lexicographic decision-making does not play a large role in the observed choices and it cannot explain the prominence of the single attributes in each class. These results thus suggest that the latent segments indeed have different strategies to deal with uncertainty in travel times.

These findings might be affected by the rather complex experimental set-up, i.e. the way travel time reliability is presented and the number of attributes. Although we based our format for presenting TRR on the best scoring format found in empirical research, we cannot rule out that its meaning is hard grasp for part of the respondents. Furthermore, the use of an Internet panel might lead to biased results as well. On the other hand, in reality public transport travellers are confronted with choice situations that are even more complex than this stated choice experiment. SP studies that have limited the description of choice situations to reliability, travel time and travel cost only, ignore important characteristics as transfers and frequencies and may therefore arrive at less valid models. Moreover, there is a large variety in public transport travellers, so the concept that they might use different strategies to cope with these complex situations seems certainly plausible.

Regarding the observation that the $RBT_{scheduled}$ showed better results than the SD, we should acknowledge that this could be biased by the presentation of the five possible travel times. Also the fact that the experimental design is optimized for one of the RBT-models could have biased the result. However, the Benwell & Black-effect suggested a shortcoming of standard deviation as an indicator of travel time reliability. Nevertheless, corroboration in future research that applying different representations of travel time reliability is needed before firm statements about the best TTR indicator can be made.

6. Conclusions

This paper examined travellers' preferences regarding travel time reliability in public transport route choice. The main question addressed was how best to represent travel time reliability in utility models. To answer this question, five model alternatives have been developed and estimated from choices observed in a stated choice experiment. To address heterogeneity, latent class choice model were estimated. The results suggest that RBT (Reliability Buffer Time) is a better indicator for inclusion of TTR (travel time reliability) in utility models than the SD (standard deviation). This suggests that when travellers are presented with a range of possible travel times that are all equally likely, most travellers do not "calculate" a standard deviation and react to that outcome, but they rather consider the acceptability and probability of the delays. Of the three tested RBT variants, the variant with only the $RBT_{scheduled}$ parameter appeared to provide the best results. This indicator is the buffer time calculated by the difference between the 80^{th}-percentile of the five presented travel times with the scheduled travel time. A possible interpretation is that the buffer time could be regarded as the additional time travellers add to the scheduled travel time in order to be on time in most of the trips they make. More specifically, the 80^{th} percentile would then mean that if they add this buffer to the scheduled time, they are on time in 80% of all trips they make, while they are late in 20% of the trips. Therefore, the $RBT_{scheduled}$ represents a degree of acceptability of being too late. The decision for the 80^{th} percentile is, however, somewhat arbitrarily and should be further researched.

A four-segment latent class model appeared to fit the data best, which showed that considerable heterogeneity in preferences exists among respondents. The four segments suggest different strategies in dealing with the complex interplay of factors that are related to travel time reliability, that is variability of arrival time, having to make transfers and frequency of service that determines waiting time. While buffer time was not statistically significant in the first two classes, the main strategies in making PT route choices in those classes still seem to be related with travel time reliability: the main strategy of the first segment is to avoid making transfers, while the main strategy of the second segment is to avoid low service frequency (hence, minimizing waiting time). Recall that missing a transfer due to irregular public transport services might lead to substantial longer travel times. In the latter two segments that represent the majority of travellers (56.3%), buffer time is statistically significant, while these two segments have more conventional strategies in making PT choices: the third segments mainly focuses on minimizing travel costs, while the fourth segment mainly focuses on minimizing travel time (although this strategy is less pronounced). Overall, the results of this study suggest that when it comes to TTR in PT route choice behaviour it is worthwhile to estimate latent class models, since this significantly improves the model fit and the interpretation of the results. This heterogeneity could not be captured by segmenting on observable characteristics or by mixed logit models. Also the inclusion of PT-specific attributes, such as transfer and frequency, is recommended in future research, since these attributes are closely interrelated with TTR in a public transport context.

It should be noted that the results presented in this paper are found in the context of a stated choice experiment. Travel time reliability is operationalized in a specific way by presenting five equally likely outcomes. Hence, this represents a situation where the traveller is aware of travel

time reliability and has full knowledge on the likelihood and the magnitude of the travel time outcomes. It may be obvious that this may not be the case in the real world. Although it seems reasonable to assume that most travellers are aware of travel time variability, they may not be aware of the magnitude. Moreover, travellers' perceived distributions of TTR may differ from true values. Hence, the way travel time variability is stored in their memory may influence the performance of TTR indicators. Hence, an interesting direction for future research is to examine the performance of TTR in a revealed choice context. In the Netherlands some unique data sets exist that may offer opportunities for conducting such a research. Public Transport Chipcard data allows for the analysis of public transport trip making over time, while GOVI-data provides detailed information on the actual (operational) quality of the services offered (see also Van Oort *et al.* and GOVI (n.d.)). By analysing route choice behaviour of a specific group of travellers and linking their choices to the actual operational performance of the public transport system, more detailed insight can be obtained in the way travellers perceive TTV, and what kind of strategies they apply when they are faced with a transport system that inherently reveals variability in vehicle times and in which they experience the discrete impacts when missing a vehicle or missing a transfer.

Furthermore, and as argued before, it is highly recommended that such a research considers estimating latent class models to capture heterogeneity. Our results suggests that different classes exhibit different behaviours, that is, they apply different decision weights to arrive at decisions. That we could not find interpretable MNL and mixed logit models in this study suggests that, in contrast to widespread belief, MNL models do not always produce robust results in case of heterogeneity.

Acknowledgements

This research is performed in cooperation with BRU, the transit authority in the region Utrecht, the Netherlands, and Delft University of Technology, Department of Transport & Planning.

References

Arentze, T.A. and Molin, E.J.E. (2013). Travelers' preferences in multimodal networks: Design and results of a comprehensive series of choice experiments. *Transportation Research Part A: Policy and Practice*, 58, 15-28.

Asensio, J. and Matas, A. (2008). Commuters' valuation of travel time variability. *Transportation Research Part E*, 44, 1074–1085.

Bates, J., Polak, R.B., Jones, P. and Cook, A. (2001). The valuation of reliability for personal travel *Transportation Research Part E*, 37, 191-229.

Benwell, M. and Black, I. (1984). *Train service reliability on BR intercity services. Report 3: passenger attitudes,* Cranfield University.

Bliemer, M.C.J., Rose, J.M. and Hensher, D.A. (2009). Efficient stated choice experiments for estimating nested logit models. *Transportation Research Part B: Methodological*, 43, 19-35.

Börjesson, M. and Eliassion, J. (2011). Experiences from the Swedish value of time study, *CTS working paper*.

Carrion, C. and Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, 46(4), 720–741.

ChoiceMetrics (n.d.) *Ngene 1.1.2 User manual & reference guide.*

Fosgerau, M., and Karlström, A. (2010). The value of reliability. *Transportation Research Part B: Methodologica*l, 44(1), 38-49.

Furth, P.G. and Muller, T.H.J. (2006). Service Reliability and Hidden Waiting Time: Insights from AVL Data. *Transportation Research Record*, 1955, 79-87.

Ghosh, A. (2001). *Valuing Time and Reliability: Commuters' Mode Choice from a Real Time Congestion Pricing Experiment*. Ph.D. thesis, University of California.

GOVI (n.d.) *Public transport information without frontiers*, http://www.govi.nu/index.php?lang=english, accessed November 10, 2015.

Gupta, S. and Chintagunta, P.K. (1994). On using demographic variables to determine segment membership. Journal of Marketing Research, 31, 128-136.

Hollander, Y. (2006) Direct versus indirect models for the effects of unreliability. *Transportation Research Part A*, 40, 699-711.

Hoogendoorn-Lanser, S., Bovy, P.H.L. and Uges, R. (2005). Modelling route choice behaviour in Multimodal transport networks. *Transportation,* 32, 341–368.

Jackson, B.W. and Jucker, J.V. (1982). An empirical study of travel time variability and travel choice behavior. *Transportation Science*, 16, 460-475.

Kouwenhoven M., De Jong, G.C., Koster, P., Van Den Berg V.A.C., Verhoef, E.T., Bates, J. and Warffemius, P.M.J. (2014). New values of time and reliability in passenger transport in The Netherlands, *Research in Transport Economics*, 47, 37-49.

Lee, A., van Oort, N. and Van Nes, R. (2014). Service reliability in a network context, *Transportation Research Record*, No. 2417, 18-26.

Li, Z., Hensher, D.A. and Rose, J.M. (2010). Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation Research Part E: Logistics and Transportation Review,* 46, 384-403.

Liu, H.X., Recker, W. and Chen, A. (2004) Uncovering the contribution of travel time reliability to dynamic route choice using real-time loop data. *Transportation Research Part A*, 38, 435-453.

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

Ma, Z., Ferreira, L. and Mesbah, M. (2014). Measuring Service Reliability: Using Automatic Vehicle Location Data. *Mathematical Problems in Engineering*.

Mai, E., List, G. and Hranac, R. (2012). Simulating the travel time impact of missed transit connections. *Transportation Research Record,* 2274, 69-76.

Nylund, K.L., Asparouhov, T. and Muthen, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte-Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J.M., Wardman, M.R., Shires, J.D. and White, P. (2006). The demand for public transport: The effects of fares, quality of service, income and car ownership. *Transport Policy*, 13, 295-306.

Peer, S., Verhoef, E., Knockaert, J., Koster, P. and Tseng Y.Y. (2015). Long-Run Versus Short-Run Perspectives On Consumer Scheduling: Evidence From A Revealed-Preference Experiment Among Peak-Hour Road Commuters. *International Economic Review*, 56(1), 303-323.

Shires, J.D. and De Jong G.C. (2009). An international meta-analysis of value of travel time savings. *Evaluation and Program Planning*, 32, 315-325.

Significance, VU University Amsterdam, John Bates Services (2013). *Values of time and reliability in passenger and freight transport in the Netherlands*. Den Haag: Ministry of Infrastrcture and Environment.

Small, K.A. (1982). The scheduling of consumer activities: work trips. *American Economic Review*, 72, 467-479.

Small, K.A., Winston, C. and Yan, J. (2005). Uncovering the distribution of motorists' preferences on travel time and reliability. *Econometrica*, 73, 1367-1382.

Train, K., and Wilson, W.W. (2008). *Estimation on stated-preference experiments constructed from revealed-preference choices.* Transportation Research Part B, 42, 191–203.

Tseng, Y. (2008) *Valuation of Travel Time Reliability in Passenger Transport*. VU Amsterdam, Amsterdam.

Tuinenga, J.G. (2014). ANTONIN, een model voor de regio Parijs in Dutch. *Platos Colloquium 2014*. Den Haag.

Uniman, D.L. (2010). *Service Reliability Measurement Framework using Smart Card Data: Application to the London Underground*. University of California, Berkeley.

Van Loon, R., Rietveld, P. and Brons, M. (2011). Travel-time reliability impacts on railway passenger demand: a revealed preference analysis. *Journal of Transport Geography*, 19, 917-925.

Van Oort, N. (2011). *Service Reliability and Urban Public Transport Design*. TRAIL Thesis Series 2011/2, TRAIL, Delft (http://repository.tudelft.nl/view/ir/uuid%3A68f6dd34-53cf-4792-81e7-799c3d552b94/).

Van Oort, N., Brands, T., De Romph, E. and Flores, J.A. (2015). Unreliability effects in public transport modelling, *International Journal of Transportation,* 3(1), 113-130.

Van Oort, N., Sparing, D., Brands, T. and Goverde, R. (2015). Data Driven Improvements in Public Transport: the Dutch Example, *Public Transport,* 7(3), 369-389.