Dwell-time station-service analysis using a Rasch analysis technique

Kuipers, Ruben; Tortainchai, Natchaya; Tony, Neba C; Fujiyama, Taku

Published in:
Transportation Research Interdisciplinary Perspectives

DOI:
10.1016/j.trip.2024.101119

2024

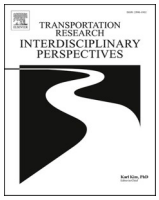
Document Version:
Publisher's PDF, also known as Version of record

Transportation Research Interdisciplinary Perspectives 25 (2024) 101119

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



Dwell-time station-service analysis using a Rasch analysis technique

Ruben Alaric Kuipers [a,b], Natchaya Tortainchai [c,*], Neba C Tony [d,e], Taku Fujiyama [d]

[a] Department of Technology and Society, Lund University, Lund, Sweden
[b] K2 The Swedish Knowledge Centre for Public Transport, Lund, Sweden
[c] The Cluster of Logistics and Rail Engineering, Faculty of Engineering, Mahidol University, Thailand
[d] Department of Civil, Environmental and Geomatic Engineering, University College London, United Kingdom
[e] Transportation Research and Injury Prevention Centre, Indian Institute of Technology Delhi, India

ARTICLE INFO

Keywords:
Dwell time
Rasch analysis
Timetable
Planning
Railways
Commuter trains

ABSTRACT

In order to ensure punctual and robust service, it is vital to have a good understanding of the current performance of a railway network. Several approaches to doing so exist but lack the ability to compare both service and station performance in a single dimension. The study presented here proposes the use of the Rasch analysis technique within an operational context to compare the relative dwell time performance of stations and services. To do so, we make use of data from commuter trains in Sweden and the UK. The results from the study suggest that the method can be used to study dwell times on a line level and can capture the variability in dwell times. Assessing the model output also shows that the approach adequately reflects the expected variability in both service performance and station difficulty. Comparing the model output to more commonly used indicators for dwell time, we find that the Rasch analysis allows us to better identify cases where planners can make adjustments to reduce the likelihood of dwell time delays. In addition to this, we highlight that a common assumption that more passengers lead to worse dwell times does not hold. Having more in-depth insights into where dwell time performance is troublesome can help planners to make more informed decisions which helps towards improving overall dwell time performance, reducing delays, and improving the attractiveness of trains as a mode of transport.

Introduction

With the increasing popularity and frequency of trains comes a growing pressure to make optimal use of available capacity. Capacity is limited in most railway networks, while there is an ambition to run more trains with a higher frequency at the same time. In the Netherlands, for example, there is the ambition to run six trains an hour on some busy corridors (Ministerie van Infrastructuur en Waterstaat, 2022). High land prices and the hefty costs of upgrading the existing infrastructure mean that further utilization of existing infrastructure through better operation planning is key. It is vital to implement better and more robust timetable planning practices that include realistic and appropriate times for all processes that make up a timetable (Hansen, 2010; Harris et al., 2014; Vieira et al., 2018) to accommodate the higher number of trains desired on the existing infrastructure.

Broadly speaking, we can divide the timetable into two main processes: the run time of trains between stations and the dwell time of trains at stations. The study we present here focuses on the latter, which is the dwell time of trains at stations. Dwell time refers to the time needed for trains to halt at a station to allow passengers to alight and board. Dwell times and dwell time delays are relevant to study as their impact on the overall punctuality of railways can be strong. In their study on how to monitor punctuality improvements, Palmqvist and Kristoffersson (2022) highlight the relationship between the frequency of dwell time delays and the overall punctuality of trains. The authors state that reducing the frequency of dwell time delays is one way to achieve an improvement in the punctuality of railways. Dwell time delays also reduce the effective capacity of a railway network, as dwell times accumulate to a large portion of journey time (Christoforou et al., 2020). Large variations in dwell times also reduce the robustness of train operations (van den Heuvel, 2016), as trains dwelling longer than scheduled can cause knock-on delays by occupying the platform, meaning that the following train cannot enter the station (Yamamura et al., 2012). This makes dwell time scheduling an important aspect of timetables since scheduling too little time can lead to delays, for the aforementioned reasons, whereas scheduling too much time will lead to

* Corresponding author.
E-mail address: natchaya.tor@mahidol.ac.th (N. Tortainchai).

https://doi.org/10.1016/j.trip.2024.101119
Received 27 September 2023; Received in revised form 10 April 2024; Accepted 20 May 2024
Available online 24 May 2024

a capacity loss (Goverde, 2005).

When making improvements in dwell time punctuality, either through a new scheduling regime or by making investments in the infrastructure, it is vital to have a deep understanding of how dwell times perform on a network level. This means that it is necessary to understand dwell times for both stations and services. A lack of information and understanding will hinder the implementation of strategies to improve dwell time performance (Pritchard et al., 2021). Having in-depth information on how dwell times perform on a network level makes it possible to identify whether the problem of actual dwell times exceeding scheduled dwell times is isolated to a small set of stations or services or if there is a network-wide problem. If the former is the case, isolated and targeted interventions will be beneficial, whereas a more system-wide approach would be desirable in the latter case.

In this study, we propose a Rasch analysis-based approach to perform an analysis of dwell time performance on a line level, directly comparing the relative dwell time performance of stations and services both separately and in combination with one another. Here we use the term *service* to refer to a unique train running between its origin and destination. For instance, we consider two trains operating between the same origin and destination, but operating on different days or at different times of the day, as two distinct services. For our study, we make use of data collected for both Swedish and UK-based commuter trains and apply the method to both cases. It is important to emphasize that our study does not compare the two networks, but rather utilizes data from both networks to better understand the applicability of a Rasch analysis-based approach for dwell time research. The proposed method can serve as a way for planners to perform a network search in order to identify stations and lines that perform poorly and require attention. The raw outcome of the Rasch analysis does not show how to improve dwell times, but rather how stations and services perform relative to each other. We, therefore, also propose how to use the output of the Rasch analysis in relation to dwell time evaluation on a line level and how this can be used to gain a deeper understanding of dwell time performance within a railway network.

The remainder of this paper is structured as follows: Section 2 provides an overview of how dwell time delays arise and insight into current dwell time scheduling and analysis techniques. In Section 3, we describe the Rasch analysis technique along with the steps we took to adapt dwell time data to fit the Rasch analysis technique. Section 4 provides an overview of our case study. In Section 5, we show the results, both in terms of the applicability of the method as well as how to use the output from a Rasch analysis in relation to dwell time evaluations. The discussion and conclusion are provided in Section 6 and Section 7, respectively.

Literature review

Dwell times

Dwell times refer to the time a train is stationary at a station and are measured as the difference between the arrival and departure times of a train (Li et al., 2014). Although we often speak of dwell time as a single process, it consists of several subprocesses (Buchmueller et al., 2008; Goverde, 2005). A schematic overview of the dwell time process is shown in Fig. 1. The dwell time process includes both static and dynamic time elements (Seriani et al., 2019). The static and dynamic time elements are indicated with the different colours in Fig. 1. The static time elements are governed by the technical aspects of the railway system, these being the time it takes for the doors to open and close as well as the dispatching time. These elements are relatively easy to schedule, as the time required to complete these steps can be seen as consistent. The dynamic time element of dwell times consists of the time needed to complete the boarding and alighting process and the arrival time of a train, or to be more precise, the arrival punctuality. The impact of the arrival punctuality of a train on the duration of dwell times has been highlighted in the past (Coulaud et al., 2023; Kecman and Goverde, 2015), where dwell times are longer for trains that arrive early. In fact, this can be attributed to trains having to wait until their departure time when arriving early, thus having an extended dwell time.

The other dynamic time element, the boarding and alighting time, is influenced by the volume and behaviour of passengers during the boarding and alighting process. As passenger volumes increase, so does the time it takes for the boarding and alighting process to be finished. This is; however, not a simple linear relationship (Kuipers and Palmqvist, 2022), making it hard to schedule even when passenger volumes are known. In addition to the volume of passengers, studies found the spread of passengers (Oliveira et al., 2019), the ratio between boarding and alighting passengers (Seriani et al., 2019), and the way in which passengers behave during the boarding and alighting process to affect the time it takes for passengers to alight and board. Passengers are, for example, more likely to show antisocial behaviour when boarding a train in situations where the platforms become crowded in order to secure a seat (Hirsch and Thompson, 2014). This results in passengers starting to board before the alighting process is completed. The subsequent friction between both alighting and boarding flows of passengers slows down the process (Harris, 2005), meaning more time is needed for its completion.

Dwell time scheduling

When scheduling a dwell time, it is important to consider setting an achievable dwell time, as small delays in dwell time can affect service capacity significantly. Delays in dwell time can result in a bunching effect, which is when the following train must wait outside the station until the preceding train leaves the platform. This also leads to passenger journey time delays (Krause, 2014; Ding, 2016; Wu et al., 2018). Although several studies have highlighted how the dwell time process is rather dynamic in its nature, current dwell time scheduling practices are still mostly based on generalized assumptions and rules of thumb (Christoforou et al., 2020; Palmqvist, 2019). Swedish timetable practices commonly set dwell time to approximately one minute at most stations and two minutes when a larger volume of passengers is expected (Palmqvist, 2019), for example. This static approach to dwell time scheduling results in actual dwell times exceeding scheduled dwell times on a regular basis (Goverde and Hansen, 2001; Nash et al., 2006). This is whilst statistical models to predict the dwell time for scheduling purposes exist, of which regression analysis is the most commonly used

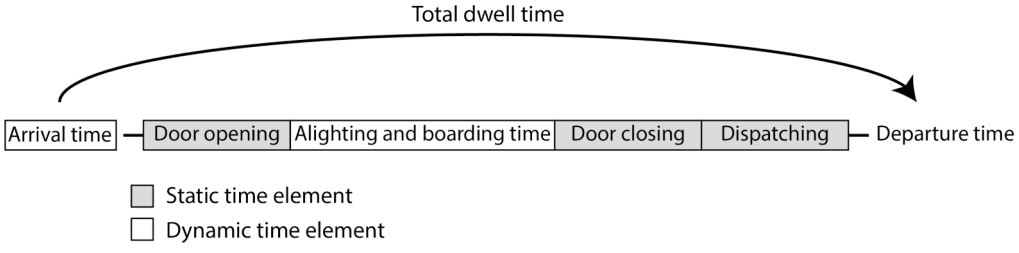


Fig. 1. Schematic overview of the dwell time process, with the colours indicating the static and dynamic time elements making up the total dwelling process.

method (Yang et al., 2019). However, the practical application of such models is rare (Volovski et al., 2021).

Quantifying dwell time delays

In order to improve dwell time scheduling principles is it important to quantify the problem of dwell time delays, which is not a straightforward task. Dwell time delays are relatively small in nature making them difficult to quantify using commonly used delay metrics. In Sweden, for example, 80 % of the delays at stations are less than three minutes (Palmqvist, 2019). This means that dwell time delays rarely exceed the national threshold for delays, which is set at five minutes and fifty-nine seconds in Sweden, and their actual impact and frequency are likely to remain hidden within the normal delay metrics. Several ways have been proposed to quantify dwell time delays more accurately. In their study on travel-time reliability, van Loon et al. (2011) provide several indicators to measure punctuality. The authors mention using the probability of a train arriving with a delay of more than three or nine minutes, measured at the final station, the average delay minutes, the standard deviation of the arrival and departure time, and using the difference between the 80th and 50th percentile of the arrival and departure time distribution. More specifically related to dwell times, Gysin (2018) mentions the use of the mean and standard deviation, and Christoforou et al. (2020) state using the ratio between the mean and 95th percentile of the observed dwell times. Tortainchai (2023) uses a risk-based evaluation approach to evaluate the impacts of different amounts of dwell time delay. In addition to defining a more accurate metric to measure dwell time delays, distribution models have been widely used. Such models enable an insight into the variability of the different aspects of dwell times. In this case, random processes are established by fitting observation data with a theoretical distribution model, the fitted model can then be used to perform analyses regarding dwell time delays (Lessan et al., 2018; Yuan et al., 2010). It is worth noting that most of the proposed metrics assess dwell times from the point of view of either individual stations or individual services, whilst it is likely that a combined impact of stations and services can influence the amount of dwell time delays.

Analysing dwell time performance within a network

So far, we have shown how scheduling dwell time is a non-trivial task due to the dynamic and stochastic nature of the dwelling process, and how dwell time delays are often not captured in commonly used delay metrics. In order to improve the scheduling regime of dwell times, it is important to take the actual situation into account. Scheduling dwell times that accurately reflect the necessary time can help to improve both punctuality and the effective use of the available station capacity. Doing so requires a deeper understanding with regard to the current situation. It is, however, not practical or viable to measure and then schedule dwell time for each station and each service individually. It is, therefore, of more interest to identify subsets of services or stations that perform in a similar pattern. It allows planners to use a similar approach to dwell time scheduling across a subset of the stations and services within a network.

Several different approaches to classifying or grouping stations can be found in the literature. These range from rather simplistic approaches where stations are grouped based on their function in a network to more complex clustering techniques. An example of the former is the categorization used in Switzerland, where stations range from so-called national traffic hubs which have the highest importance to more regional stations (SVI, 2013). Such an approach does, however, not take into account the actual situation at each station as the categorization is merely based on the function desired and assumed by the operator. A different approach was proposed by Havlena et al. (2014). The researchers made use of a point-based scoring system to rank stations to determine their importance within the Czech railway network.

A more common approach to group stations found in the literature is the use of clustering algorithms. Zemp et al. (2011), for example, performed a cluster analysis on data from the Swiss railway network with a special focus on both the demand and conditions at the stations. Their study identified seven different station classes, differing in use and passenger volumes. A similar approach was used by Stoilova and Nikolova (2017), who made use of clustering analysis to classify stations in Bulgaria based on the area in which the station is located and the characteristics of passengers. The area in which a station is located was also used by Reusser et al. (2008), who distinguished stations in terms of their connectedness with other places, also known as the node, and the possible activities around the stations, known as the place. Recently, improvements in the collection of passenger flow data through automatic passenger counts of fare card data have allowed consideration of passenger flow characteristics when analysing railway stations. Zhou et al. (2022), for example, clustered and classified metro stations in Beijing based on the criticality of the network based on ridership. However, these clustering techniques are not suitable for taking into account different types of data, for example, passenger demand and service performance data.

To somewhat overcome this issue, some studies have made use of Data Envelopment Analysis approaches. Tortainchai et al. (2022), for example, studied the relative performance of stations in London based on the volume and ratio of boarding and alighting passengers. Another example is the study by Khadem Sameni et al. (2016), who determined the relative technical efficiency and service effectiveness of stations in Great Britain, taking into account the numbers of passenger entries and exits. However, while Data Envelopment Analyses and other data frontier approaches can evaluate the efficiency of each factor, they cannot represent interactions between different factors, which could, in the case of analysis of dwell time performance, include both stations and train services. Since dwell times vary according to both stations and services, both of these factors need to be considered in timetabling. There is thus a need for a tool that can consider both station and service performance simultaneously. The method proposed in this study is the use of a Rasch analysis technique.

The Rasch analysis technique

The Rasch analysis technique, initially developed by Rasch (1980), is a model which is often used within the domain of item response theory, referring to a set of statistical models that allow to model the relationship between item responses and a latent variable (Zheng and Rabe-Hesketh, 2007). It is a tool to create accurate and reliable measurements which can measure individual items, regardless of their characteristics, therefore Rasch analysis assumes a latent unidimensionality trait (Combrinck, 2020). Another benefit of Rasch over other classical measure theories is that it can predict the probability of a particular ability level getting a certain difficulty level (Hambleton and Cook, 1997). The Rasch analysis technique was developed as a psychometric technique and has been widely used in medical and health sciences (see for example: (Gothwal et al., 2009; Hart and Wright, 2002; Jette et al., 2002; Lamoureux et al., 2006; Massof and Rubin, 2001; Pearce et al., 2011; Pesudovs et al., 2003; Prieto et al., 2003; Turano et al., 1999). For an example of a more in-depth overview of the use of the Rasch analysis technique in the domain of healthcare, we refer to the recent work done by Stolt et al. (2022) who identified 88 papers using the Rasch analysis technique in nursing research.

In the field of transport research, the Rasch analysis technique has mostly been used to assess the interaction between people and the transport system. It is a valuable tool for measuring passenger satisfaction in public transport systems. The model provides a probabilistic framework to convert ordinal raw-score data into a linear scale by converting scores into logits, enhancing the understanding of latent traits like satisfaction to be compared on the same scale. Gallo (2011) developed a Rasch analysis model to evaluate passenger satisfaction in

public transport and utilises Analysis of Means (ANOM) to study satisfaction levels among different passenger groups, which are characterised based on age, profession, and sex. Cheng and Chen (2015) applied the Rasch analysis technique to assess the accessibility of two cities in Taiwan, for example, while Kim et al. (2018) used the Rasch analysis technique to analyse urban transit interchanges. Another example of the use of the Rasch analysis technique within the field of transportation is the study by Chan (2018), who developed an instrument to measure the ability of passengers with low vision and limited mobility to use public transport. In this research, Rasch analysis considers that different variables vary in difficulty and compares the difficulty of each item with respect to the passenger's vision ability.

In contrast to the common use of the Rasch analysis technique, we propose its use in an operational context. Rasch analysis is beneficial in developing new scales or modifying existing ones, offering a method to handle ratio or quantitative data effectively for precise measurement outcomes and providing linear scores, enabling straightforward comparison of measures (Combrinck, 2020). Here, we apply a Rasch analysis technique to dwell time data, focusing on the distribution of dwell time delays for different train services at different stations within the same line. In this case, train services and stations are regarded as persons and items, respectively. We thus look at the success or failure of a train service to dwell within its scheduled time at a given station and compare this to the difficulty of all train services on the line to dwell within their scheduled time at said station.

Benefits of the Rasch analysis technique

A benefit of using the Rasch analysis technique over some of the methods mentioned in Section 2.4 is that it allows us to study the dwell time performance of train services and stations in combination with one another on the same linear dimension. In contrast to other approaches that assume different items with the same weight or difficulty, a Rasch analysis shows the probability that a service with a certain level of performance will achieve the scheduled time for specific station difficulty levels. This means that it is possible to not only identify problematic train services or stations separately but also jointly examine stations at which these services perform poorly, and vice versa. This is rather cumbersome when using other clustering methods, as they often only allow clustering along a single dimension, which is either the relative dwell time performance of the train services or that of the railway stations. Being able to assess both service and station performance in a single dimension is relevant for timetabling since the allocation of dwell times needs to consider both train services and stations simultaneously. In addition, the Rasch analysis technique evaluates the "goodness of fit" between station performance and station difficulty, thus serving as a criterion for assessing the structure of responses rather than solely providing a statistical description of the responses (Kim et al., 2018).

Furthermore, the Rasch analysis technique can account for any unequal difficulty across the test items (Boone, 2016), or stations in our case, rather than directly comparing raw scores, such as the average dwell time at a station. As Brush and Soutar (2022) state, Rasch scores should work the same way across different ethnicities, demographics, or levels of experience. This is relevant since different station characteristics, such as differences in platform layout, can influence dwell times differently and comparing raw test scores would not account for this effect.

In addition to this, the Rasch analysis allows us to go beyond a measure of central tendency to group stations and services in terms of their dwell time performance. Using a measure of central tendency, for example, can lead to incorrect assumptions because it does not account for the underlying relationship between service performance and station difficulty. In addition, outliers can skew the outcome of such analyses, leading to wrongful conclusions. While some statistical methods, such as an Analysis of Variance (ANOVA) or a Kruskal-Wallis test, can compare dwell times across multiple stations and service types, such methods still rely on the difference between either the mean or median.

Contributions of this study

While the Rasch analysis technique has been widely applied outside of the domain of transportation, its use within the domain of transportation is somewhat limited. Furthermore, most of these studies applied the Rasch analysis technique to study the interaction between people and the transport system. So far, the Rasch analysis technique has, to the best of our knowledge, not been used within an operational context. The contribution of this paper to knowledge is as follows:

1) Showing the applicability of the Rasch analysis technique within an operational context with a specific focus on dwell time evaluation for commuter trains.
2) Highlight how the output of a Rasch analysis can be used to investigate the dwell time performance of stations and services on a line level.

Data availability

Case study description

To study the applicability of the Rasch analysis technique for dwell time analyses we made use of a case study consisting of data from both Sweden and the UK. It should be noted that the Rasch analysis is applied to two datasets with different systems to test the applicability of the Rasch analysis technique for dwell time data. The analysis thus does not provide a comparison between both cases. A direct comparison between both systems based on the output from the Rasch analysis is not possible since the Rasch scores show the relative score within each system and not across both systems.

For the Swedish dataset, we consider the service between Helsingborg Central Station and Trelleborg Central Station in the southern region of Scania during the morning peaks on weekdays. The data is taken from the 1st of January to the 31st of December 2019. The data originates from the onboard system for the commuter trains on this line, recording the dwell time in a magnitude of seconds. Dwell times are scheduled to be 60 s at most stations on this line with the exception of two larger stations, these being Lund and Malmö central station, where dwell times are set at 120 s. The service runs at an hourly interval through the morning peak, with the exception of Monday and Friday during which additional train services are operated. The data included information on 941 train services.

For the UK dataset, we consider the service from Shenfield to London Liverpool Street of London's Elizabeth line. We also make use of the dwell times observed during the morning peak on weekdays. The data is taken from 13th May 2019 to 20th November 2019. The data included 897 train services passing through 12 stations. Most stations in the UK dataset have a scheduled dwell time of 30 s, with two major stations having a scheduled dwell time of 60 s, these being Stratford and Ilford.

Overview of the data

The variable of interest for the study presented here is the dwell time deviation, i.e., the difference between the scheduled and actual dwell time. Table 1 shows some descriptive statistics for the dwell time deviations for both the Swedish and UK data. In addition to this, a frequency distribution of the dwell time delays for the Swedish (left side) and UK data (right side) is shown in Fig. 2. In the Swedish case, we can observe that stations with longer scheduled dwell times are found to have shorter dwell time deviations on average but with a larger standard deviation. In the UK, this trend is reversed, where stations with short scheduled dwell times have shorter delays on average, with a smaller standard deviation. Looking at Fig. 2, we can observe that dwell time

Table 1
Descriptive statistics for the dwell time deviation (in seconds) for the Swedish and UK datasets.

	Swedish data		UK data	
	Scheduled dwell time: 60 s *(n = 13365)*	Scheduled dwell time: 120 s *(n = 2020)*	Scheduled dwell time: 30 s *(n = 8970)*	Scheduled dwell time: 60 s *(n = 1796)*
Mean	20 s	34 s	31 s	49 s
Standard deviation	36 s	28 s	20 s	25 s
Median	15 s	33 s	31 s	52 s

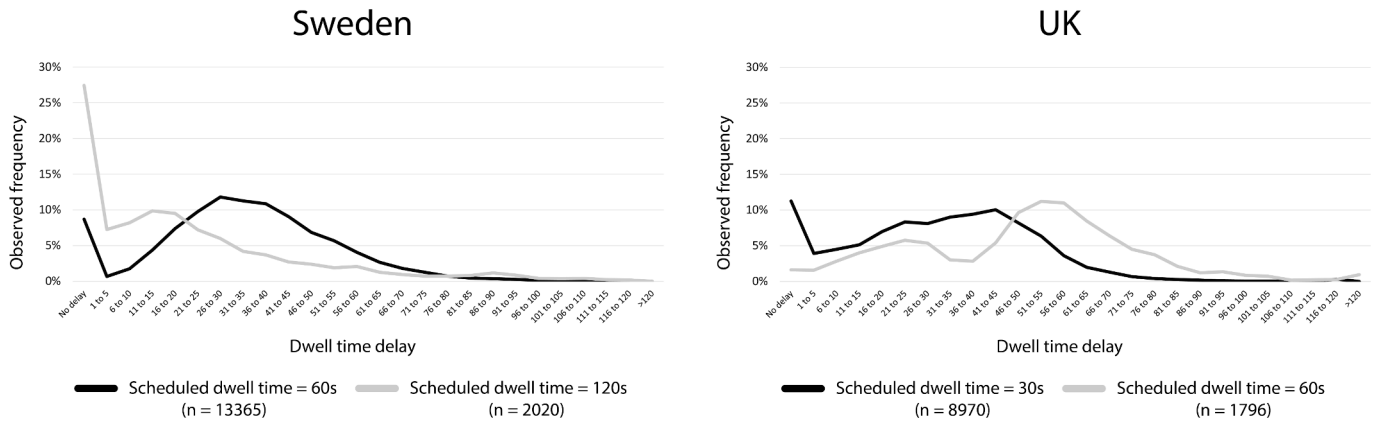


Fig. 2. Frequency distribution of dwell time delays for Sweden (left) and the UK (right). Dwell time delays are grouped in steps of 5 s for legibility reasons. No delay indicates dwell time delays of 0 s or less. Frequency distributions are split based on scheduled dwell times.

delays in Sweden are relatively small, which is in line with the mean dwell time deviation shown in Table 2. Delays of more than 90 s (i.e., a minute and a half) are not common here. In the case of the UK, we can observe that relatively large dwell time delays are more common, especially for stations with a scheduled dwell time of 60 s, where the most common delay is found to be around 60 s (i.e., one minute).

Method

Rasch analysis technique

In the Rasch model, shown in Eq. (1), respondents are referred to as persons, and the tasks they undertake are referred to as items. The model is used to calculate interval estimates that represent person locations (i.e., the ability of a person represented by θ_n) and item locations (i.e., the general item difficulty to perform tasks represented by δ_i) on a linear scale or dimension. In the simplest form of the Rasch model, the response to an item is the dependent variable, and the ability of a person and the difficulty of a test item are the independent variables (McCamey, 2014). The relationships between observed responses and underlying latent traits are thus defined based on the scores for the *person ability* and *item difficulty,* which are obtained from the Rasch model. These estimates are measured in logits, which are calculated by taking the natural logarithms of the odds ratios of success or failure when a person attempts an item. This means that persons and items are assigned a score measured on the same scale that represents the latent variable, allowing much easier comparison (Cappelleri et al., 2014). In contrast to other item response models, the Rasch analysis can be seen as prescriptive, where it asks the data to fit the model rather than the model to fit the data (Tesio et al., 2024). The Rasch model was first developed to be used with dichotomous variables but was later generalised to make use of polytomous variables by Andrich (1978) and Masters (1982).

$$ln\left(\frac{P_{n_i(x_i=k)}}{P_{n_i(x_i=k-1)}}\right) = \theta_n - \delta_i - \tau_{ik} \quad (1)$$

Where:

θ_n: The capability of service n capability to dwell within its scheduled dwell time.
δ_i: The difficulty for all trains to dwell within their scheduled dwell time at station i.
τ_{ik}: Thresholds for station I for a correct or positive response to level k.
$P_{n_i(x_i=k)}$: Probability of service n at station i to achieve a correct or positive response to the level k, i.e. to dwell within its scheduled dwell time.

Table 2
Labelling regime to transform continuous dwell time data to polytomous data for use in the Rasch analysis technique for both the Swedish and UK data.

Swedish Data		UK Data	
Adjusted dwell time deviation	Label	Adjusted dwell time deviation	Label
≤0 s	5	≤0 s	5
>1 and ≤30 s	4	>1 and ≤15 s	4
>31 and ≤60 s	3	>16 and ≤30 s	3
>61 and ≤90 s	2	>31 and ≤45 s	2
>91 and ≤120 s	1	>46 and ≤60 s	1
>120 s	0	>60 s	0

Implementing the Rasch model for dwell time deviations

In the study presented here, we apply the Rasch model to analyse dwell time deviations. Using the Rasch analysis to study dwell time deviations required some additional steps compared to the more common application of the Rasch analysis technique, which is to study the ability of people based on questionnaire data. The first step is a change to the definition of the parameters included in the model. The parameters in the classical applications of the Rasch analysis are defined as "person ability" and "item difficulty" as shown in Eq. (1). In our case, the person is defined by the train services, and the item is defined by the

station. In our case, we define θ_n (service performance) as the ability of service n to dwell within its scheduled dwell time, and δ_i (station difficulty) as the difficulty for all trains to dwell within their scheduled dwell time at station i. $P_{n_i(x_i=k)}$ is then defined as the probability of service n at station i to achieve a correct or positive response to the level k, i.e., to dwell within its scheduled dwell time. This probability is a direct result of the difference between the ability of a service and the difficulty of stations, where the larger the ability level is compared to the difficulty, the larger the probability of a successful response. In our case, this means that the probability of a train service having a dwell time closer to the scheduled dwell time at a given station is thus larger for train services for which the performance is better than the difficulty of a given station, whereas the opposite is true for services where the performance is worse than the difficulty of a given station. This concept is central to the Rasch model.

Data preparation

The variable of interest for the Rasch model is the dwell time deviation of a service at a station, which we use as an indicator of dwell time performance. Eq. (2) calculates dwell time deviations, defining them as the difference between the scheduled and actual dwell time, which can have both negative and positive values. A negative deviation indicates that the actual dwell time is shorter than scheduled and can occur when a train arrives with a delay and can thus dwell shorter than scheduled. A dwell time delay occurs when the deviation is positive, indicating that the train dwelled for longer than scheduled. We make use of the dwell time deviation rather than the total dwell time to better show dwell time performance, as it is an indication of the accuracy of the scheduled dwell times with respect to the actual dwell times. The dwell time deviation is calculated using the following formula:

$$Dwelltimedeviation = SD - (T_{departure} - T_{arrival}) \quad (2)$$

Where:

SD : The scheduled dwell time at a given station.
$T_{departure}$: The departure time of a given train at a given station.
$T_{arrival}$: The arrival time of a given train at a given station.

When making use of dwell time deviations as a measure of performance, it is important to correct for early arriving trains. Early arriving trains, indicated by having a negative arrival delay, have to dwell longer than scheduled whilst waiting for their scheduled departure time. This does not consider a dwell time delay and therefore requires correction. Here, we do so by following the protocol shown in Fig. 3. Omitting this correction can lead to an overestimation of the size and frequency of dwell time delays. The adjusted dwell time deviation is calculated as follows:

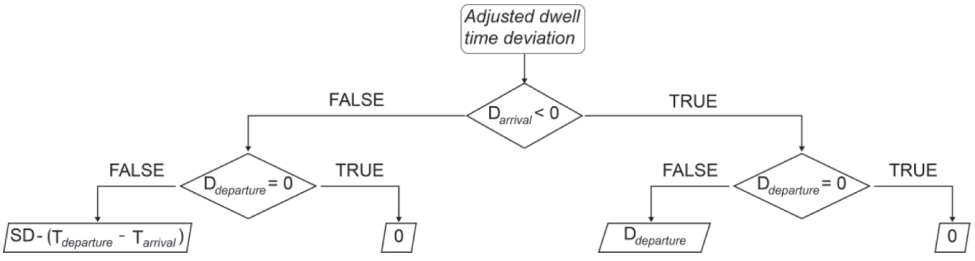
Where:

SD : The scheduled dwell time at a given station.
$T_{departure}$: The departure time of a given train at a given station.
$T_{arrival}$: The arrival time of a given train at a given station.
$D_{arrival}$: The arrival delay of a given train at a given station
$D_{departure}$: The departure delay of a given train at a given station

The next step of the data preparation consisted of limiting the number of dwell time deviations to be included in the analysis. In line with Pritchard et al. (2021), we limit the delay size to a maximum of 180 s. By setting this upper limit for the dwell time deviation, we focus on cases that can be seen as normal deviations. Although this means that extreme cases are excluded, we deem those cases to be of lesser interest with regard to scheduling dwell times due to the rare occasion in which they occur.

The final step of the data preparation consists of converting the continuous dwell time deviation data to polytomous data. Although using dwell time deviations provides a good metric to judge the dwell time performance of a station or service, the data cannot be directly used for the Rasch analysis. This is because the Rasch analysis technique requires polytomous data, such as responses on a Likert scale, rather than continuous data as input. To convert the continuous dwell time deviation data to polytomous data, we labelled the size of the deviation in either 30-second intervals (as is the case for the Swedish data) or 15-second intervals (as is the case for the data from the UK). The labels are shown in Table 2. The size of the buckets was chosen to reflect both operational and local constraints while still having sufficient granularity to capture the nature of dwell time deviations. We determined the latter of these through an iterative process, evaluating different bucket sizes based on model outputs and fit statistics.

The scheduling regime in Sweden makes use of steps of 60 s, and in the UK, the schedule makes use of steps of 30 s. These sizes were initially used as bucket sizes, but it was found that using these steps resulted in too few labels to be used for the Rasch analysis technique. The number of labels typically ranges from three to six labels with more labels being better. Another option was to use bucket sizes of 5 s, but this resulted in too many labels and does not accurately reflect scheduling principles as changes of such magnitude cannot be made when scheduling dwell times. Bucket sizes of 30 and 15 s for the Swedish and UK data respectively, were found to both result in sufficient labels to be used for the Rasch analysis while still resembling actual railway scheduling practices in both countries.



Where:

SD:	The scheduled dwell time at a given station.
$T_{departure}$:	The departure time of a given train at a given station.
$T_{arrival}$:	The arrival time of a given train at a given station.
$D_{arrival}$:	The arrival delay of a given train at a given station
$D_{departure}$:	The departure delay of a given train at a given station

Fig. 3. Flowchart of the protocol to correct the dwell time deviation based on the arrival punctuality of a train.

Performing the Rasch analysis

We use the Partial Credit Model (PCM), also known as the Rating Scale Model, for the Rasch analysis due to the conversion of continuous dwell time deviation data into polytomous data (Masters, 1982). In order to run the Partial Credit Model, we employed the PCM function within the eRM package in R. This package uses conditional maximum likelihood estimation to find the model's parameters for polytomous item responses (Mair et al., 2021). For a more in-depth discussion on the use of conditional maximum likelihood estimations for the Rasch analysis, we refer to the work done by Mair and Hatzinger (2007). The input for the PCM function in the eRM package consists of a matrix with the rows representing the dwell time of the services and the columns representing the stations (Mair et al., 2021). An example of this is shown in Table 3. The continuous dwell time deviation for a given service at a given station was transformed into polytomous data using the labelling regime shown in Table 2.

Results

Model fit assessment

Two metrics are used to assess the model fit. The first metric uses the separation and reliability scores for the person (service in our case) and item (station in our case). These scores are used to assess the ability of the model to differentiate between trains with different service ability levels (i.e., the ability of a service to dwell within its scheduled time at all stations) and stations with varying levels of difficulty (i.e., the difficulty for trains to dwell within the scheduled time at that station). It is an acceptable level of separation to consider the minimum required to divide the sample into two distinct strata i.e., low, and high ability (Souza et al., 2018). The reliability or separation index indicates the consistency of ranking relative to person and item location (Cappelleri et al., 2014). A high reliability of persons or items means that there is a high probability that persons or items estimated with high measures do have higher measures than persons or items estimated with low measures (Linacre, 1997).

A *service separation score* of less than two indicates that the developed model may not be sensitive enough to distinguish between long and short dwell time deviations. A *station separation score* lower than three indicates that the sample size of services is not large enough to confirm the difficulty hierarchy. *Service and station reliability scores* of less than 0.9 reflect that the model may not be sensitive enough to discriminate between the variances of the services and station difficulty, the length of the rating scale, and the number of categories per item (Bond and Fox, 2007). The service reliability score is found to be 0.99, and the separation score is 9.94 in both models. These scores indicate that both models are sensitive enough to distinguish between high and low performers and that the sample is large enough to confirm the item difficulty hierarchy. The high service and station reliability scores of 0.99 for both the Swedish and UK models reflect that the wide range of stations includes different capability levels and that appropriate difficulty levels were present in the model.

The second metric to assess model fit makes use of the Mean Square statistics from the model output. Fit statistics, defined by infit and outfit, are analysed to highlight any unexpected participant responses (Bond and Fox, 2007). Infit reflects the difference between observed and expected responses for those items that have a difficulty level near the person's ability level. Outfit includes the differences for all items, irrespective of how far away the item difficulty is from the person's ability (Linacre, 2002; Tennant and Conaghan, 2007). Fit statistics are generated for both items and persons. Both infit and outfit are expressed in the form of mean-square fit statistics (MSQ) which is the chi-square statistic divided by its degrees of freedom.

The in and outfit scores are shown in Fig. 4, with the triangular markers indicating the data points for the stations, and the data points for the services are shown as dots. The metric reflects the randomness present in the data, with an expected value close to 1.0. Generally, mean square values ranging from 0.5 to 1.5 are considered acceptable. Values lower than 0.5 and between 1.5 and 2 are noted to not result in an improvement or a decline in model performance. Values larger than 2 should be treated as troublesome (Tran et al., 2018).

As we can see in Fig. 4, all MSQ values for the station fit are within the ideal bandwidth. This, however, is not the case with the service fit statistics. In the Swedish data, we find that 15 % of the observations for services have an MSQ value greater than 2. In the case of the data from commuter trains in the UK, this value is slightly larger, with 19 % of the observations having an MSQ value of more than 2. This indicates that not all observations for the service scores fall within the bandwidth of what is considered to be a good model fit. This is, however, only a small portion of the observations, and the overall model fit is still deemed to be good. A further analysis was conducted to understand the cases that fall outside of the desired bandwidth, but no clear pattern in these observations was found.

Person-Item maps

The output of Rasch analysis is called a person-item map or Wright map, which represents each station in relation to its train dwell time. The map provides both person measures and item measures on the same linear scale so that researchers can determine how well the test items (stations in our case) are distributed regarding the ability level of test takers (trains in our case) (Boone, 2016). Fig. 5 depicts the person-item map for the stations and services of the Swedish commuter train and the person-item map for the UK is shown in Fig. 6. The person-item map shows the location of the service performance and station difficulty along the same latent dimension, in this case, the dwell time punctuality. The distribution of the service ability is shown in the histogram at the top of the figure and the station difficulty is shown on the bottom half of the figures where dark circles show the location of station difficulties, and the thresholds of adjacent categories are depicted with the open circles. In this case, a higher score in the histogram shown on top indicates a better ability of a train service to dwell close to its scheduled dwell time. The ability and difficulty are measured in logits, meaning that the scales in the person-item map are additive. Services located at a logit score of 0 have average ability, whereas services located on the right-hand side of the distribution show better ability, and vice versa for services on the left-hand side of the distribution. The opposite is true for the station difficulty scores, where a higher score indicates that a train is more likely to suffer a delay at that station.

Concerning the model fit, we can observe a spread in the data points for both the stations and services in the person-item map for both the Swedish and UK models, suggesting that there were variations in dwell

Table 3
Example of the data matrix used as input for the Rasch analysis, with the different services on each row and the different stations in each column. Values represent the converted observed dwell time deviation for each service at each station.

	Station A	Station B	Station C	Station D	Station E	Station F	Station G
Service 1	0	2	3	5	2	0	3
Service 2	3	5	1	3	0	1	2
Service 3	1	0	2	4	2	4	3
Service 4	1	4	3	5	0	0	1

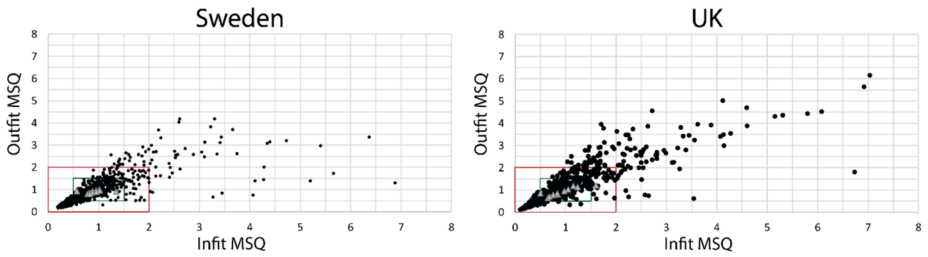


Fig. 4. Infit and outfit Mean Square statistics scores for Rasch model based on Swedish (left) and UK (right) commuter train data. The triangular markers represent the station data points, while the dots represent the service data points. The colored boxes show the threshold for model fit assessment, with the green box showing the range in which data points are not improving or declining model performance and the red box showing the threshold for troublesome values.

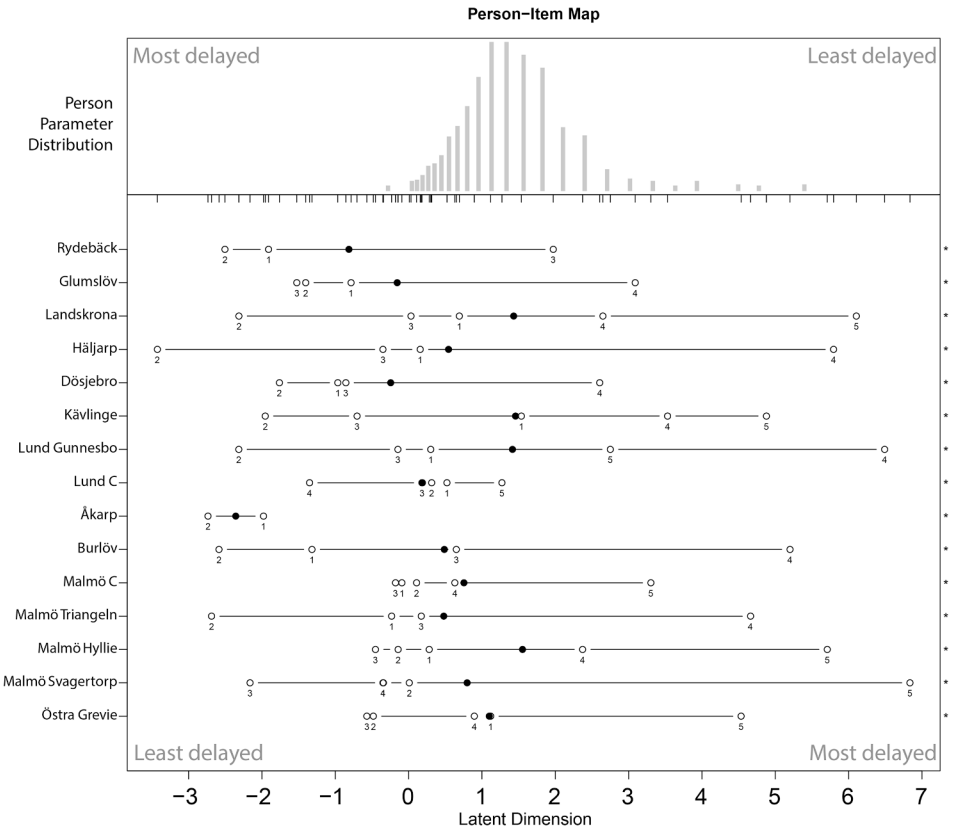


Fig. 5. Person-Item map for the Swedish commuter train data showing the service ability and station difficulty along the latent variable, this being the dwell time punctuality. The histogram on top shows the distribution of the service ability scores from most delayed (left) to least delayed (right). The bottom of the figure shows the station difficulty scores where the black dots indicate the location of the service difficulty and the thresholds of the adjacent categories are depicted with open circles.

time performance within both that were captured by the model. The spread found in the data point also suggests that the bucket sizes used to convert the continuous data to polytomous data have a sufficient level of granularity to reflect the changes in dwell times in both the Swedish and UK models.

Service performance scores

The following can be noted considering the service performance scores derived from the Rasch analysis. As can be seen in the histograms at the top of both Fig. 5 and Fig. 6, most service performance scores are above 0, indicating an above-average ability to adhere to the scheduled dwell times. In both the Swedish and UK cases, there are only a few services on the most right-hand side of the distribution; these services showed the greatest ability to adhere to the scheduled dwell time. The same is true for the frequency of services on the most left-hand side of the distribution, thus indicating that only a few services perform very poorly.

To understand differences in the service performance scores between services, Fig. 7 shows the scores for each operational service in both the Swedish and UK datasets. As expected, the services do not perform in a uniform manner in terms of the service performance score. In the case of commuter trains in Sweden, it can be observed that the median service performance score is relatively similar across the board, while the variability shows a stronger difference between services. The Rasch analysis hypothesizes that the average measure for all item parameters is fixed at zero logits, making this the comparative basis for interval scales (Kim et al., 2018). The median service performance score across all observations is 1.33 (IQR 0.79–1.56), indicating that most services run with good performance where the actual dwell time is close to the scheduled dwell time. No extreme negative values are found for the service performance scores in Sweden, indicating that it is not likely for

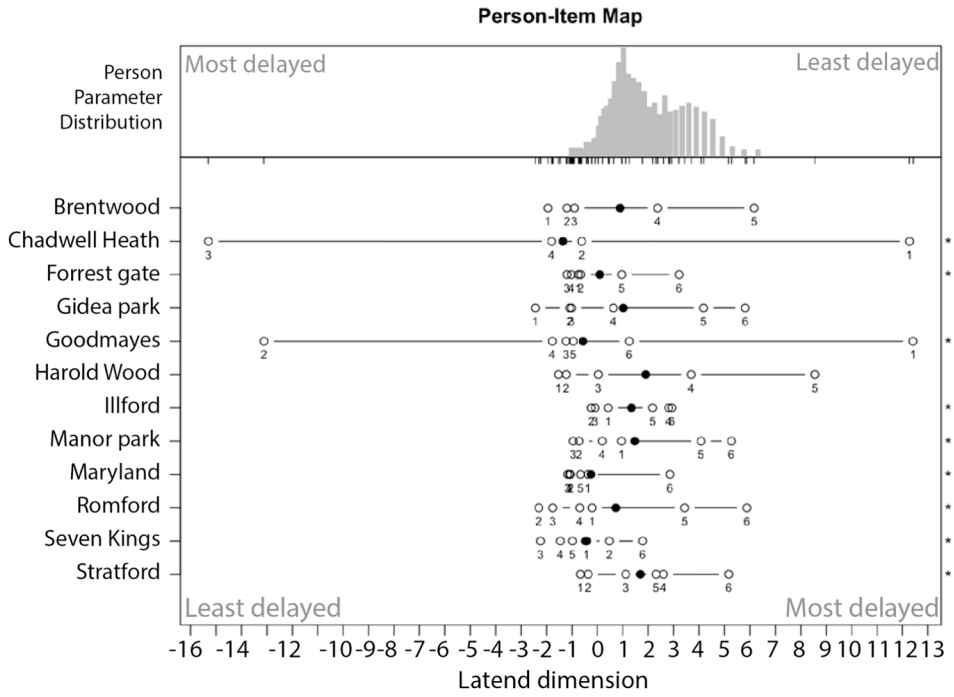


Fig. 6. Person-Item map for the UK commuter train data showing the service ability and station difficulty along the latent variable, this being the dwell time punctuality. The histogram on top shows the distribution of the service ability scores from most delayed (left) to least delayed (right). The bottom of the figure shows the station difficulty scores where the black dots indicate the location of the service difficulty and the thresholds of the adjacent categories are depicted with open circles.

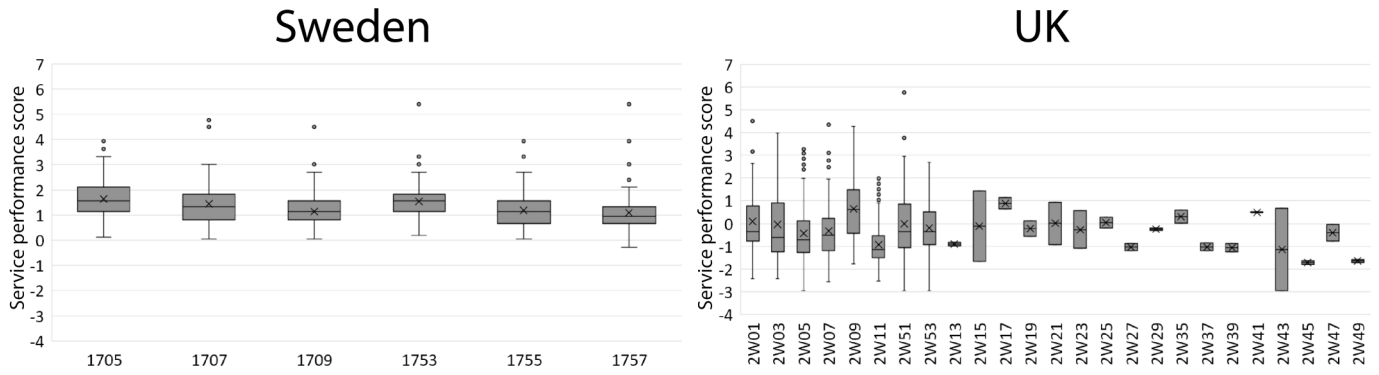


Fig. 7. Boxplots showing the service performance scores obtained from the Rasch analysis for the different services in use in Sweden (left) and the UK (right). Service performance scores reflect the ability of a service to dwell within its scheduled time and a higher score indicates a better performance, meaning that a train is less likely to incur a dwell time delay.

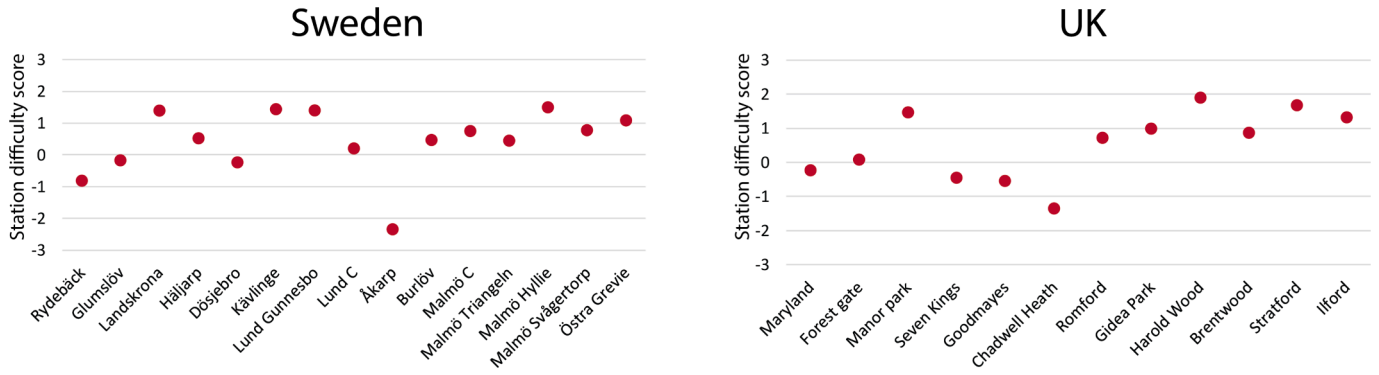


Fig. 8. Station difficulty scores obtained from the Rasch analysis for stations in Sweden (left) and the UK (right). Station difficulty scores reflect the likelihood of a service being delayed at a given station. Higher station difficulty scores indicate a greater likelihood of a train being delayed and vice versa.

a service to perform extremely poorly during the dwelling process.

The services in operation in the UK show a larger spread in terms of their respective service performance scores. The median scores are found to be negative for some services, indicating a relatively poor dwell time for these services. This is also reflected in the overall median value for the service performance score across all UK data which is 0.03 (IQR −1.03 – 1.31), indicating that a relatively sizable portion of services are likely to have a longer dwell time than what is scheduled.

Station difficulty scores

The station difficulty scores based on the output from the Rasch analysis are shown in Fig. 8. A higher difficulty score means that services are more likely to be delayed at that specific station, whereas a lower difficulty score indicates that it is less likely for services to be delayed at that station. Most stations in the Swedish case have a relatively high difficulty score, with the exception of Rydebäck, Glumslöv, Dösjebro and Åkarp station. This indicates that most stations are likely to be characterized by dwell time delays. A similar pattern is found for the stations in the UK, where all stations except for Maryland, Seven Kings, Goodmayes, and Chadwell Heath are found to have higher station difficulty scores. In both cases, we thus find some stations that perform exceptionally well relative to the other stations, as well as stations that perform quite poorly in terms of dwell times. By identifying stations with high difficulty scores, the study offers significant information for focused operations. Stations that consistently show high levels of difficulty should be given priority for infrastructure enhancements such as platform redesign or the implementation of methods to regulate passenger flow. In addition, it may be worth considering making scheduling adjustments by providing more time at certain stations to accommodate the difficulties.

Service-station performance

So far, we have shown both the service performance and station difficulty scores from the Rasch analysis. Although this provides some insights into the overall dwell time performance of the two datasets under consideration here, the real strength of the Rasch analysis technique is its ability to assess both service performance and station difficulty scores on a single dimension. To do so, it is necessary to calculate the performance of a service relative to the difficulty of a given station. This is done using the following equation (Tesio et al., 2024):

$$\mathrm{P}\,(\text{dwell time delay}) = \mathrm{f}\,(\theta - \delta) \tag{3}$$

Where:

θ : The service ability score.
δ : The station difficulty score.

This can be read as the probability P of a dwell time delay being observed as a function of the difference between the ability and difficulty (Tesio et al., 2024), in our case the service ability and station difficulty scores. We call this indicator the *dwell time performance score*. A train with a service score of 2 halting at a station with a difficulty score of 1.5 will have a dwell time performance score of 0.5 (2–1.5 = 0.5), for example. Where the outcome of Eq. (3) shows the probability of a dwell time delay (a range between 0 and 1), the difference between the service ability and station difficulty can take on values between – infinity to +infinity. The Rasch analysis, therefore, makes use of logits and Eq. (3) takes the following form in a Rasch analysis (Tesio et al., 2024):

$$\ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \theta - \delta = \text{logit} \tag{4}$$

Where:

θ : The service ability score.
δ : The station difficulty score.

The dwell time performance score can be interpreted as follows: a higher dwell time performance score indicates situations where actual dwell times are closer to the scheduled dwell time, whereas a lower dwell time performance score indicates the opposite. As Tesio et al. (2024) state, it is important to note that a change in logits remains invariant across its span and works the same as changes in degree centigrade or changes in kilograms. This means that a reduction from 3 to 1 logit is similar to a reduction from 6 to 4 logit and a change in dwell time score from 1 to 2 is the same as a change from 3 to 4.

The dwell time performance scores are visually represented using a heat map for both the Swedish and UK cases, shown in Figs. 9 and 10 respectively. For legibility reasons, we only display aggregated values for the unique service scores rather than each individual service or all possible service scores. The presented heat maps can be used to visually identify hotspots for both poor and good dwell time performances along each line for a given service performance score. With regards to the dwell time performance score for the Swedish case, we can observe that Rydebäck and Åkarp are stations that perform well across the board and Chadwell Heath is the best-performing station in the UK case. Most dwell time performance scores are above zero at these stations, indicating it is more likely that services halting at those stations have a dwell time close to the scheduled time. We can also observe that there is no such thing as a station for which all services are likely to have a dwell time that is extremely long compared to the scheduled dwell time, given that all stations have at least one instance where the dwell time performance score is relatively high.

It is also worth noting that the majority of the services perform relatively well across the entire line. Even the worst performing services, those with a service score of −0.5 or lower, have a dwell time performance score of zero or higher for at least one station along the line. Note that the median service performance score in the UK case was found to be 0.03. Looking at the dwell time performance scores for a service performance score of 0.03, we can observe that it is expected that a train is more likely to incur a dwell time delay at Manor Park, and at all the stations between Romford and Ilford. For the Swedish dataset, the median service performance score is 1.33. Again, looking at the dwell time performance scores for a service performance score of 1.33, we can observe that Landskrona, Kävlinge, Lund Gunnesbo, and Malmö Hyllie can be considered to be troublesome stations along the line. This example indicates how visually representing the dwell time performance scores provides insights into the performance of a service on a line level by being able to identify hotspots of poor dwell time performances.

Rasch output compared to common indicators of dwell time performance

A common indicator used to assess dwell time performance is the historical dwell time at a station. To compare the output from the Rasch analysis to the historical dwell times of all stations on the selected lines, we plot the median dwell time deviation at each station along with the station difficulty score in Fig. 11. The red dots in the figure show the station difficulty score and the grey bars indicate the median dwell time deviation observed at each station. Here we can observe that the station difficulty score follows a similar pattern to the median observed dwell time deviation for some stations. However, we can also observe cases where the median observed dwell time is positive, indicating that the actual dwell times commonly exceed the scheduled dwell time whilst the station difficulty is relatively low, thus indicating that it is less likely for a service to incur a delay at these stations.

While the pattern for the station difficulty scores thus appears to follow the median dwell time deviation values, there are some stations for which this is not the case. In the Swedish case, for example, Rydebäck has a relatively low difficulty score while the median dwell time deviation is high here. A similar observation is made for Maryland, Seven

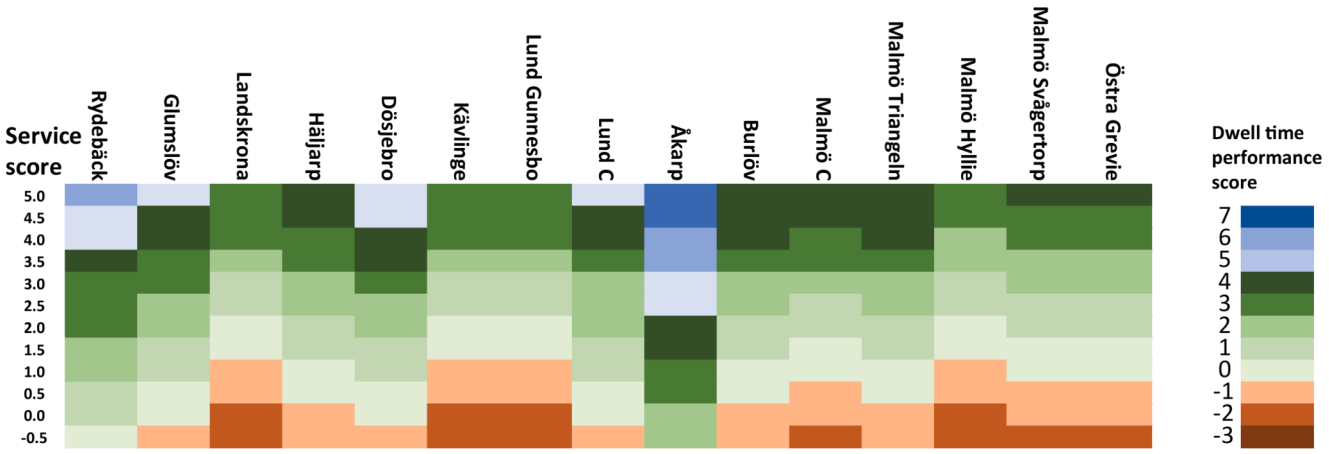


Fig. 9. Heat map showing the dwell time performance scores for the Swedish case with the service ability on the y-axis and station names on the x-axis. Station difficulty scores are omitted from the figure for clarity reasons. Dwell time performance scores were calculated by subtracting the station difficulty scores from the service performance scores. A higher dwell time performance score indicates situations where actual dwell times are closer to the scheduled dwell time, whereas a lower dwell time performance score indicates the opposite.

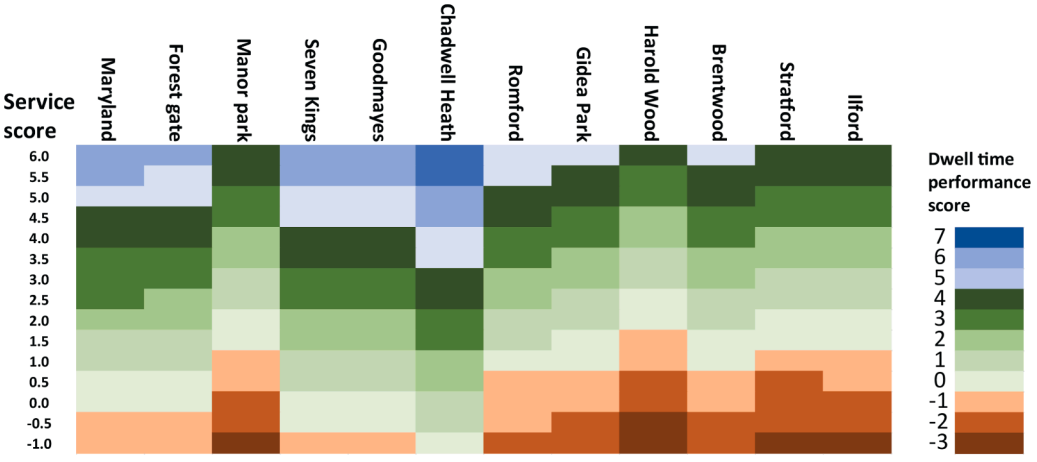


Fig. 10. Heat map showing the dwell time performance scores for the UK case with the service ability on the y-axis and station names on the x-axis. Station difficulty scores are omitted from the figure for clarity reasons. Dwell time performance scores were calculated by subtracting the station difficulty scores from the service performance scores. A higher dwell time performance score indicates situations where actual dwell times are closer to the scheduled dwell time, whereas a lower dwell time performance score indicates the opposite.

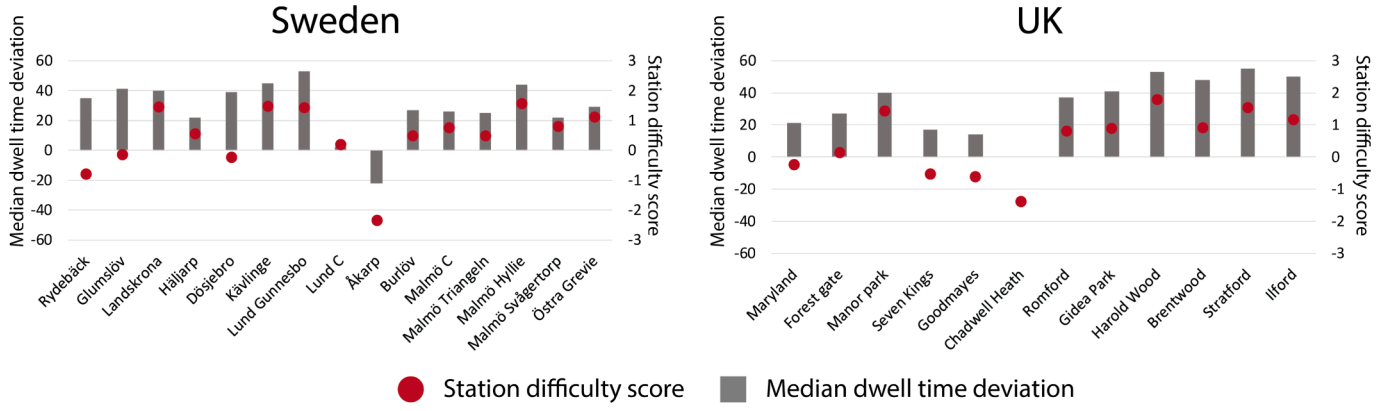


Fig. 11. Station difficulty scores (red dots) and median observed dwell time deviations (grey bars) for stations in Sweden (left) and the UK (right), stations are ordered based on their order along the line from left to right.

Kings, and Goodmayes stations in the UK. One reason for this can be due to the Rasch analysis taking into account all observations, including the possibility of services being on time at these stations. This is in contrast to the utilization of median values, which solely relies on the 50th percentile of the data. Another possible explanation is that few services are very delayed here, inflating the median delay time, while most services dwell close to the scheduled dwell time, thus lowering the station difficulty score.

In addition to the comparison to the median dwell time deviation, we compare dwell time performance scores from the Rasch analysis to the volume of boarding and alighting passengers. The volume of passengers is often used as an indicator when scheduling dwell times, where longer dwell times are expected when passenger volumes are higher. To understand the relation between the Rasch output and passenger volumes we first plot the station difficulty score in relation to the average volume of boarding and alighting passengers at each station. The results of this are shown in Fig. 12 for Sweden and Fig. 13 for the stations in the UK. As previously mentioned, it is commonly assumed that a larger number of boarding and alighting passengers will lead to longer dwell times. If this is the case, this would be reflected in higher station difficulty scores for stations with higher passenger volumes. However, the results in Fig. 12 and Fig. 13 show that stations with a higher volume of passengers do not necessarily have a higher station difficulty score. Furthermore, we observe that some stations with similar passenger volumes have different station difficulty scores. This is the case in both Sweden and the UK and indicates that there is no clear relationship between passenger volumes and station difficulty scores. This effect is in line with findings presented by Kuipers (2024) in which it is argued that higher passenger volumes are not necessarily the cause of dwell time delays but other station-specific characteristics play a more important role. Such station-specific characteristics could be the way passengers spread out across the platform, for example, or trains being more prone to late departures as a result of late arriving passengers, trains waiting for connections, or dispatching decisions. An analysis of such factors falls outside of the scope of the paper presented here, however.

Fig. 14 shows the service performance score in relation to the average volume of boarding and alighting passengers on a service level. This is only plotted for the Swedish commuter train data since passenger volumes are only collected on a station level in the UK case. Looking at Fig. 14, we can observe that there is somewhat of a relationship between the average number of passengers and the service performance scores, as indicated by the labels. In this case, larger average volumes of both boarding and alighting passengers result in lower service performance scores and vice versa. The results derived from Figs. 12–14, which examine the correlation between dwell time performance analysed using Rasch analysis and passenger volume, indicate that the service performance scores align with the common assumption that dwell times are worse when passenger volumes are higher. This is in contrast to the station difficulty scores where a different pattern is found.

Discussion

This study aims to 1) show the applicability of the Rasch analysis technique within an operational context with a specific focus on dwell time evaluation for commuter trains and 2) highlight how the output of a Rasch analysis can be used to investigate the dwell time performance of stations and services on a line level. To do so, we made use of data on the dwell time performance of commuter trains in Sweden and the UK. The study presents an analysis of the output of a Rasch analysis and highlights its potential applications in evaluating dwell times on a line level.

Applicability of the Rasch analysis technique within an operational context

In terms of the applicability of the Rasch analysis technique within an operational context, our findings demonstrate that it can be a suitable method to study dwell time performance. The high reliability scores for both Swedish and UK data indicate a good model fit. The high separation scores indicate that the models are sensitive enough to distinguish between both high and low-performing services and stations. Although the Mean Square statistics scores indicate that the model fit for services is not perfect, with 15 % and 19 % of the observations falling outside the desired threshold, we argue that the model fit is still sufficient given the small number of poorly fitting observations. Comparing both the service performance and station difficulty scores to more commonly used indicators of dwell time performance shows somewhat of a mismatch. For

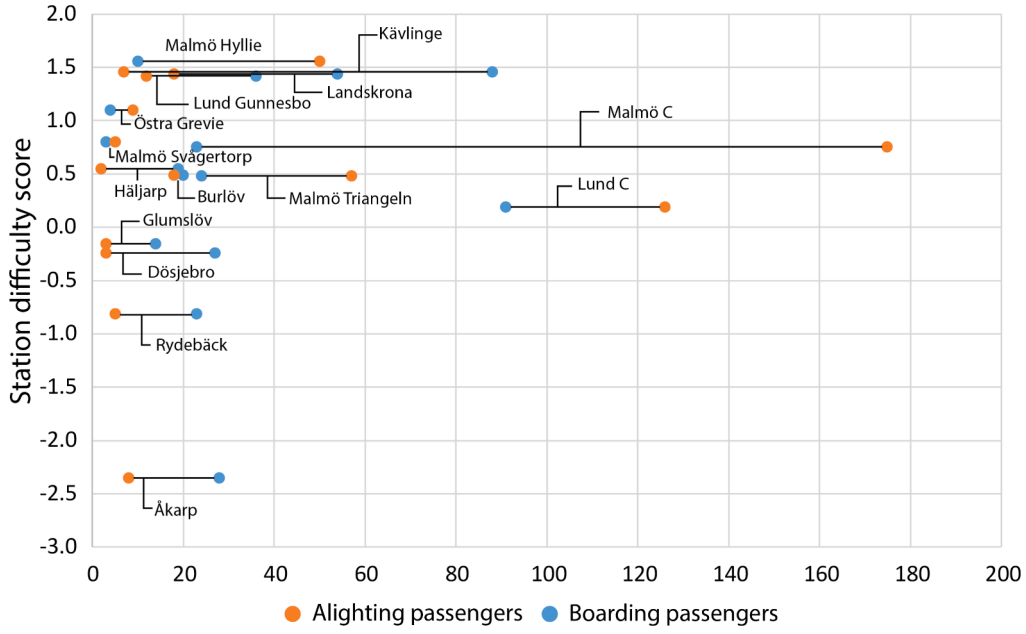


Fig. 12. Relation between station difficulty scores (y-axis) and average volume of passengers (x-axis) for stations in Sweden. The orange dots indicate the average volume of alighting passengers and the blue dots indicate the average volume of boarding passengers for a given station. Station difficulty scores reflect the likelihood of a service being delayed at a given station. Higher station difficulty scores indicate a greater likelihood of a train being delayed and vice versa.

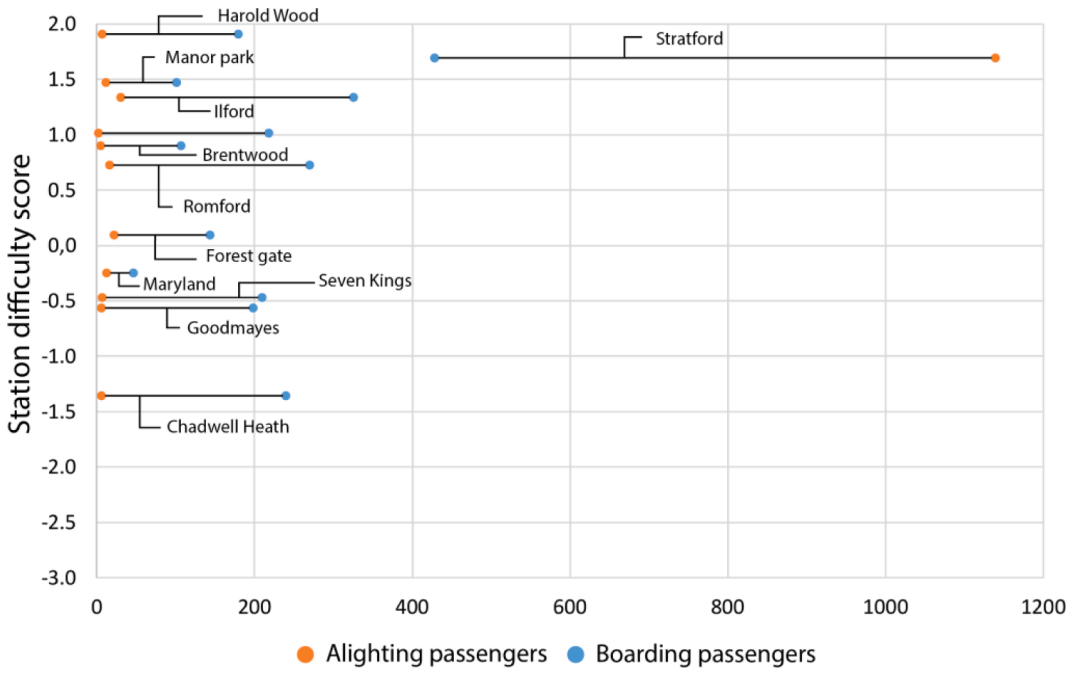


Fig. 13. Relation between station difficulty scores (y-axis) and average volume of passengers (x-axis) for stations in the UK. The orange dots indicate the average volume of alighting passengers and the blue dots indicate the average volume of boarding passengers for a given station. Station difficulty scores reflect the likelihood of a service being delayed at a given station. Higher station difficulty scores indicate a greater likelihood of a train being delayed and vice versa.

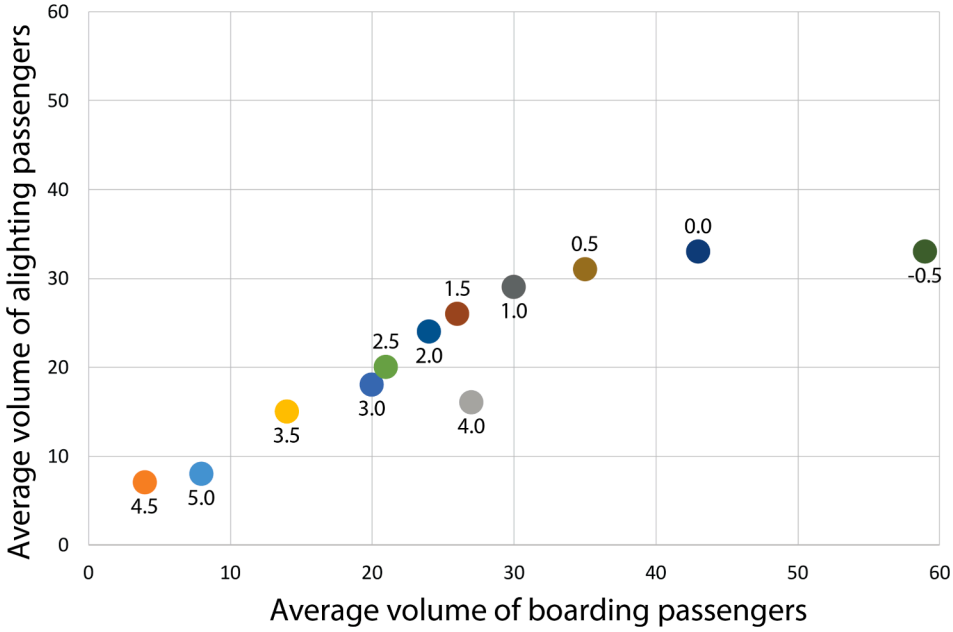


Fig. 14. Service performance score in relation to the average volume of boarding (x-axis) and alighting passengers (y-axis) for Swedish commuter trains. The service performance score is indicated by the labels in the figure and a higher score indicates a better performance and vice versa.

example, while higher passenger volumes generally lead to lower service performance scores this correlation does not match with station difficulty scores. Furthermore, comparing station difficulty scores with median observed dwell time deviations shows that some stations have high dwell time deviations but low difficulty scores. One reason for this is that the Rasch analysis includes the full spectrum of performance aspects by considering the probability of multiple levels of delays for each service at each station, which is not the case when making use of the median dwell time deviation. Another possible reason for this mismatch could be due to a few trains having very large delays, inflating the median dwell times, whilst most services have a dwell time close to the scheduled time, which results in a lower station difficulty score.

Two aspects that are important to the use of the Rasch analysis technique for dwell times and operational research are worth discussing further. The first point is the use of adjusted dwell time deviation over the use of the length of dwell times to study dwell time performance. In our case, we make use of the difference between the scheduled and actual dwell time to reflect dwell time performance in relation to the scheduled dwell time. This measure is preferred over the use of the total dwell time since the latter is heavily influenced by the scheduled dwell

time, where dwell times are expected to be longer when more dwell time is scheduled. Using the dwell time deviation, somewhat, accounts for this. When making use of the dwell time deviation it is, however, vital to ensure that this time is adjusted based on the arrival punctuality of trains. Trains arriving early will have a longer dwell time since it is not possible to depart early. Such a deviation is, however, not a delay as such and if this is not accounted for it can result in an overestimation of the number and size of dwell time delays. This is, however, specific to studying dwell times and might not be required in other operational contexts in which a Rasch analysis can be applied.

The second aspect that is worth discussing is the necessity to convert the continuous dwell time deviations to polytomous data before it can be used in a Rasch analysis. This conversion is not needed in the more traditional way in which Rasch analyses are performed, given that the data input often consists of Likert scale-like data in such cases. This is, however, not the case when using operational data such as dwell times. In our case, we converted the continuous data to polytomous data by making use of different buckets for the size of the dwell time deviations. Different ways in which this conversion can be made, such as the use of standardized z-scores have been explored but resulted in worse model performance and insufficient spread of the data points. We, therefore, argue that the use of buckets is a fitting way in which this conversion from continuous to polytomous data can be made. The decision of the size of these buckets is not a trivial choice and will differ from case to case and needs to be based on the operational and real-world constraints of the specific case. In our case, we define the size of the buckets based on the scheduling regimes used in Sweden and the UK and the number of labels needed to perform a Rasch analysis. Using a Rasch analysis within a different case might require a finer or more coarse definition for the bucket sizes.

Using the output of a Rasch analysis to study dwell times

There are various ways to use the output of the Rasch analysis during the dwell time scheduling process. As Kim et al. (2018) mention, items on both ends of the difficulty scale are important from a policy perspective since items at the top of the scale should be maintained, while those at the bottom of the scale should be addressed. This is also true in our case, with the exception that stations cannot simply be removed. In this case, the station difficulty scores show which stations require the most attention to reduce overall dwell time delays and which stations perform well in terms of dwell times. The output of the Rasch analysis also allows for an analysis on a case-by-case basis (Brush and Soutar, 2022). In practice, this means that the output from the Rasch analysis could be used to study the performance of a single service or group of services across all stations, thereby identifying where additional efforts need to be made for that specific service or group of services.

The key output in terms of using the Rasch analysis to study dwell times is the ability to address both the service performance and station difficulty scores in a single dimension, an indicator that we call the *dwell time performance score* in this study. This combined score provides richer insights into the effectiveness of scheduled dwell times on a line level compared to metrics that address the deviations between the scheduled and actual dwell time on either a service or station level separately, allowing for a better indication of hotspots of poor dwell time performance. By combining both the relative performance of stations and services in a single dimension, it is possible to highlight which services are likely to incur a dwell time delay at which station, which can help guide efforts to improve dwell time punctuality. For example, the dwell time performance scores show that none of the stations included in the cases presented here are likely to experience delays across all services. Furthermore, the dwell time performance scores show that the higher median dwell time deviations at some stations with low station difficulty are likely due to a few services performing poorly. In a practical context, this means that efforts to understand why dwell time delays arise can be focused on those services in combination with the stations, rather than focusing on all services that halt at stations where the median dwell time deviation indicates a delay. Furthermore, using the dwell time performance scores, it shows that it is important to differentiate between services halting at stations when scheduling dwell times.

Limitations

Although using a Rasch analysis to study dwell times has several benefits over the use of measures of central tendency or the volume of passengers, there are some limitations. The output from the Rasch analysis presented here is limited to identifying which services and stations perform poorly in terms of dwell time, for example. While the method is valuable, it does not provide a cause for the respective performances, or the size of the delay incurred at a given moment. Identifying the cause of the dwell time performance of a specific service at a given station calls for more in-depth analyses, which fall outside of the scope of the study presented here, as is the case for identifying the size of the dwell time delays. Using the Rasch analysis does not provide a full analysis of dwell times on a network, but it guides planners in determining where to focus their efforts on understanding delay causes. Another limitation of the Rasch analysis presented here is the potential information loss when converting the continuous dwell time data to polytomous data. While the bucket sizes were carefully chosen through an iterative process, the nuances of having continuous dwell time data are lost as delays of 30 and 59 s are treated the same way in the Swedish case, for example. This problem is less present in the UK case, where the buckets used are smaller, and hence nuances in the data are better conserved, but potential information loss is nevertheless present.

Conclusion

This study highlights the applicability of the Rasch analysis technique within an operational context. To achieve this, continuous dwell time deviation data was converted into polytomous data, as required for the Rasch analysis technique, using buckets to label the size of the dwell time delay deviation. The results of the study indicate that the Rasch model is a suitable method to study dwell time performance and that the chosen bucket sizes provide sufficient spread in the data.

The Rasch analysis output offers insights into the dwell time performance for Swedish and UK commuter trains on a line level, revealing the difference between using a Rasch model compared to commonly used indicators of dwell time performance. The first is that the notion that an increase in passenger volumes leads to longer dwell times does not necessarily hold on a station level, where stations with similar passenger volumes were found to have different station difficulty scores. This indicates that other aspects likely have a larger effect on dwell times on a station level. The effect of passenger volumes on a service level does indicate that an increased volume of passengers leads to worse performance. The major benefit of using a Rasch analysis over the more common way to study dwell time performance is the ability to combine both service performance and station difficulty in a measure we call the dwell time performance score.

While the findings presented in this study are based on case studies and thus limited generalizability in this study area, Future research could include different case studies to further understand the applicability of the Rasch analysis technique in an operational context. Another approach for future studies is to predict the service performance score of services and compare this to the known station difficulty scores. In this way, it is not necessary for planners to perform a complete Rasch analysis for every specific case. The model fit suggests that converting continuous data to polytomous data allows the use of a Rasch analysis outside of the scope of Likert-like scale data for which it is more commonly used. To further understand the applicability of the Rasch analysis technique, future studies can focus on implementing the approach proposed here in different contexts. This can be done both

within transportation, such as the operations of a bus network, as well as other operational processes.

CRediT authorship contribution statement

Ruben Alaric Kuipers: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Data curation, Project administration. **Natchaya Tortainchai:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. **Neba C Tony:** Formal analysis, Investigation, Methodology, Software. **Taku Fujiyama:** Conceptualization, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ruben Kuipers reports financial support was provided by K2 The Swedish Knowledge Centre for Public Transport. Taku Fujiyama reports financial support was provided by City Partnership Programme of UCL, KTH and Lund University.

Data availability

The data that has been used is confidential.

Acknowledgements

The authors would like to thank Carl-William Palmqvist for acquiring part of the funding for this study and for his role in organizing a workshop in the early stage of this study. The author, affiliated with the Cluster of Logistics and Rail Engineering, Mahidol University, would also like to thank the university for facilitating the work on this project. This work was partly funded by K2 The Swedish Knowledge Centre for Public Transport, as well as the City Partnership Programme of UCL (UK), KTH (Sweden) and Lund University (Sweden).

References

Andrich, D., 1978. A rating formulation for ordered response categories. Psychometrika 43 (4), 561–573. https://doi.org/10.1007/BF02293814.

Boone, W.J., 2016. Rasch analysis for instrument development: why, when, and how? CBE—life sciences. Education 15 (4), rm4. https://doi.org/10.1187/cbe.16-04-0148.

Brush, G.J., Soutar, G.N., 2022. A Rasch analysis of service performance in a tourism context. J. Bus. Res. 139, 338–353. https://doi.org/10.1016/j.jbusres.2021.09.038.

Buchmueller, S., Weidmann, U., Nash, A., 2008. Development of a dwell time calculation model for timetable planning. WIT Trans. Built Environ. 103, 525–534. https://doi.org/10.2495/CR080511.

Cappelleri, J.C., Lundy, J.J., Hays, R.D., 2014. Overview of classical test theory and item response theory for quantitative assessment of items in developing patient- reported outcome measures. Clin. Ther. 36, 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006.

Chan, N., 2018. Development of an Instrument to assess Transport Ability for people with low vision and limited mobility. UCL (University College London). Doctoral thesis (Ph.D).

Cheng, Y.H., Chen, S.Y., 2015. Perceived accessibility, mobility, and connectivity of public transportation systems. Transp. Res. A Policy Pract. 77, 386–403. https://doi.org/10.1016/j.tra.2015.05.003.

Christoforou, Z., Chandakas, E., Kaparias, I., 2020. Investigating the impact of dwell time on the reliability of urban light rail operations. Urban Rail Transit. https://doi.org/10.1007/s40864-020-00128-1.

Combrinck, C., 2020. Is this a useful instrument? An introduction to Rasch measurement models. In S. Kramer, S. Laher, A. Fynn, & H. H. Janse van Vuuren (Eds.), Online Readings in Research Methods. Psychological Society of South Africa: Johannesburg. https://doi.org/10.17605/OSF.IO/ BNPFS.

Coulaud, R., Keribin, C., Stoltz, G., 2023. Modeling dwell time in a data-rich railway environment: With operations and passenger flows data. Transp. Res. Part C 146, 103980. https://doi.org/10.1016/j.trc.2022.103980.

Ding, Z. (2016). Headway Control Schemes to Resist Bus Bunching. [Ph.D. Thesis, Georgia Institute of Technology].

Gallo, M., 2011. Measuring passenger satisfaction: a strategy based on Rasch Analysis and the ANOM. J. Appl. Quant. Methods 6 (2), 27–35.

Gothwal, V.K., Wright, T.A., Lamoureux, E.L., Pesudovs, K., 2009. Visual Activities Questionnaire: Assessment of subscale validity for cataract surgery outcomes. J Cataract Refract Surg 35 (11), 1961–1969. https://doi.org/10.1016/j.jcrs.2009.05.058.

Goverde, R.M.P., Hansen, I.A, Hooghiemstra, G., Lopuhaa, H.P., 2001. Delay Distributions in Railway Stations. The 9th World Conference on Transport Research.

Goverde, R.M.P., 2005. Punctuality of Railway Operations and Timetable Stability Analysis [PhD thesis, TU Delft]. https://repository.tudelft.nl/islandora/object/uuid%3Aa40ae4f1-1732-4bf3-bbf5-fdb8dfd635e7.

Gysin, K., 2018. An Investigation of the Influences on Train Dwell Time [Master thesis]. Swiss Federal Institute of Technology, ETH.

Hambleton, R.K., Cook, L.L., 1997. Latent trait models and their use in the analysis of educational test data. J. Educ. Meas. 14, 75–96.

Hansen, I.A., 2010. Railway network timetabling and dynamic traffic management. Int. J. Civil Eng. 8 (1), 14.

Harris, N.G., 2005. Train boarding and alighting rates at high passenger loads. J. Adv. Transp. 40 (3), 249–263. https://doi.org/10.1002/atr.5670400302.

Harris, N. G., Risan, Ø., Schrader, S.-J., 2014. The impact of differing door widths on passenger movement rates. 53–63. https://doi.org/10.2495/CRS140051.

Hart, D.L., Wright, B.D., 2002. Development of an index of physical functional health status in rehabilitation. Arch. Phys. Med. Rehabil. 83 (5), 655–665. https://doi.org/10.1053/apmr.2002.31178.

Havlena, O., Jacura, M., Javořík, T., Svetlík, M., Týfa, L., 2014. Parameters of passenger facilities according to railway station characteristics. Transport Problems 19 (4), 8.

Hirsch, L., Thompson, K., 2014. I can sit but I'd rather stand: Commuter's experience of crowdedness and fellow passenger behaviour in carriages on Australian metropolitan trains. ATRF 2011 - 34th Australasian Transport Research Forum, January.

Jette, A.M., Haley, S.M., Coster, W.J., Kooyoomjian, J.T., Levenson, S., Heeren, T., Ashba, J., 2002. Late life function and disability instrument: I. Development and evaluation of the disability component. J. Gerontol. A Biol. Sci. Med. Sci. 57 (4), M209–M216. https://doi.org/10.1093/gerona/57.4.M209.

Kecman, P., Goverde, R.M.P., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transport 7 (3), 295–319. https://doi.org/10.1007/s12469-015-0106-7.

Khadem Sameni, M., Preston, J., Khadem Sameni, M., 2016. Evaluating efficiency of passenger railway stations: A DEA approach. Res. Transp. Bus. Manag. 20, 33–38. https://doi.org/10.1016/j.rtbm.2016.06.001.

Kim, J., Schmöcker, J.-D., Yu, J.W., Choi, J.Y., 2018. Service quality evaluation for urban rail transfer facilities with Rasch analysis. Travel Behav. Soc. 13, 26–35. https://doi.org/10.1016/j.tbs.2018.05.002.

Krause, C., 2014. Simulation of dynamic station dwell time delays on high frequency rail transport systems. [Master thesis, KTH Royal Institute of Technology]. Stockholm.

Kuipers, R.A., Palmqvist, C.-W., 2022. Passenger volumes and dwell times for commuter trains: A case study using automatic passenger count data in Stockholm. Appl. Sci. 12 (12), 5983. https://doi.org/10.3390/app12125983.

Kuipers, R.A., 2024. Dwell time delays for commuter trains: An analysis of the influence of passengers on dwell time delays [PhD thesis, Lunds Tekniska Högskola]. https://portal.research.lu.se/sv/publications/dwell-time-delays-for-commuter-trains-an-analysis-of-the-influenc.

Lamoureux, E.L., Pallant, J.F., Pesudovs, K., Hassell, J.B., Keeffe, J.E., 2006. The impact of vision impairment questionnaire: an evaluation of its measurement properties using Rasch analysis. Invest. Opthalmol. Visual Sci. 47 (11), 4732. https://doi.org/10.1167/iovs.06-0220.

Lessan, J., Fu, L., Wen, C., Huang, P., Jiang, C., 2018. Stochastic model of train running time and arrival delay: A case study of Wuhan-Guangzhou high-speed rail. Transp. Res. Rec. 2672 (10), 215–223. https://doi.org/10.1177/0361198118780830.

Li, D., Goverde, R.M.P., Daamen, W., He, H., 2014. Train dwell time distributions at short stop stations. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 2410–2415. https://doi.org/10.1109/ITSC.2014.6958076.

Linacre, J.M., 1997. KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the "Truth"? Rasch Meas. Trans. 573–586.

Mair, P., Hatzinger, R., Maier, M.J., 2021. eRm: Extended Rasch Modeling (1.0-2) [Computer software]. https://cran.r-project.org/package=eRm.

Mair, P., Hatzinger, R., 2007. CML based estimation of extended Rasch models with the eRm package in R. Psychol. Sci. 49 (1), 26–43.

Massof, R.W., Rubin, G.S., 2001. Visual function assessment questionnaires. Surv. Ophthalmol. 45 (6), 531–548. https://doi.org/10.1016/S0039-6257(01)00194-1.

Masters, G.N., 1982. A rasch model for partial credit scoring. Psychometrika 47 (2), 149–174. https://doi.org/10.1007/BF02296272.

McCamey, R., 2014. A Primer on the One-Parameter Rasch Model. Am. J. Econ. Bus. Admin. 6 (4), 159–163. https://doi.org/10.3844/ajebasp.2014.159.163.

Ministerie van Infrastructuur en Waterstaat. (2022). Programma Hoogfrequent Spoorvervoer—Achtste Voortgangsrapportage (4). Ministerie van Infrastructuur en Waterstaat. https://www.rijksoverheid.nl/documenten/rapporten/2022/10/11/bijlage-1-phs-vgr-7-2022-1.

Nash, A., Weidmann, U., Bollinger, S., Luethi, M., Buchmueller, S., 2006. Increasing Schedule Reliability on the S-Bahn in Zurich, Switzerland. Transp. Res. Rec. 1955 (1), 17–25. https://doi.org/10.1177/0361198106195500103.

Oliveira, L.C., Fox, C., Birrell, S., Cain, R., 2019. Analysing passengers' behaviours when boarding trains to improve rail infrastructure and technology. Rob. Comput. Integr. Manuf. 57, 282–291. https://doi.org/10.1016/j.rcim.2018.12.008.

Palmqvist, C.W., Kristoffersson, I., 2022. A Methodology for monitoring rail punctuality improvements. IEEE Open J. Intell. Transp. Syst. 3, 388–396. https://doi.org/10.1109/OJITS.2022.3172509.

Palmqvist, C.-W., 2019. Delays and Timetabling for Passenger Trains [Doctoral thesis, Lund University Faculty of Engineering, Technology and Society, Transport and

Roads]. http://portal.research.lu.se/ws/files/70626078/Carl_William_Palmqvist_web.pdf.
Pearce, E., Crossland, M.D., Rubin, G.S., 2011. The efficacy of low vision device training in a hospital-based low vision clinic. Br. J. Ophthalmol. 95 (1), 105–108. https://doi.org/10.1136/bjo.2009.175703.
Pesudovs, K., Garamendi, E., Keeves, J.P., Elliott, D.B., 2003. The activities of daily vision scale for cataract surgery outcomes: re-evaluating validity with Rasch analysis. Invest. Opthalmol. Visual Sci. 44 (7), 2892. https://doi.org/10.1167/iovs.02-1075.
Prieto, L., Alonso, J., Lamarca, R., 2003. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. Health Qual. Life Outcomes 1 (1), 27. https://doi.org/10.1186/1477-7525-1-27.
Pritchard, J., Sadler, J., Blainey, S., Waldock, I., Austin, J., 2021. Predicting and mitigating small fluctuations in station dwell times. J. Rail Transp. Plann. Manage. 18, 100249 https://doi.org/10.1016/j.jrtpm.2021.100249.
Rasch, G., 1980. Probabilistic models for some intelligence and attainment tests. University of Chicago Press.
Reusser, D.E., Loukopoulos, P., Stauffacher, M., Scholz, R.W., 2008. Classifying railway stations for sustainable transitions – balancing node and place functions. J. Transp. Geogr. 16 (3), 191–202. https://doi.org/10.1016/j.jtrangeo.2007.05.004.
Seriani, S., Fernandez, R., Luangboriboon, N., Fujiyama, T., 2019. Exploring the Effect of Boarding and Alighting Ratio on Passengers' Behaviour at Metro Stations by Laboratory Experiments. J. Adv. Transp. 2019 https://doi.org/10.1155/2019/6530897.
Souza, A.C.S., Bittencourt, L., Taco, P.W.G., 2018. Women's perspective in pedestrian mobility planning: the case of Brasília. Transp. Res. Procedia 33, 131–138. https://doi.org/10.1016/J.TRPRO.2018.10.085.
Stoilova, S., Nikolova, R., 2017. Classifying railway passenger stations for use transport planning – application to Bulgarian railway network. Transport Problems 11 (2), 143–155. https://doi.org/10.20858/tp.2016.11.2.14.
SVI, 2013. Bahnhöfe und Haltestellen: Typisierung – Ausgestaltung – Kooperation. https://www.svi.ch/media/upload/publications_de/af30a80d_SVI_Leitfaden_2013_01_Bahnh%C3%B6fe_Haltestellen_131029.pdf.
Tennant, A., Conaghan, P.G., 2007. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care Res (Hoboken) 57, 1358–1362. https://doi.org/10.1002/ART.23108.
Tesio, L., Caronni, A., Kumbhare, D., Scarano, S., 2024. Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. Disabil. Rehabil. 46 (3), 591–603. https://doi.org/10.1080/09638288.2023.2169771.
Tortainchai, N., Wong, H., Winslett, D., Fujiyama, T., 2022. Train dwell time efficiency evaluation with data envelopment analysis: case study of London underground Victoria line. Transp. Res. Rec. 2676 (3), 728–739. https://doi.org/10.1177/03611981211056640.
Tortainchai, N., 2023. Train Dwell Time Evaluation at High Passenger Volume Stations. Doctoral thesis (Ph.D), UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10167384/.
Tran, V.D., Dorofeeva, V.V., Loskutova, E.E., 2018. Development and validation of a scale to measure the quality of patient medication counseling using Rasch model. Pharm. Pract. 16 (4), 1327. https://doi.org/10.18549/PharmPract.2018.04.1327.
Turano, K.A., Geruschat, D.R., Stahl, J.W., Massof, R.W., 1999. Perceived visual ability for independent mobility in persons with retinitis pigmentosa. Invest. Ophthalmol. Vis. Sci. 40 (5), 865–877.
van den Heuvel, J., 2016. Field experiments with train stopping positions at Schiphol airport train station in Amsterdam, Netherlands. Transp. Res. Rec. 2546 (1), 24–32. https://doi.org/10.3141/2546-04.
van Loon, R., Rietveld, P., Brons, M., 2011. Travel-time reliability impacts on railway passenger demand: A revealed preference analysis. J. Transp. Geogr. 19 (4), 917–925. https://doi.org/10.1016/j.jtrangeo.2010.11.009.
Vieira, A.P., Christofoletti, L.M., Vilela, P.R.S., 2018. Analyzing Railway Capacity Using a Planning Tool. In: 2018 Joint Rail Conference. https://doi.org/10.1115/JRC2018-6160.
Wu, W., Liu, R., Jin, W., 2018. Integrating bus holding control strategies and schedule recovery: simulation-based comparison and recommendation. J. Adv. Transp.. Hindawi Limited 2018. https://doi.org/10.1155/2018/9407801.
Yamamura, A., Koresawa, M., Adachi, S., Tomii, N., 2012. Identification of causes of delays in urban railways. 403–414. https://doi.org/10.2495/CR120341.
Yang, J., Shiwakoti, N., Tay, R., 2019. Train dwell time models – development in the past forty years. Australasian Transport Research Forum 2019 Proceedings, 12.
Yuan, J., Goverde, R.M.P., Hansen, I.A., 2010. Evaluating stochastic train process time distribution models on the basis of empirical detection data. WIT Trans. State Art Sci. Eng. 40, 95–104. https://doi.org/10.2495/978-1-84564.
Zemp, S., Stauffacher, M., Lang, D.J., Scholz, R.W., 2011. Classifying railway stations for strategic transport and land use planning: Context matters! J. Transp. Geogr. 19 (4), 670–679. https://doi.org/10.1016/j.jtrangeo.2010.08.008.
Zheng, X., Rabe-Hesketh, S., 2007. Estimating parameters of dichotomous and ordinal item response models with gllamm. The Stata Journal: Promoting Communications on Statistics and Stata 7 (3), 313–333. https://doi.org/10.1177/1536867X0700700302.
Zhou, Y., Zheng, S., Hu, Z., Chen, Y., 2022. Metro station risk classification based on smart card data: A case study in Beijing. Physica A 594, 127019. https://doi.org/10.1016/j.physa.2022.127019.