Transportation Research Part F: Psychology and Behaviour 116 (2026) 103418

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf

EXID: Development and validation of the experienced impact of disruptions scale

Marloes Vredenborg [a,*], Marit Bentvelzen [a], Anouk van Kasteren [a], Christine Bauer [b], Judith Masthoff [a]

[a] Utrecht University, Princetonplein 5, Utrecht, 3584 CC, The Netherlands
[b] University of Salzburg, Jakob-Haringer-Strasse 1, Salzburg, 5020, Austria

ARTICLE INFO

Keywords:
Disruptions scale
Validation
Disruption experience
Public transport
Disruption
Questionnaire

ABSTRACT

Disruptions in public transport can significantly affect how travellers experience their journeys. While several attempts have been made to improve this experience, reliably measuring the experienced impact of disruptions remains a challenge. To address this gap, we developed and validated the EXperienced Impact of Disruptions Scale (EXID) scale, designed to capture the multifaceted nature of disruption experience from the traveller's perspective. To construct and validate the scale, we followed a structured, multi-stage process. We first derived an initial pool of items based on prior literature, which were then refined through expert review ($n = 3$). We pre-tested the selected items through cognitive interviews ($n = 5$) and conducted an exploratory factor analysis on survey data ($n = 350$) to refine the scale and assess reliability. Based on a second survey ($n = 209$), we performed a confirmatory factor analysis and evaluated construct validity. Finally, test-retest validity was assessed with a third sample ($n = 22$). The analysis revealed six interrelated but conceptually distinct factors that constitute the experienced impact of disruption: *agency*, *anxiety*, *frustration*, *disorientation*, *time-related stress*, and *travel behaviour change*. These factors reflect the complex and interconnected nature of travellers' responses to disruptions. The EXID scale is suitable for both research and practice. It can be applied in real-time during a disruption, in retrospect, or for hypothetical travel disruption scenarios. Furthermore, it can be used to evaluate interventions aimed to improve the disruption experience, such as real-time information or alternative transport strategies. Ultimately, EXID provides a robust instrument to quantitatively measure a disruption's impact on people's experience and supports the design and evaluation of mitigation strategies in practice.

1. Introduction

When travelling by public transport, travellers generally expect a smooth and pleasant journey. However, disruptions may occur, leading to dissatisfaction (Currie and Muir, 2017; Lunke, 2020), anxiety (Cheng, 2010), stress (Cantwell et al., 2009), frustration (Rezapour and Ferraro, 2021) and a decreased feeling of control (Cantwell et al., 2009). Unfortunately, negative experiences tend to stay with people (Bougie et al., 2003), and repeated negative experiences may lead travellers to seek alternative modes of transport (Sarker et al., 2019). While disruptions will continue to occur, how travellers experience them can be improved. Various interventions

* Corresponding author.
E-mail address: m.t.r.vredenborg@uu.nl (M. Vredenborg).
URL: marloesvredenborg.com (M. Vredenborg)

https://doi.org/10.1016/j.trf.2025.103418
Received 21 March 2025; Received in revised form 16 October 2025; Accepted 16 October 2025

have been proposed to improve the traveller experience during disruptions (Dziekan and Kottenhoff, 2007; Ferreira et al., 2018; Gault et al., 2019; Zografos et al., 2009). Although these interventions are promising, no validated instrument currently exists to systematically assess travellers' experienced impact of disruptions, nor to evaluate the effectiveness of interventions in shaping those experiences.

Existing scales in the field of public transport have primarily focused on the overall travel experience (Carreira et al., 2014; Ittamalla and Kumar, 2021), as well as passenger satisfaction (Ettema et al., 2011; Kökalan and Tutan, 2021), mood (Glasgow et al., 2018), and well-being (Singleton and Clifton, 2021). Other studies have focused on passengers' perceptions of public transport service quality (Hu and Jen, 2006; Mahapatra and Bellamkonda, 2023). While these scales provide a suitable identification of the overall travel experience, there is a lack of a standardised and validated measurement tool focusing specifically on public transport disruptions. A scale focusing on disruptions allows us to assess how travellers experience the impact of disruptions, how they perceive their severity and which specific factors contribute to this severity. Moreover, it allows us to see whether specific interventions during a disruption positively or negatively affect how travellers experience the impact of disruptions and which aspects of their experience are affected. Rezapour and Ferraro (2021) examine how train delays and real-time information influence commuters' physiological and psychological experiences, and their perceptions of the passenger information system. Using factor analysis and structural equation modelling, they identify four latent factors. However, they focus on analysing these effects rather than on developing a validated scale.

This paper proposes a measurement instrument—the EXperienced Impact of Disruptions Scale (EXID), developed to assess the experienced impact of disruptions on public transport travellers. Scale development involves creating a reliable and valid set of items (e.g., "The disruption made me feel nervous" or "During the disruption, I felt capable of finding solutions") to measure a specific, unobservable psychological or behavioural construct (e.g., impact of disruption), organised into sub-scales that reflect distinct dimensions (e.g., frustration, feeling in control). Following a structured process (Boateng et al., 2018), we first investigated past work on the experienced impact of disruptions on travellers and existing relevant scales. Then, we generated an initial list of items and evaluated these with expert reviews. We used Exploratory Factor Analysis (EFA) to reduce the number of items and to explore the scale's factor structure and dimensionality. Afterwards, we investigated the correctness of our model using Confirmatory Factor Analysis (CFA) and validated the final scale through several assessments. The resulting EXID scale captures six factors of the experienced impact of public transport disruptions: *agency*, *anxiety*, *frustration*, *disorientation*, *time-related stress*, and *travel behaviour change*. It provides a multifaceted view of how travellers experience the impact of disruptions and how these experiences differ across individuals, contexts, and types of disruptions. To our knowledge, our work offers the first validated scale for assessing this impact, both during and after the journey.

The scale can be applied in various contexts, including immediately after a disruption, when recalling past disruptions, or in hypothetical vignette studies. It supports researchers and practitioners in quantitatively measuring the impact of disruptions on people's experience, comparing the effects of several disruption management strategies, or assessing the effectiveness of a newly implemented mitigation strategy in practice. As such, the scale can help determine whether specific interventions (e.g., providing additional information during disruptions) influence how people perceive these disruptions.

2. Related work

In this section, we first show how disruptions impact a traveller's experience (Section 2.1). Next, we discuss related scales and show a need for an instrument that measures the traveller's experience of public transport disruptions (Section 2.2).

2.1. Disruptions and travellers' experience

Disruptions in public transport occur when operations deviate from their usual schedule, affecting multiple stakeholders, including transport operators, policymakers, and travellers (Ge et al., 2022; Yap and Cats, 2021). These disruptions may be planned, such as maintenance work or strikes, or occur unplanned, e.g., resulting from technical failures, severe weather conditions, or accidents. Researchers commonly distinguish between recurrent and non-recurrent disruptions (Durand, 2017; Yap and Cats, 2021). Recurrent disruptions, such as minor malfunctions or short delays, happen frequently but have a limited impact and are sometimes referred to as disturbances. In contrast, non-recurrent disruptions, including signal failures or derailments, are less common but often more severe.

Disruptions can extend travel time and force travellers to re-plan their journeys (Ibrahim et al., 2020; Van Lierop et al., 2018). They may affect the predictability of a journey, leading to increased stress and more complex decision-making. For example, Bhat and Sardesai (2006) found that travel time reliability is an important predictor in modelling travel-mode choice behaviour. Furthermore, when a journey feels unpredictable, commuters tend to report higher levels of stress (Cantwell et al., 2009; Evans et al., 2002), particularly during peak hours or on longer trips (Evans and Wener, 2006; Wener et al., 2003). A perceived lack of control or autonomy may further intensify stress (e.g., Bollini et al., 2004; Dijkstra and Homan, 2016; Spector, 1986), which researchers have linked to reduced commuter well-being (Koslowsky et al., 2013). For instance, in their survey study, Sposato et al. (2012) found that perceived control is a strong predictor of commuter stress. Moreover, when travellers are uncertain about arriving on time, workplace expectations for punctuality can heighten anxiety and potentially affect professional performance (Gobind, 2018).

The cognitive demands associated with disruptions also influence the traveller's experience. Cognitive Load Theory (Sweller, 1988, 2010) suggests that a person can only process a certain amount of information at a given time. In public transport, travellers continuously process information to navigate their journeys (Armougum et al., 2020; Grotenhuis et al., 2007). When disruptions occur, travellers must quickly re-plan their journeys, process new information, and make decisions under pressure, which can increase cognitive load and negatively affect their travel experience. However, researchers suggest that providing clear and timely information can mitigate this effect and improve satisfaction (Ibraeva and de Sousa, 2014; Mouwen, 2015; Pruyn and Smidts, 1998; Romero et al., 2023).

Moreover, disruptions can alter how travellers perceive time itself. While the duration of a journey influences stress and satisfaction (Cantwell et al., 2009; Evans and Wener, 2006; Rüger et al., 2017), the perception of time can also be distorted. Research shows that the perception of time is subjective and can be influenced by various factors (Dewulf et al., 2012; Hess et al., 2004; Psarros et al., 2011; Wittmann, 2016). For instance, the perception of waiting time is influenced by service delivery quality (e.g., reliability, punctuality) (Casado Diaz and Más Ruíz, 2002; Dubé et al., 1991; Taylor, 1994). Moreover, environmental stimuli such as colour and sound can influence the waiting experience (Van Hagen et al., 2014), while comfort, crowding, and weather conditions further affect how travellers perceive waiting time (Beirão and Cabral, 2007). In case of unreliable public transport services, such as delays, travellers perceive travel time as unreasonably long (Li, 2003), further shaping overall passenger satisfaction (Krygsman et al., 2004).

In addition, literature revealed that several social and environmental factors further contribute to the disruption experience. In general, research shows that traveller satisfaction is influenced by the behaviour of others, such as the presence and support of staff (Mouwen, 2015; Van Lierop et al., 2018). For example, when disruptions occur, staff have to respond well by taking travellers seriously and offering them sufficient and useful information (van Hagen and van Oort, 2019). When this support is lacking during a disruption, travellers may feel isolated and unsupported, negatively affecting their experience. However, disruptions may also lead to crowded conditions, invading travellers' personal space. Such conditions can impact stress levels (Cox et al., 2006; Evans and Wener, 2007) and influence a traveller's feeling of safety (Mouwen, 2015). Crowdedness often leads to discomfort, negatively affecting the overall travel experience (Cantwell et al., 2009; Mahudin et al., 2012). Contributing factors include the need to stand, limited space for personal activities, and close proximity to other passengers (Grison et al., 2017; Haywood et al., 2017). As Cox et al. (2006) highlight, crowding is not just a matter of physical density but a psychological experience shaped by cognitive, social, and environmental factors. One other example related to environmental factors concerns unfamiliar situations. Several studies have found that disruptions can raise safety concerns, as they can place passengers in unfamiliar or vulnerable situations (Friman et al., 2001; Ibrahim et al., 2020; Mouwen, 2015; Van Lierop et al., 2018).

Beyond these immediate impacts, disruptions can have long-term consequences on travellers' perceptions and attitudes towards public transport travel (Currie and Muir, 2017; He et al., 2024). As disruptions become more frequent and severe, travellers may alter their future behaviour, such as departing earlier, taking different routes, or changing their mode of transport (Lin et al., 2016; Papangelis et al., 2013). For instance, a study in Krakow by Drabicki et al. (2021) revealed that 77 % of respondents made long-term adjustments to their travel behaviour due to experiencing disruptions. Similarly, Nichols et al. (2024) showed that delays influence future decisions regarding public transport use. In extreme cases, travellers have adopted unconventional coping strategies, such as storing spare clothes at alternative locations or even relocating residences (Papangelis et al., 2013).

Overall, the literature consistently links transport disruptions to lower traveller satisfaction (Friman et al., 2001; Oliveira et al., 2023). For example, Currie and Muir (2017) found that Melbourne rail passengers who experienced unplanned disruptions reported significantly lower satisfaction levels, both with their overall journey and with the operator's response. Likewise, van Kasteren et al. (2024) highlighted that travellers reported low satisfaction with the information provided during disruptions. A Malaysian study (Islam et al., 2014) identified service quality attributes, travel and waiting time, and reliability of services as key predictors of traveller satisfaction. In the UK, Monsuur et al. (2021) found that delays exceeding 30 min or cancellations, along with the provision of information regarding these disruptions, had a particularly negative impact on satisfaction.

In summary, the reviewed literature shows that various aspects influence how travellers *experience* public transport disruptions. To build on this, our work develops a validated scale to systematically measure the experienced impact of disruptions.

2.2. Measuring the public transport disruption experience

Several validated scales were developed to measure travellers' experiences while travelling with public transport. Many of these focus on service quality, satisfaction, and travel mood, which could indirectly capture elements of the disruption experience.

A widely used scale for measuring perceptions of service quality (SERVQUAL) developed by Parasuraman et al. (1988) has also been applied in various public transport research (e.g., Mapunda, 2021; Randheer et al., 2011; Sam et al., 2018). Moreover, several attempts to tailor the SERVQUAL scale to the transportation domain have been made (Caro and García, 2007; Hu and Jen, 2006; Lai and Chen, 2011; Prasad and Shekhar, 2010; Sánchez Pérez et al., 2007; Wen et al., 2005). Due to the limitations of these scales, Bakti and Sumaedi (2015) developed the P-TRANSQUAL scale based on these earlier adaptations of the SERVQUAL scale. The authors defined four dimensions to measure perceived service quality: *comfort*, *tangible*, *personnel*, and *reliability*. Similarly, the Flight Quality Scale (FliQual) measures the perceived service quality among air passengers (Mahapatra and Bellamkonda, 2023), using five factors: *available and accessible services*, *food and beverage services*, *staff services* (e.g., staff are easy to contact), *staff courtesy and behaviour*, and *updated technical services* (e.g., the airport having sufficient electronic display boards). While these scales do not directly measure the impact of disruptions, factors such as staff services may offer insights into travellers' needs during such events.

Beyond service quality, some scales take a broader perspective by assessing the overall passenger experience. Ittamalla and Kumar (2021) developed the Holistic Passenger Experience (HPX) scale, which evaluates experience across seven factors. Four of these—*vehicle maintenance* (e.g., vehicle cleanliness), *comfort and security*, *off-board service* (e.g., the station's toilet facilities), and *travel information*—are aspects they argue can be directly managed by the service provider. In contrast, they argue that the remaining three—*social environment*, *supporting services*, and *accessibility*—are not fully controllable by service providers. Similarly, Carreira

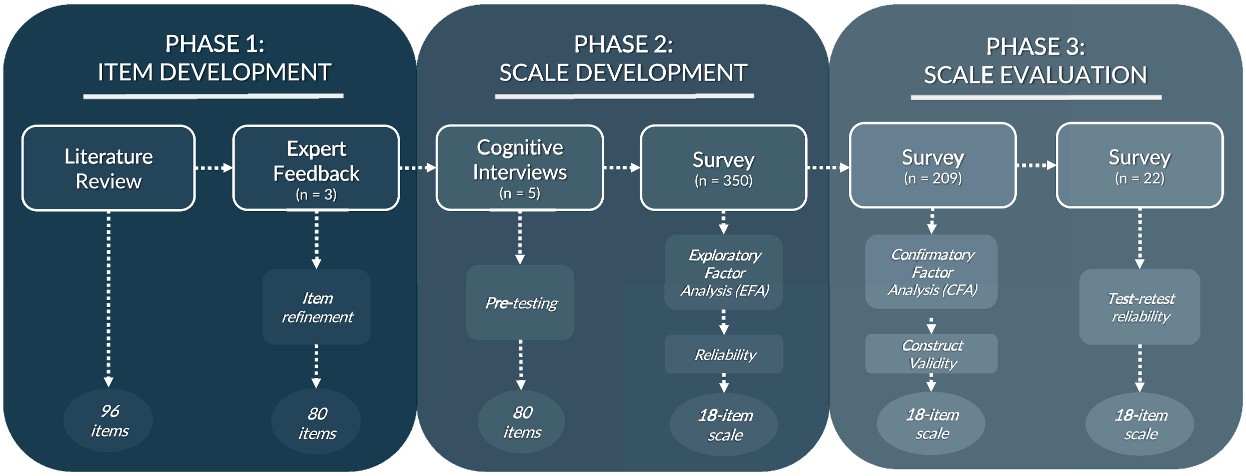

Fig. 1. A detailed overview of the EXID scale formation process, which consists of three phases: (1) item development, (2) scale development, and (3) scale evaluation.

et al. (2014) proposed a holistic framework with seven factors, including *individual space*, *information provision*, *staff skills*, *social environment*, *vehicle maintenance*, *off-board facilities*, and *ticket line service*. Additionally, the Passenger Satisfaction Scale (Kökalan and Tutan, 2021) comprises 22 items categorised into four factors: *technical satisfaction*, *service satisfaction*, *comfort satisfaction*, and *cleanliness satisfaction*.

Furthermore, research has broadened the assessment of travel experience by incorporating both cognitive and affective components. Ettema et al. (2011) developed the nine-item Satisfaction with Travel Scale (STS), which distinguishes between two pairs of affective states: *positive deactivation* (e.g., relaxed) and *negative activation* (e.g., time-pressed) as well as *positive activation* (e.g., alert) and *negative deactivation* (e.g., tired), each assessed using three items. The remaining three items evaluate *overall transport quality and efficiency*. Focusing more specifically on mood, Glasgow et al. (2018) introduced the Travel Mood Scale to assess travellers' mood immediately after their journey, differentiating between two factors: *general* and *relaxation*. Finally, Singleton and Clifton (2021) created a measurement tool to assess the affective and eudaimonic subjective well-being in the travel domain. Their four-factor model captures both travel affect, including *enjoyment*, *attentiveness*, *distress*, and *fear* and travel eudaimonia, which encompasses *health*, *competence*, *autonomy*, and *security*. Different approaches have been used in the development of these scales, reflecting the variety of methods applied to identify and refine relevant items as well as to establish underlying model structures. See Table 1 for an overview of the methodologies and results of these studies.

While existing scales enable the measurement of the various aspects of the public transport experience, they do not directly assess the experienced impact of disruptions on travellers. Instead, these scales primarily focus on broader factors such as service quality, satisfaction, and mood, which reflect the journey as a whole rather than the specific effects of disruptions. A dedicated and validated scale would offer researchers and practitioners more insights into travellers' experienced impact of disruptions, the factors contributing to these perceptions, and the perceived severity of disruptions.

3. Methodology

Several researchers provide guidance and best practices for developing and validating scales (e.g., Boateng et al., 2018; Carpenter, 2018; DeVellis and Thorpe, 2021). This process can be roughly divided into three main phases (Boateng et al., 2018): (1) item development, (2) scale development, and (3) scale evaluation. Following these established guidelines and phases, we developed and evaluated the EXID scale (see Fig. 1).

This study was allowed to proceed by Utrecht University's Research Institute of Information and Computing Sciences on the basis of an Ethics and Privacy Quick Scan.

3.1. Phase 1: Item development

The item development phase involves identifying the domain in order to specify its boundaries. To achieve this, we reviewed the literature to define the main dimensions relevant to the experienced impact on travellers of public transport disruptions and generated an initial list of items. Afterwards, the item list was evaluated and refined based on expert reviews.

3.1.1. Item generation: Literature review

As discussed in Section 2.1, the literature demonstrates that public transport disruptions affect travellers in multiple ways. Following the guidelines for the scale development process outlined by Boateng et al. (2018), we first reviewed the literature to identify and specify dimensions relevant to the experienced impact of disruptions. This process was supported by iterative brainstorming and

Table 1
Overview of methodological approaches used in the development of travel experience-related scales. The table summarises the measured constructs, item generation and refinement processes, analytical techniques applied, and the resulting factor structures.

Ref.	Construct and Scale name	Item generation and Scale construction	Analytical techniques	Final factor structure
Bakti and Sumaedi (2015)	Public land transport service quality* (P-TRANSQUAL)	23 items based on literature → Refined and reduced through EFA → 18 final items	EFA → Reliability via Cronbach's α → CFA → Reliability via Cronbach's α → SEM → Convergent, discriminant, and criterion-related validity → Model stability	4 factors: Comfort, Tangible, Personnel, Reliability
Carreira et al. (2014)	Traveller experience	70 items derived from literature and interviews → Refined and reduced through pilot survey and EFA → 28 final items	EFA (PCA with varimax rotation) → Reliability via Cronbach's α → CFA → Convergent and discriminant validity → SEM	7 factors: Individual space, Information provision, Staff's skills, Social environment, Vehicle maintenance, Off-board facilities, Ticket line service
Ettema et al. (2011)	Satisfaction with travel (STS)	9 items based on literature	Reliability via Cronbach's α → ANOVA, t-tests → Concurrent validity	3 factors: Cognitive evaluation, Positive deactivation–negative activation, Positive activation–negative deactivation
Glasgow et al. (2018)	Travel mood	9 bipolar word-pairs based on literature → Refined and reduced to 7 final items	EFA (PCA and Oblimin rotation) → Reliability via Cronbach's α → CFA → Convergent validity → Regression	2 factors: General Mood, Relaxation
Ittamalla and Kumar (2021)	Traveller experience (HPX)	68 items from literature and interviews → Refined and reduced through expert review, item-total correlation and EFA → 22 final items	EFA (PCA with varimax rotation) → Reliability via Cronbach's α → CFA → Convergent and discriminant validity → Measurement invariance tests → Nomological validation → SEM → Test-retest reliability	7 factors: Social environments, Vehicle maintenance, Comfort & safety, Supporting services, Travel information, Accessibility, Off-board services
Kökalan and Tutan (2021)	Passenger satisfaction	30 items from literature → Refined and reduced through expert review and EFA → 22 final items	EFA (PCA with varimax rotation) → CFA → Reliability via Cronbach's α	4 factors: Service, Technical, Comfort, Cleanliness
Mahapatra and Bellamkonda (2023)	Service quality in air travel (FliQual)	114 items based on critical incidences and literature → Refined and reduced through focus groups, survey, expert reviews and EFA → 16 final items	Inter-item correlations → EFA (PCA with varimax rotation) → Reliability via Cronbach's α → Construct, discriminant, convergent, known-group validity → CFA → SEM and predictive model testing → Multi-group CFA for cross-validation → Nomological validation → Fuzzy Delphi	5 factors: Available and accessible services, Food and beverage services, Staff services, Staff courtesy and behaviour, Updated technical services.
Singleton and Clifton (2021)	Affective and eudaimonic subjective well-being	Travel affect: 120 adjectives based on literature → Refined and reduced through survey, interviews and EFA → 18 final items	EFA → CFA → Reliability via Cronbach's α	4 factors: Enjoyment, Attentiveness, Distress, Fear
		Travel eudaimonia: 75 adjectives based on literature → Refined and reduced through surveys, interviews and EFA → 16 final items		4 factors: Health, Competence, Autonomy, Security

*This scale is based on multiple other adaptations of the SERVQUAL scale. Therefore, these earlier adaptations are not included in this table.

Table 2
Overview of the conceptual dimensions, including definitions and key references used for item generation. The 'References' column indicates whether the items for each dimension were adapted from existing validated scales ('adapted from') or informed by broader theoretical or empirical literature during development ('built upon').

Dimension	Description	References
Physical demands	The extent to which the disruption imposes physical strain, effort, or discomfort on the traveller.	Adapted from Rezapour and Ferraro (2021) and built upon Beirão and Cabral (2007), Cantwell et al. (2009), Grison et al. (2017), Hart and Staveland (1988), Haywood et al. (2017), Ibrahim et al. (2020), Mahudin et al. (2012), Van Lierop et al. (2018)
Temporal demands	The degree to which the disruption alters or intensifies the traveller's perception and management of time.	Adapted from Ettema et al. (2011) and built upon Bhat and Sardesai (2006), Cantwell et al. (2009), Evans et al. (2002), Hart and Staveland (1988), Li (2003), Rezapour and Ferraro (2021), Van Lierop et al. (2018)
Mental demands	The cognitive burden imposed by the disruption, reflecting the mental effort required to navigate, plan, or adapt.	Built upon Armougum et al. (2020), Grotenhuis et al. (2007), Hart and Staveland (1988), Mouwen (2015), Sweller (1988, 2010).
Emotional demands	The emotional impact of the disruption, capturing the intensity and nature of emotional reactions.	Adapted from Ettema et al. (2011), Rezapour and Ferraro (2021) and built upon Cantwell et al. (2009)
Frustration	The degree of irritation, dissatisfaction, or anger caused by the disruption.	Adapted from Ettema et al. (2011), Rezapour and Ferraro (2021) and built upon Currie and Muir (2017), Friman et al. (2001)
Social environment	The perceived quality and supportiveness of the social context during the disruption.	Built upon van Hagen and van Oort (2019), Mouwen (2015), Van Lierop et al. (2018).
Autonomy and control	The traveller's perceived ability to make choices, remain independent, and effectively manage the situation.	Adapted from Friman and Olsson (2023) and built upon Bollini et al. (2004), Dijkstra and Homan (2016), Gobind (2018), Koslowsky et al. (2013), Spector (1986), Sposato et al. (2012), Van Lierop et al. (2018).
Safety	The extent to which the disruption leads to feelings of personal insecurity or heightened fear of harm.	Built upon Currie and Muir (2017), Friman et al. (2001), Ibrahim et al. (2020), Mouwen (2015), Van Lierop et al. (2018).
Long-term effect	The anticipated lasting consequences of the disruption.	Adapted from Friman and Olsson (2023) and built upon Currie and Muir (2017), Drabicki et al. (2021), He et al. (2024), Lin et al. (2016), Nichols et al. (2024), Papangelis et al. (2013), Van Lierop et al. (2018).

discussion within the research team. Ultimately, this resulted in nine dimensions (see Table 2). We expected the dimensions to be distinct yet closely related, reflecting the complex nature of disruption experiences described in prior research. These dimensions formed the conceptual basis for our scale development, informing both item generation and subsequent analysis.

To generate an initial pool of items, we reviewed the literature on the experienced impact of disruptions on travellers as described in Section 2.1. Although not all sources discussed disruptions explicitly, we used findings from these sources to inspire initial brainstorm sessions (see Table 2). Drawing on this literature and adapting from existing scales, we brainstormed items to capture disruptions' perceived impact. The resulting conceptual dimensions, along with definitions and references used in item development, are summarised in Table 2. Subsequently, these items were discussed, and in cases of similarity, one item was retained. Through an iterative process, we refined the items by establishing criteria: (1) We ensured that items were not overly specific, which could make it difficult for participants to relate to them if they had not experienced a particular situation. (2) We formulated the items subjectively, avoiding objective statements, since the scale aims to measure personal experience, and objective items make it difficult to understand how the person feels about the situation. (3) We incorporated reverse-scored items to control for agreement bias. (4) We split long items for clarity, and (5) we tried to use accessible language so that individuals with a basic understanding of English could easily comprehend the items.

The item generation process resulted in 96 initial items, which were subject to further analysis. The full list of initial items is available in the supplementary material.

3.1.2. Item refinement: Expert reviews

We conducted expert reviews to assess whether each item adequately represented the domain of interest, ensuring content validity. Additionally, experts evaluated whether the items measured their intended constructs, confirming face validity. The reviews also aimed to identify gaps and assess whether additional items were needed for completeness.

We recruited three experts ($n = 3$) to review and provide feedback on the initial item list. To ensure diverse perspectives, our experts included an associate professor in psychology and two employees from a major railway company: a researcher with a PhD in traveller experience and a service designer specialising in traveller needs during public transport disruptions. The reviews were conducted online, with each session lasting about one hour.

After obtaining informed consent, we asked the experts to review each item in the list, evaluating its clarity, comprehensibility, and relevance to the overall construct of the scale. Additionally, they identified any missing concepts. Audio recordings were made throughout the sessions. Based on their feedback, we created an overview summarising comments for each item, allowing us to identify problematic items, propose new ones, and determine necessary revisions or removals. This iterative process led to a refined list of 80 items for further analysis. The full item list, following expert review, is available in the supplementary material.

The expert reviews indicated general agreement that the proposed dimensions were conceptually sound and that the items adequately represented the domain of interest. No changes to the dimensions were recommended, though experts suggested adding a few items to better capture the full experienced impact of disruptions. One expert noted that travel disruptions may, in some cases, promote a sense of social connectedness with others. Therefore, an item reflecting this perspective was added to the social environment dimension. Moreover, experts recommended including items addressing underload states (e.g., feeling overly relaxed or drowsy) and contrasting experience of heightened attentional focus within the mental demands dimension. Finally, they noted that some items were difficult to interpret and should be simplified, and they emphasised the need to highlight subtle differences in wording to improve clarity.

3.2. Phase 2: Scale development

This phase involves the evaluation of the set of items after the expert reviews. This included assessing the clarity of the items, performing item reduction and identifying the underlying factor structure. We first conducted cognitive interviews to evaluate whether participants interpreted the items as intended. Next, we conducted an online survey to collect data for exploratory factor analysis and reliability tests.

3.2.1. Pre-testing items: Cognitive interviews

We conducted cognitive interviews to ensure content and construct validity among potential scale users. Cognitive interviewing involves administering draft survey questions while collecting additional verbal information about participants' thinking processes. This helps evaluate response quality and determine whether the items elicit the intended information (Beatty and Willis, 2007).

Five participants ($n = 5$) reviewed the item list during in-person sessions, each lasting approximately 20 min. Participants were asked to recall a personal experience of a public transport disruption, describe it aloud, and then complete a printed version of the scale based on that event. They were encouraged to think aloud, identify confusing items, and provide general feedback. Based on their input, we created a list of potential improvements, leading to revisions that enhanced clarity, while maintaining the total number of items (80).

The cognitive interviews indicated that participants found the use of the present tense confusing, so we revised all items to the past tense for subsequent versions. Participants also emphasised grouping all the items that shared the same opening phrase to improve readability and help process the questions more easily. Therefore, items beginning with phrases like "The disruption made me feel…" or "During the disruption, I felt…" were presented together in blocks. Additionally, participants found some items too vague, so we clarified the wording to improve interpretability.

3.2.2. Survey

In the next stage of our process, we designed an online survey using the Qualtrics platform[1] to collect data for EFA and item reduction Before launching the survey, we conducted a pre-test to identify potential issues in the protocol and estimate the time required to finish the survey. Seventeen participants ($n = 17$) completed the pre-test survey, leading to minor changes in the instructions. For instance, we further clarified key terminology used in the survey by adding more concise definitions (e.g., specifying what was meant by "disruption"). On average, participants took approximately 8 min to complete the pre-test.

Participants. We recruited participants using Prolific[2] and reimbursed them with £1.20.[3] Crowdsourcing platforms such as Prolific and Amazon Mechanical Turk are widely used in scale development (Ford and Scandura, 2023) and are recognised as versatile and effective tools for participant recruitment. We required participants to speak English fluently, use public transport for commuting to work, and have experienced at least one public transport disruption, ensuring they were familiar with the topic. We included two attention checks to ensure data quality, which also aligns with Prolific's recommendations. These were simple questions designed to assess whether participants were reading items carefully and following instructions (Oppenheimer et al., 2009). One item instructed participants to select "strongly agree" regardless of content, and another presented an obviously false statement ("I live on a cloud and commute to work by riding a rainbow"). We excluded participants who failed the attention checks within the survey ($n = 7$). Moreover, we informed all participants that study participation was voluntary and they could leave at any point if they felt uncomfortable. We also informed them that the data collected would be anonymised. The survey was conducted online and could be completed in 8 min.

In the end, we collected a total of $n = 350$ valid responses, which is considered "good" for scale development (Clark and Watson, 2016; Comrey, 1988). The participants had an average age of 30.9 years ($SD = 8.4$ years), with 172 identifying as female (49.1%) and 178 as male. Their nationalities included 226 Europeans, 92 Africans, 15 Asians, 6 North Americans, 3 South Americans and 3 Oceanic. In terms of current residence, 247 participants lived in Europe, 85 in Africa, 10 in Asia, 4 in North America, 2 in South America, and 2 in Oceania. When asked about their public transport usage, 52 % of the participants reported using it daily, 33 % used it 3–4 times per week, 10 % used it 1–2 times per week, 4 % used it 1–2 times per month and 0.3 % only once every 2–3 months.

[1] https://www.qualtrics.com/
[2] https://www.prolific.com/
[3] Equivalent to an hourly rate of £9

Apparatus. We used the Qualtrics platform to design the online survey; the full EFA survey is available in the supplementary material. To evaluate the items, we asked participants to recall a public transport disruption they had experienced and provide a brief description of the situation. Reflecting on the described disruption, they then rated their agreement with each item from our final list on a 7-point Likert scale ranging from “strongly agree” to “strongly disagree.” The statements were randomised for each participant to reduce potential order effects.

3.2.3. Exploratory factor analysis

We employed EFA to iteratively reduce the items of the scale as described by Boateng et al. (2018). Moreover, we explored the scale’s factor structure and dimensionality at this stage. For factor analysis and related computations, we used the packages `pandas`, `numpy`, `matplotlib`, `factor_analyzer`, `pingouin`, and `scipy`. For detailed formulas and computational procedures, see Appendix A.

Initial examination and data suitability. Before conducting the EFA, we reviewed the item pool to ensure suitability for factor analysis and to improve the quality of the final scale. Following guidance from Classical Test Theory (CTT), we removed items that performed poorly to achieve a reliable and efficient scale (Boateng et al., 2018).

We first inspected the *inter-item correlation matrix*, to assess how strongly each item is related to the others. Items with correlations below 0.3 were considered weak and less desirable because they have insufficient contribution to measuring the same latent construct (Boateng et al., 2018; Carpenter, 2018). These weak items can reduce the consistency of the scale and were therefore removed. Moreover, we also looked for items that were too strongly correlated (above 0.7), as they could suggest multicollinearity and potential redundancy between items (Bergqvist et al., 2020; Field, 2012). In the end, the inspection of the inter-item correlation matrix resulted in the removal of four items.[4]

We then assessed the *item-total correlation matrix* to evaluate each item’s contribution to the overall scale. Items with correlations below 0.3 were excluded for their limited contribution to the overall construct and are therefore unlikely to support reliable factor structure development (Boateng et al., 2018; Carpenter, 2018). This step resulted in the deletion of an additional 15 items, leaving us with 61 items.

Afterwards, we confirmed data suitability using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett’s sphericity test. KMO = 0.938, and Bartlett’s test of sphericity results reached significance ($p < .01$). Hence, the data was considered appropriate for conducting EFA (Bartlett, 1951; Kaiser, 1970), and all items were retained for further analysis.

Factor Structure. We conducted an EFA on the 61 items to examine the underlying factor structure. To determine the optimal number of factors, we assessed both the scree plot and eigenvalues. The examination of the eigenvalues indicated that 12 factors had an eigenvalue above 1. However, many methodologists caution against relying solely on the eigenvalues, as it is sensitive to the number of items and can lead to over- or under-extraction (Carpenter, 2018; Field, 2012; Zwick and Velicer, 1986). Therefore, we examined the scree plot, which is often regarded as a more reliable indicator (Pett et al., 2003; Reise et al., 2000). The scree plot suggested an optimal solution of six factors.

We also explored models with five, seven, eight, and nine factors to align with our initial conceptual framework and best practices that recommend considering alternative nearby solutions during EFA (Worthington and Whittaker, 2006). These alternatives produced less coherent and less interpretable results, whereas the six-factor solution yielded the clearest structure with stronger loadings. We therefore selected the six-factor model as both empirically and theoretically most appropriate.

To better interpret the factor structure, we applied Oblimin rotation, which allows for potential correlation between factors (Fabrigar et al., 1999; Field, 2012; Osborne, 2014). This choice was guided by the theoretical expectation that the factors would be interrelated. The Oblimin rotation was performed with the gamma (γ) parameter set to 0, indicating a moderate degree of obliqueness (Clark and Watson, 2016). The choice of gamma was based on standard practice to allow moderate correlations without overfitting.

We continued our analysis by examining the factor loadings. Consistent with established guidelines (Boateng et al., 2018; Carpenter, 2018; Stevens et al., 2002), we adopted an iterative approach, initially removing items with loadings below 0.40 or with cross-loadings above 0.40 as they are generally considered inadequate.

Following Worthington and Whittaker (2006), we further reduced the number of items to enhance the scale’s practical utility while maintaining validity. Prior studies have successfully demonstrated the feasibility of this approach (Woźniak et al., 2023, 2021), showing that reliable measurement can be achieved with fewer items. Although longer scales may increase internal consistency, they can compromise response quality under fatigue or situational stress (Worthington and Whittaker, 2006). As our scale is intended for use during or after stressful disruptions, lengthy questionnaires could overburden travellers. We therefore aimed to make the scale as short as possible while preserving measurement quality. To balance these considerations, we removed items with the lowest factor loadings and those lacking conceptual coherence with the underlying construct, as suggested by Worthington and Whittaker (2006). To ensure sufficient reliability, we followed recommendations to retain at least three items per factor (Costello and Osborne, 2019).

This resulted in an interpretable factor structure and a practically applicable final set of 18 items, with three items per factor. Table 3 shows the factor loadings after Oblimin rotation. The theoretical model explained 56.1 % of the variance, and all items—except one—had a commonality above 0.40, which is considered adequate (Costello and Osborne, 2019). One slightly lower item (0.36) was retained, given the model’s overall robustness and the importance of maintaining three items per factor (Field, 2012).

[4] Four pairs of items showed inter-item correlations above 0.7. In each pair, one item was removed.

Table 3
Overview of the factors and items of the EXID scale following item reduction and EFA. Items are ranked on a 7-point Likert scale ranging from "strongly agree" to "strongly disagree". We report Cronbach's α for the full scale and each factor, while factor loadings are provided for each item within these factors.

Factor	#	Item	Loading	Cronbach's α
Agency	Q1	During the disruption, I felt in control of the situation (R)	0.74	0.80
	Q2	During the disruption, I felt capable of finding solutions (R)	0.74	
	Q3	During the disruption, I felt able to plan ahead (R)	0.77	
Anxiety	Q4	The disruption made me feel worried	0.67	0.83
	Q5	The disruption made me feel emotionally uncomfortable	0.76	
	Q6	The disruption made me feel nervous	0.79	
Frustration	Q7	The disruption made me feel unsatisfied	0.70	0.80
	Q8	The disruption made me feel angry	0.63	
	Q9	The disruption made me feel annoyed	0.88	
Disorientation	Q10	The disruption made me feel it was hard to decide what action to take	0.68	0.79
	Q11	The disruption made me feel I did not know where to go	0.81	
	Q12	The disruption made me feel lost	0.67	
Time-related stress	Q13	The disruption made me feel hurried	0.72	0.78
	Q14	The disruption made me feel time pressured	0.84	
	Q15	The disruption made me feel managing my time was difficult	0.55	
Travel behaviour change	Q16	The disruption made me think twice about the next journey I will take	0.79	0.82
	Q17	The disruption made me think of changing the way I travel	0.75	
	Q18	The disruption made me avoid this route in the future	0.71	
Total Cronbach's α				0.88

Reliability Analysis. Subsequently, we examined Cronbach's α for each factor to test the scale's reliability (Field, 2012). The 18-item scale showed good internal consistency with a total Cronbach's α of 0.88. The Cronbach's α for the subscales ranged from 0.78 to 0.83 (see Table 3).

3.3. Phase 3: Scale evaluation

This phase assesses the scale's dimensionality and hypothesised structure using CFA. Moreover, we evaluate construct validity by testing the scale's ability to differentiate between 'known groups'. A second survey was conducted to assess the temporal stability through test-retest reliability.

3.3.1. Survey

In the next stage of our process, we designed an online survey using the Qualtrics platform to collect data for CFA and to measure construct validity and reliability.

Participants. We recruited a total of $n = 209$ participants for an online survey. This sample size is considered appropriate for CFA,[5] as it adheres to the commonly recommended guideline of 10 participants per item (Boateng et al., 2018), and aligns with sample sizes used in similar scale development studies (Bentvelzen et al., 2021; Ettema et al., 2011; Woźniak et al., 2021). We used Prolific to recruit participants and compensated them with £0.45.[6] We required participants to speak English fluently, use public transport for work, and have experienced at least one public transport disruption, ensuring they were sufficiently familiar with the topic. We informed them that participation was voluntary, they could withdraw at any time, and their data would be anonymised. The survey could be completed in 3 min.

The participants had an average age of 30.4 years ($SD = 8.2$ years), with 106 identifying as female (50.7 %) and 103 as male. Their nationalities included 122 Europeans, 17 Asians, 39 Africans, 6 North Americans, and 25 South Americans. In terms of current residence, 134 participants lived in Europe, 35 in Africa, 25 in South America, 10 in Asia, 4 in North America and 1 in Oceania. When asked about public transport usage, 56 % of the participants reported using it daily, 27 % used it 3–4 times per week, 12 % used it 1–2 times per week, 4 % used it 1–2 times per month, and 1 % used it only a few times per year.

Apparatus. We used the Qualtrics platform to design the online survey; the full CFA survey is available in the supplementary material. To evaluate the scale, we developed two scenarios depicting a low-impact and a high-impact public transport disruption (see Table 4). We assumed that a disruption involving multiple cancellations and greater overall inconvenience would lead to a higher perceived impact. Consequently, we created these two groups to test whether the scale could effectively distinguish between different levels of

[5] The survey used for the CFA consisted of the 18 items shown in Table 3.
[6] Equivalent to an hourly rate of £9

Table 4
The scenarios used in the CFA survey. Participants were asked to rank their perception of the disruption on the EXID scale.

Severity level	Scenario description
Low	Imagine you are travelling with public transport when an announcement comes on saying there will be a short delay due to a minor technical issue. You will need to wait for about 10 min, but you are seated. After a brief wait, the vehicle resumes its journey, and you arrive only slightly later than planned.
High	Imagine you are travelling with public transport when there is a sudden disruption. The vehicle comes to a complete stop and you are informed over the loudspeaker that all services have been suspended due to an emergency. You are asked to disembark and find alternative transportation. It is very crowded, and the noise levels increase as people start to look for other options. Unfortunately, you are far from other transport options, and there is no clear guidance on alternative routes. With limited options available, the chance of arriving much later is likely.

Table 5
Comparison of fit indices across four confirmatory factor analysis models for the EXID scale. The single-order correlated factors model demonstrated the best overall fit. **Bold** values indicate the best or equal-best performance for each fit index.

Model	χ^2(df)	p-value	CFI	TLI	RMSEA	SRMR
Unidimensional			0.75	0.72	0.13	0.09
Single-order (correlated)	**247.58 (120)**	< 0.001	**0.93**	**0.91**	**0.07**	**0.06**
Second-order (correlated)	273.41 (129)	< 0.001	0.92	0.90	**0.07**	0.07
Multi-hierarchical	270.11 (126)	< 0.001	0.92	0.90	**0.07**	0.07

severity. Unlike in the EFA survey, we did not rely on participants' past experiences; instead, we presented hypothetical disruptions to ensure there was a clear contrast in impact. This approach allowed us to further verify whether the scale could successfully differentiate between 'known groups', thus groups that are already expected to differ on the measured construct (as described by Boateng et al. (2018)). This method is commonly used to evaluate the construct validity of scales (Bentvelzen et al., 2021; Ettema et al., 2011; Woźniak et al., 2023, 2021).

We randomly presented each participant with one of the two scenarios and a short description: *"We would like you to read the scenario below carefully"*. Afterwards, we asked participants to *"imagine you were in the scenario described above"* and to indicate how much they agreed with each item of our final scale about the presented disruption on a 7-point Likert scale (strongly agree to strongly disagree).

3.3.2. Confirmatory factor analysis

To investigate the correctness of the model created in the EFA, we conducted a CFA. The CFA aimed to assess the degree to which the observed data fit several measurement models. We employed Python 3.9 using Jupyter Notebook for data analysis. For factor analysis and related computations, we used the packages `pandas`, `numpy`, `matplotlib`, `factor_analyzer`, `pingouin`, `scipy`, and `semopy`. For detailed formulas and computational procedures, see Appendix A.

We assessed model fit using a set of fit indices as recommended by established guidelines (Boateng et al., 2018; Hu and Bentler, 1999). Specifically, the overall fit of the model was checked by the chi-square test (χ^2), followed by the comparative fit index (CFI), Tucker-Lewis index (TLI), Root Mean Square Error of Approximation (RMSEA) and Standardised Root Mean Square Residual (SRMR). While a non-significant χ^2 is a conventional index of absolute fit, its sensitivity to sample size and data characteristics limits its reliability as a sole indicator (Alavi et al., 2020; Boateng et al., 2018). Cutoff values were based on recommendations from the literature (Bentler and Bonett, 1980; Boateng et al., 2018; Browne and Cudeck, 1992; Hu and Bentler, 1999), with generally accepted thresholds of CFI and TLI ≥ 0.90 for adequate fit and ≥ 0.95 for good fit, RMSEA ≤ 0.08 for acceptable fit and ≤ 0.06 for good fit, and SRMR ≤ 0.08.

In the end, we tested four models: 1) a unidimensional model, 2) a six-factor correlated model, 3) a second-order model, and 4) a multi-hierarchical model (see Table 5).

Our first model was unidimensional, where each of the 18 items loaded onto a single factor representing the impact of disruptions. The model demonstrated poor fit overall, as indicated by a low CFI = 0.75 and TLI = 0.72. Although the RMSEA = 0.13 and the SRMR = 0.09 were within acceptable ranges, the overall pattern of fit indices suggests that the unidimensional model did not fit adequately.

The second model was single-order, correlated, where each of the 18 items loaded onto one of six factors identified through EFA. The hypothesised six-factor model revealed an acceptable fit based on the goodness-of-fit statistics: $\chi^2(120) = 247.58$, $p < 0.001$, CFI = 0.93, TLI = 0.91, RMSEA = 0.07, SRMR = 0.06. This result suggests that the scale was internally consistent. Additionally, the model showed several moderate to high correlations between the latent variables, showing that the overall construct, as proposed, was valid.

The third model represents a second-order model structure. Specifically, six first-order factors are loaded onto a single higher-order latent factor. This model also showed an acceptable fit: $\chi^2(129) = 273.41$, $p < 0.001$, CFI = 0.92, TLI = 0.90, RMSEA = 0.07, and SRMR = 0.07 . While slightly less optimal than the single-order model, the fit statistics suggest that the second-order structure remains an acceptable representation of the data.

Table 6
Overview of the differentiation results between 'known groups' using Mann-Whitney U tests. Bonferroni-adjusted α levels of 0.007.

Scale/subscale	Mrankminor	SDrankminor	Mrankmajor	SDrankmajor	U	z	p
EXID	76.37	19.01	89.26	15.20	7594.5	4.88	< 0.001
EXID (agency)	12.81	3.93	12.83	3.49	5594.5	0.31	0.760
EXID (anxiety)	13.47	4.69	15.36	3.80	6719.5	2.88	< 0.007
EXID (frustration)	15.11	4.01	17.88	2.86	7710.0	5.15	< 0.001
EXID (disorientation)	9.76	4.76	13.70	4.01	8004.0	5.82	< .001
EXID (time-related)	14.46	4.07	16.62	2.85	7231.5	4.05	< .001
EXID (travel behaviour)	10.76	4.74	12.88	4.04	6879.0	3.25	< .007

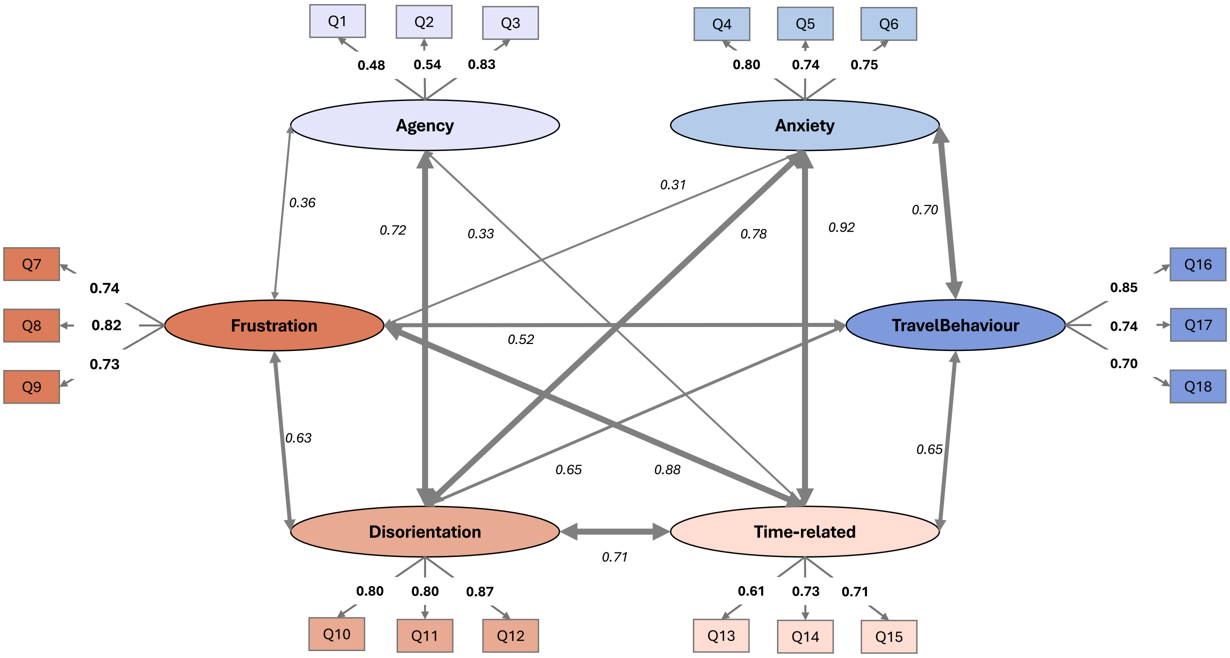

Fig. 2. Path diagram illustrating the relationships between scale factors (ellipses) and scale items (rectangles). Bold numbers indicate factor loadings between factors and items, while italicised numbers represent correlations between latent factors. Arrow thickness corresponds to the magnitude of the relationships, with thicker lines indicating stronger correlations.

The fourth model represents a multi-hierarchical structure. Given the flexibility in specifying hierarchical relationships, we tested several theoretically plausible configurations. The final structure was selected based on its conceptual coherence and performance. Specifically, we grouped factors related to immediate emotional reactions, namely frustration, anxiety, and time-related stress, under a single higher-order factor. Moreover, we grouped cognitive responses such as disorientation and agency together under a separate factor. Long-term impacts were modelled as a distinct factor, capturing more enduring effects of disruption. This configuration resulted in an acceptable model fit: $\chi^2(126) = 270.11$, $p < 0.001$, CFI = 0.92, TLI = 0.90, and RMSEA = 0.07, and SRMR = 0.07.

Among the evaluated models, the single-order six-factor structure proved to be the most appropriate, both theoretically and empirically. While the second-order and multi-hierarchical models offered conceptually plausible alternatives and demonstrated acceptable fit, they introduced additional complexity without providing meaningful improvement in model performance. The single-order model preserved the distinctiveness of each construct and allowed for a more straightforward interpretation of their interrelationships, making it the most suitable representation of the EXID scale's structure. Fig. 2 shows the final CFA model.

3.3.3. *Construct validity*

Next, we evaluate the scale's construct validity, i.e., whether the scale could differentiate between 'known groups'. We compared the perceived impact of the low-severity versus the high-severity disruption. We hypothesised that a high-severity disruption would be significantly more impactful than a low-severity disruption. Shapiro-Wilk tests revealed that data in both the low-severity and high-severity disruption groups were not normally distributed. Thus, we applied non-parametric statistics. The Mann-Whitney U test with a Bonferroni-adjusted α level of 0.007 (0.05/7) indicated that the low-severity disruptions were perceived as significantly less impactful than the high-severity disruptions ($p < .007$). Looking at the individual subscales, these significant differences appear for every subscale except 'Agency'. Table 6 provides an overview of these results.

To assess the stability of internal consistency across scenarios, Cronbach's α was calculated separately for each subscale in the low- and high-severity disruption conditions. The reliability estimates were largely consistent, with most subscales showing acceptable to good internal consistency in both contexts. Anxiety ($\alpha = 0.74$ vs. .84), Disorientation ($\alpha = 0.81$ vs. .86), and Behaviour ($\alpha = 0.78$ vs. .82) showed slightly higher reliability under high-severity conditions, suggesting these constructs may be more consistently expressed when disruptions are more impactful. Agency ($\alpha = 0.64$ vs. .69) and Time ($\alpha = 0.65$ vs. .72) also showed modest increases across conditions but remained in the lower acceptable range. Frustration remained relatively stable ($\alpha = 0.78$ vs. .77). These findings support the scale's internal reliability across varying levels of disruption severity.

3.3.4. Test-retest reliability

To assess the consistency of the EXID scale over time, we conducted a test-retest reliability study using an online survey administered via Qualtrics; the full test-retest survey is available in the supplementary material. The same participants completed the survey twice, with a minimum 14-day interval between sessions.

We recruited a sample of $n = 22$ participants (15 female, 7 male), with an average age of $M = 29$ years ($SD = 5.31$). In both surveys, participants were presented with the high-severity scenario description (Table 4), as we expected its impact to be stronger and less affected by external factors. A low-severity disruption might lead to more variable responses, making it harder to assess whether the scale produces stable results over time. Before responding to the items, they received the following description: "*We would like you to imagine you were in the scenario described above. Please respond to the statements below.*"

As suggested by Boateng et al. (2018), we used the intra-class correlation coefficient to measure test-retest reliability. In line with recommendations by Koo and Li (2016), we used a single rating, absolute-agreement, 2-way mixed-effects model with a Bonferroni-adjusted α level of 0.007 (0.05/7). We obtained substantial reliability for the full scale with EXID: $\kappa = 0.79$, $p < 0.001$, good reliability. Among the subscales, moderate and statistically significant reliability was observed for EXID (frustration), $\kappa = 0.73$, $p < 0.001$; EXID (disorientation), $\kappa = 0.62$, $p < 0.001$; and EXID (anxiety), $\kappa = 0.58$, $p < 0.002$.

In contrast, reliability for the remaining subscales ranged from moderate to poor but was not considered statistically significant: EXID (time-related), $\kappa = 0.58$, $p < 0.01$; EXID (travel behaviour), $\kappa = 0.50$, $p < 0.01$; and EXID (agency), $\kappa = 0.40$, $p < 0.02$.

4. Discussion

In this section, we detail the potential use cases of the EXID scale (Section 4.1) and discuss its limitations (Section 4.2). The results of our study indicate that there are six factors for measuring the experienced impact of disruptions in public transport: *agency*, *anxiety*, *frustration*, *disorientation*, *time-related stress*, and *travel behaviour change*.

Our analysis showed that these latent variables are not independent of each other. Instead, almost all are significantly correlated, supporting the expectation that the different aspects of the experienced impact of public transport disruption are interconnected. A particularly strong correlation was found between *Time-related stress* and *Anxiety*, and *Frustration* and *Time-related stress*. This suggests that the experience of time-related stress during a disruption is related to feelings of anxiety and frustration, as also reflected in the literature (Cheng, 2010; Rezapour and Ferraro, 2021). We emphasise that, conceptually, these factors can be distinguished from one another, but they are closely related and can influence each other. Rather than indicating redundancy, these correlations reflect how feelings of time-related stress, anxiety, and frustration reinforce each other during a disruption. We assume that these feelings can exist independently, but the nature of the disruption and the specific situation influence both their intensity and the extent to which they become interconnected. Importantly, the CFA results showed good model fit, supporting the idea that these are distinct but strongly related constructs rather than overlapping measures of the same concept.

In terms of construct validity, our results show that the EXID scale is able to capture differences in perceived impact. However, the factor *Agency* did not show a significant difference between the minor and major disruption scenarios. Earlier research has shown that travellers often feel powerless or lacking control in unpredictable situations (Evans et al., 2002; Koslowsky et al., 2013; Sposato et al., 2012). We believe that even though the scenarios differed in impact, travellers in both cases might have felt equally limited agency about the situation, leading to similar scores. The minor disruption scenario, while less severe, could still have created a sense of powerlessness as travellers had to wait without clear information on whether or when the situation would improve. In the major disruption scenario, in contrast, travellers might have felt a lack of control due to the urgency and lack of clear guidance. In both scenarios, while the impact of the disruption differs, the lack of control and the absence of clear, reassuring information may have similarly affected participants" perceptions of their agency.

Lastly, the full EXID scale demonstrated good test-retest reliability. However, the reliabilities of the subscales were more variable, with some demonstrating significant moderate reliability, while others showed non-significant moderate or poor reliability. This could indicate that some constructs may be measured less consistently over time. Yet, the relatively small number of items per subscale and the relatively small test-retest sample could have increased variability in reliability levels. Future research should further explore the additional aspects of temporal stability of the subscales with larger samples.

4.1. Potential use cases

The EXID scale measures the experienced impact of public transport disruptions, capturing multiple facets of this experience. It can provide valuable insights into traveller experiences and how they vary across different traveller groups, contexts, and types of disruptions. These insights can support public transport providers, infrastructure operators, authorities, and policymakers when planning disruptions (e.g., for maintenance), implementing mitigation strategies to make disruptions less impactful for travellers, and enhancing public transport resilience. The scale can be applied in various contexts: (1) immediately after a traveller experiences a disruption, (2) in studies where travellers recall past disruptions, or (3) in vignette studies that present hypothetical travel scenarios involving a disruption.

The scale can also be used to evaluate interventions designed to improve the disruption experience, such as assessing the effectiveness of real-time information and alternative transport options. A significant body of research has explored personalised passenger information systems that automatically adapt to individual travellers and their specific context (Vredenborg et al., 2025). This scale could be applied to assess the impact of such systems during disruptions, allowing for comparisons of traveller experiences with different variants of support, or for evaluating changes in experience before and after an intervention has been implemented.

Disruptions are not limited to public transport, and the scale may also be applicable in other transport related contexts. For instance, it could assess the impact of disruptions such as road blockages on private transport and the effectiveness of navigation systems in providing relevant information. The scale's items appear well-suited for this purpose.

4.2. Limitations and future directions

We acknowledge several limitations in our study. First, our initial conceptual dimensions were not derived from a systematic literature review. As a result, the literature review is not exhaustive, and we do not claim to have captured all relevant studies of travellers' disruption experiences. However, we are confident that the high-level dimensions identified in our work capture the key aspects of experienced impact of disruption. This confidence is supported by the expert reviews, which affirmed the conceptual soundness of these dimensions.

Second, the scale has been developed and validated using participants recruited through Prolific, which introduces self-selection and demographic biases. The sample consisted predominantly of participants in their 20s and 30s, and Prolific users tend to have higher education levels than the general population. Additionally, participants had to speak English fluently. Future research could explore to what extent the scale items are also understandable to those with more limited English language skills, lower education levels, and older adults. Additionally, we did not specifically examine the disruption experiences of particular user groups, such as individuals with severe special needs, marginalised communities, or those living in particularly unsafe areas. Future studies could investigate whether these groups experience additional factors not captured by the scale.

Lastly, the scale was specifically developed to measure the experienced impact of disruptions in public transport. Whilst some or all of its components may also apply to other disruption contexts, further research is needed to validate its use beyond public transport. Additionally, future studies could explore whether there are context-specific factors not captured by the scale that may be relevant in other types of disruptions.

In our future studies, we will use the scale to investigate how personalised passenger information during disruptions affects their experience.

5. Conclusion

In this paper, we proposed a measurement instrument—EXperienced Impact of Disruptions Scale (EXID)—that allows researchers and practitioners to measure the experienced impact of disruptions on public transport travellers. To this end, we presented the development and evaluation of the EXID scale, following the structured process by Boateng et al. (2018). Based on a literature review on disruption experiences, expert interviews and cognitive interviews, we constructed and reduced the scale. We refined and extensively evaluated the scale with two surveys, illustrating its discriminant validity, its ability to differentiate between known groups, and its consistency over time through test-retest reliability.

The EXID scale has been developed specifically to evaluate interventions designed to improve the disruption experience for the public transport domain. It is possible that the scale may also be applicable in other transport-related contexts where disruptions occur. However, additional research is needed to investigate its usefulness in other contexts and to investigate what additional factors may need to be included.

The scale is suitable for researchers and practitioners alike, for instance, to quantitatively measure disruptions' impact on people's experience, compare the effects of several disruption management strategies, or assess the effectiveness of a newly implemented mitigation strategy in practice. Beyond its direct practical implications for practitioners, we hope that EXID will benefit and interconnect various research communities beyond the transportation field.

CRediT authorship contribution statement

Marloes Vredenborg: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Marit Bentvelzen:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Conceptualization; **Anouk van Kasteren:** Writing – review & editing, Writing – original draft, Methodology, Investigation; **Christine Bauer:** Writing – review & editing, Writing – original draft, Supervision; **Judith Masthoff:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Data availability

Data will be made available on request.

Acknowledgment

This publication was partially funded by the Dutch Railways. This publication was also supported by the Excellence in Digital Sciences and Interdisciplinary Technologies (EXDIGIT) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.

Appendix A. Formulas used in data analysis

Kaiser–Meyer–Olkin (KMO) Measure

To assess sampling adequacy for factor analysis, we used the Kaiser–Meyer–Olkin (KMO) measure. It is calculated as the ratio of the sum of squared correlations to the total of squared correlations and squared partial correlations (excluding diagonals):

$$\mathrm{KMO} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2} \tag{A.1}$$

where:

- r_{ij} is the correlation coefficient between items i and j,
- p_{ij} is the partial correlation between items i and j, computed as:

$$p_{ij} = -\frac{R_{ij}^{-1}}{\sqrt{R_{ii}^{-1} R_{jj}^{-1}}} \tag{A.2}$$

where R^{-1} is the inverse of the correlation matrix.

This implementation corresponds to the method used in the `factor_analyzer` package in Python, which excludes diagonal elements from both correlation and partial correlation matrices during computation. A value close to 1 indicates high factorability of the data, with thresholds interpreted as follows:

- 0.90s – marvellous,
- 0.80s – meritorious,
- 0.70s – middling,
- 0.60s – mediocre,
- below 0.50 – unacceptable.

Bartlett's Test of Sphericity

Bartlett's test evaluates whether the observed correlation matrix **R** significantly differs from the identity matrix, indicating the presence of latent factors. The test statistic is computed as:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \cdot \ln |\mathbf{R}| \tag{A.3}$$

Where:

- n is the number of observations,
- p is the number of variables,
- $|\mathbf{R}|$ is the determinant of the correlation matrix.

The degrees of freedom are:

$$\mathrm{df} = \frac{p(p-1)}{2} \tag{A.4}$$

The resulting χ^2 statistic is compared against a chi-square distribution to compute a p-value:

$$p = P(\chi^2 > \text{observed}) = \texttt{chi2.sf}(\chi^2, \mathrm{df}) \tag{A.5}$$

This implementation follows the procedure used in the `factor_analyzer` Python package.

Oblimin Factor Rotation

In exploratory factor analysis, we applied an Oblimin rotation, an oblique method that allows extracted factors to correlate. The objective function minimised in Oblimin rotation is:

$$\Phi(\Lambda) = \sum_{i<j} \left(\sum_k \lambda_{ik} \lambda_{jk} \right)^2 + \gamma \sum_j \left(\sum_i \lambda_{ij}^2 \right)^2 \tag{A.6}$$

Where:

- Λ is the factor loading matrix,
- λ_{ij} is the loading of item i on factor j,
- γ controls the simplicity penalty on factors (we used $\gamma = 0$, corresponding to the default Oblimin method in Python).

We implemented the Oblimin rotation using the `Rotator` class from the `factor_analyzer` Python package.

Confirmatory Factor Analysis Fit Indices

We report the following fit indices for Confirmatory Factor Analysis (CFA):

- Root Mean Square Error of Approximation (RMSEA):

$$\text{RMSEA} = \sqrt{\frac{\chi^2 - \text{df}}{\text{df}(n-1)}} \tag{A.7}$$

  where χ^2 is the chi-square statistic, df is degrees of freedom, and n is the sample size. RMSEA assesses how well the model, with unknown but optimally chosen parameter estimates, fits the population covariance matrix. Lower values indicate a better fit.
- Comparative Fit Index (CFI):

$$\text{CFI} = 1 - \frac{\max(\chi^2 - \text{df}, 0)}{\max(\chi^2_{\text{null}} - \text{df}_{\text{null}}, 0)} \tag{A.8}$$

  where χ^2 and df are for the tested model, and χ^2_{null} and df_{null} are for the null (independence) model. CFI compares the fit of the tested model relative to a baseline model; values closer to 1 indicate a better fit.
- Tucker–Lewis Index (TLI):

$$\text{TLI} = \frac{\chi^2_{\text{null}}/\text{df}_{\text{null}} - \chi^2/\text{df}}{\chi^2_{\text{null}}/\text{df}_{\text{null}} - 1} \tag{A.9}$$

  The TLI adjusts for model complexity by accounting for degrees of freedom. Like CFI, values closer to 1 indicate better model fit. TLI can exceed 1 or fall below 0 in cases of small sample sizes or poor-fitting models.
- Standardised Root Mean Square Residual (SRMR) is calculated as the standardised difference between the observed and predicted correlations, representing the average discrepancy; lower values indicate a better fit.

Intraclass Correlation Coefficient (ICC)

Used to assess test–retest reliability, the ICC for a two-way mixed-effects model is calculated as:

$$\text{ICC}(3,1) = \frac{\text{MS}_R - \text{MS}_E}{\text{MS}_R + (k-1)\text{MS}_E} \tag{A.10}$$

where MS_R is the mean square for rows (participants), MS_E is the mean square error (error variance), and k is the number of measurements (test occasions).

ICC values range from 0 to 1, with values closer to 1 indicating higher reliability (i.e., more consistency across repeated measurements), while values near 0 suggest low reliability.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.trf.2025.103418

References

Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., Cleary, M. L., 2020. Chi-square for model fit in confirmatory factor analysis. *Journal of Advanced Nursing 76* (9), 2209–2211. https://doi.org/10.1111/jan.14399

Armougum, A., Gaston-Bellegarde, A., Joie-La Marle, C., Piolino, P., 2020. Physiological investigation of cognitive load in real-life train travelers during information processing. *Applied Ergonomics 89*, 103180. https://doi.org/10.1016/j.apergo.2020.103180

Bakti, I. G. M. Y., Sumaedi, S., 2015. P-TRANSQUAL: A service quality model of public land transport services. *International Journal of Quality & Reliability Management 32* (6), 534–558. https://doi.org/10.1108/IJQRM-06-2013-0094

Bartlett, M. S., 1951. The effect of standardization on a χ 2 approximation in factor analysis. *Biometrika 38* (3/4), 337–344.

Beatty, P. C., Willis, G. B., 2007. Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly 71* (2), 287–311. https://doi.org/10.1093/poq/nfm006

Beirão, G., Cabral, J. A. S., 2007. Understanding attitudes towards public transport and private car: A qualitative study. *Transport Policy 14* (6), 478–489. https://doi.org/10.1016/j.tranpol.2007.04.009

Bentler, P. M., Bonett, D. G., 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin 88* (3), 588. https://doi.org/10.1037/0033-2909.88.3.588

Bentvelzen, M., Niess, J., Woźniak, M. P., Woźniak, P. W., 2021. The development and validation of the technology-supported reflection inventory. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445673

Bergqvist, E., Tossavainen, T., Johansson, M., 2020. An analysis of high and low intercorrelations between mathematics self-efficacy, anxiety, and achievement variables: A prerequisite for a reliable factor analysis. *Education Research International 2020* (1). https://doi.org/10.1155/2020/8878607

Bhat, C. R., Sardesai, R., 2006. The impact of stop-making and travel time reliability on commute mode choice. *Transportation Research Part B: Methodological 40* (9), 709–730. https://doi.org/10.1016/j.trb.2005.09.008

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., Young, S. L., 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health 6*. https://doi.org/10.3389/fpubh.2018.00149

Bollini, A. M., Walker, E. F., Hamann, S., Kestler, L., 2004. The influence of perceived control and locus of control on the cortisol and subjective responses to stress. *Biological Psychology 67* (3), 245–260. https://doi.org/10.1016/j.biopsycho.2003.11.002

Bougie, R., Pieters, R., Zeelenberg, M., 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science 31* (4), 377–393. https://doi.org/10.1177/0092070303254412

Browne, M. W., Cudeck, R., 1992. Alternative ways of assessing model fit. *Sociological Methods & Research 21* (2), 230–258. https://doi.org/10.1177/0049124192021002005

Cantwell, M., Caulfield, B., O'Mahony, M., 2009. Examining the factors that impact public transport commuting satisfaction. *Journal of Public Transportation 12* (2), 1–21. https://doi.org/10.5038/2375-0901.12.2.1

Caro, L. M., García, J. A. M., 2007. Measuring perceived service quality in urgent transport service. *Journal of Retailing and Consumer Services 14* (1), 60–72. https://doi.org/10.1016/j.jretconser.2006.04.001

Carpenter, S., 2018. Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures 12* (1), 25–44. https://doi.org/10.1080/19312458.2017.1396583

Carreira, R., Patrício, L., Jorge, R. N., Magee, C., 2014. Understanding the travel experience and its impact on attitudes, emotions and loyalty towards the transportation provider–a quantitative study with mid-distance bus trips. *Transport Policy 31*, 35–46. https://doi.org/10.1016/j.tranpol.2013.11.006

Casado Diaz, A. B., Más Ruíz, F. J., 2002. The consumer's reaction to delays in service. *International Journal of Service Industry Management 13* (2), 118–140. https://doi.org/10.1108/09564230210425331

Cheng, Y.-H., 2010. Exploring passenger anxiety associated with train travel. *Transportation 37*, 875–896. https://doi.org/10.1007/s11116-010-9267-z

Clark, L. A., Watson, D., 2016. Constructing validity: Basic issues in objective scale development. *sychological Assessment 7* (3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Comrey, A. L., 1988. Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology 56* (5), 754. https://doi.org/10.1037/0022-006X.56.5.754

Costello, A. B., Osborne, J., 2019. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation 10* (1), 7.

Cox, T., Houdmont, J., Griffiths, A., 2006. Rail passenger crowding, stress, health and safety in Britain. *Transportation Research Part A: Policy and Practice 40* (3), 244–258. https://doi.org/10.1016/j.tra.2005.07.001

Currie, G., Muir, C., 2017. Understanding passenger perceptions and behaviors during unplanned rail disruptions. *Transportation Research Procedia 25*, 4392–4402. https://doi.org/10.1016/j.trpro.2017.05.322

DeVellis, R. F., Thorpe, C. T., 2021. Scale development: Theory and applications. Sage publications.

Dewulf, B., Neutens, T., Van Dyck, D., De Bourdeaudhuij, I., Van de Weghe, N., 2012. Correspondence between objective and perceived walking times to urban destinations: Influence of physical activity, neighbourhood walkability, and socio-demographics. *International Journal of Health Geographics 11*, 1–10. https://doi.org/10.1186/1476-072X-11-43

Dijkstra, M. T. M., Homan, A. C., 2016. Engaging in rather than disengaging from stress: effective coping and perceived control. *Frontiers in Psychology 7*, 1415. https://doi.org/10.3389/fpsyg.2016.01415

Drabicki, A. A., Islam, M. F., Szarata, A., 2021. Investigating the impact of public transport service disruptions upon passenger travel behaviour–results from krakow city. *Energies 14* (16). https://doi.org/10.3390/en14164889

Dubé, L., Schmitt, B. H., Leclerc, F., 1991. Consumers affective response to delays at different phases of a service delivery 1 *Journal of Applied Social Psychology 21* (10), 810–820. https://doi.org/10.1111/j.1559-1816.1991.tb00444.x

Durand, A., 2017. Managing disruptions in public transport from the passenger perspective. Ph.D. thesis. Master thesis. Delft University of Technology.

Dziekan, K., Kottenhoff, K., 2007. Dynamic at-stop real-time information displays for public transport: Effects on customers. *Transportation Research Part A: Policy and Practice 41* (6), 489–501. https://doi.org/10.1016/j.tra.2006.11.006

Ettema, D., Gärling, T., Eriksson, L., Friman, M., Olsson, L. E., Fujii, S., 2011. Satisfaction with travel and subjective well-being: Development and test of a measurement tool. *Transportation Research Part F: Traffic Psychology and Behaviour 14* (3), 167–175. https://doi.org/10.1016/j.trf.2010.11.002

Evans, G. W., Wener, R. E., 2006. Rail commuting duration and passenger stress. *Health psychology 25* (3), 408–412. https://doi.org/10.1037/0278-6133.25.3.408

Evans, G. W., Wener, R. E., 2007. Crowding and personal space invasion on the train: Please don't make me sit in the middle. *Journal of Environmental Psychology 27* (1), 90–94. https://doi.org/10.1016/j.jenvp.2006.10.002

Evans, G. W., Wener, R. E., Phillips, D., 2002. The morning rush hour: Predictability and commuter stress. *Environment and Behavior 34* (4), 521–530. https://doi.org/10.1177/00116502034004007

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., Strahan, E. J., 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods 4* (3), 272. https://doi.org/10.1037/1082-989X.4.3.272

Ferreira, J. C., Silva, H., Afonso, J. A., Afonso, J. L., 2018. Context aware advisor for public transportation. *IAENG International Journal of Computer Science 45* (1), 74–81.

Field, A., 2012. Discovering statistics using r.

Ford, L. R., Scandura, T. A., 2023. The SAGE handbook of survey development and application. SAGE Publications Limited.

Friman, M., Edvardsson, B., Gärling, T., 2001. Frequency of negative critical incidents and satisfaction with public transport services. i. *Journal of Retailing and Consumer Services 8* (2), 95–104.

Friman, M., Olsson, L. E., 2023. Are we leaving some people behind? travel autonomy, perceived accessibility, and well-being among people experiencing mental and physical difficulties. *Transportation Research Part F: Traffic Psychology and Behaviour 98*, 243–253. https://doi.org/10.1016/j.trf.2023.08.009

Gault, P., Cottrill, C. D., Corsar, D., Edwards, P., Nelson, J. D., Markovic, M., Mehdi, M., Sripada, S., 2019. Travelbot: Utilising social media dialogue to provide journey disruption alerts. *Transportation Research Interdisciplinary Perspectives 3*. https://doi.org/10.1016/j.trip.2019.100062

Ge, L., Voß, S., Xie, L., 2022. Robustness and disturbances in public transport. *Public Transport 14* (1), 191–261. https://doi.org/10.1007/s12469-022-00301-8

Glasgow, T. E., Geller, E. S., Le, H. T. K., Hankey, S., 2018. Travel mood scale: Development and validation of a survey to measure mood during transportation. *Transportation Research Part F: Traffic Psychology and Behaviour 59*, 318–329. https://doi.org/10.1016/j.trf.2018.09.014

Gobind, J., 2018. Transport anxiety and work performance. *SA Journal of Human Resource Management 16* (1), 1–7.

Grison, E., Burkhardt, J.-M., Gyselinck, V., 2017. How do users choose their routes in public transport? the effect of individual profile and contextual factors. *Transportation Research Part F: Traffic Psychology and Behaviour 51*, 24–37. https://doi.org/10.1016/j.trf.2017.08.011

Grotenhuis, J.-W., Wiegmans, B. W., Rietveld, P., 2007. The desired quality of integrated multimodal travel information in public transport: Customer needs for time and effort savings. *Transport Policy 14* (1), 27–38. https://doi.org/10.1016/j.tranpol.2006.07.001

van Hagen, M., van Oort, N., 2019. Improving railway passengers experience: Two perspectives. *Journal of Traffic and Transportation Engineering 7* (3), 2328–2142. https://doi.org/10.17265/2328-2142/2019.03.001

Hart, S. G., Staveland, L. E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: *Advances in psychology*. Elsevier. *Vol. 52*, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Haywood, L., Koning, M., Monchambert, G., 2017. Crowding in public transport: Who cares and why? *Transportation Research Part A: Policy and Practice 100*, 215–227. https://doi.org/10.1016/j.tra.2017.04.022

He, S. Y., Tao, S., Sun, K. K., 2024. Attitudes towards public transport under extended disruptions and massive-scale transit dysfunction: A hong kong case study. *Transport Policy 149*, 247–258. https://doi.org/10.1016/j.tranpol.2024.02.008

Hess, D. B., Brown, J., Shoup, D., 2004. Waiting for the bus. *Journal of Public Transportation 7* (4), 67–84. https://doi.org/10.5038/2375-0901.7.4.4

Hu, K.-C., Jen, W., 2006. Passengers' Perceived service quality of city buses in taipei: Scale development and measurement. *Transport Reviews 26* (5), 645–662. https://doi.org/10.1080/01441640600679482

Hu, L.-t., Bentler, P. M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal 6* (1), 1–55. https://doi.org/10.1080/10705519909540118

Ibraeva, A., de Sousa, J. F., 2014. Marketing of public transport and public transport information provision. *Procedia-Social and Behavioral Sciences 162*, 121–128. https://doi.org/10.1016/j.sbspro.2014.12.192

Ibrahim, A. N. H., Borhan, M. N., Md Yusoff, N. I., Ismail, A., 2020. Rail-based public transport service quality and user satisfaction–a literature review. *Promet-Traffic & Transportation 32* (3), 423–435. https://doi.org/10.7307/ptt.v32i3.3270

Islam, R., Chowdhury, M. S., Sarker, M. S., Ahmed, S., 2014. Measuring customer's satisfaction on bus transportation. *American Journal of Economics and Business Administration 6* (1), 34–41. https://doi.org/10.3844/ajebasp.2014.34.41

Ittamalla, R., Kumar, D. V. S., 2021. Determinants of holistic passenger experience in public transportation: Scale development and validation. *Journal of Retailing and Consumer Services 61*. https://doi.org/10.1016/j.jretconser.2021.102564

Kaiser, H. F., 1970. A second generation little jiffy. *Psychometrika 35* (4), 401–415. https://doi.org/10.1007/BF02291817

van Kasteren, A., Vredenborg, M., Masthoff, J., 2024. Understanding commuter information needs and desires in public transport: A comparative analysis of stated and revealed preferences. In: *International conference on human-Computer interaction*. Springer, Berlin, Heidelberg, pp. 83–103. https://doi.org/10.1007/978-3-031-60480-5_5

Kökalan, Ö., Tutan, A., 2021. Passenger satisfaction scale for public transportation. *Transportation Research Record 2675* (3), 44–52. https://doi.org/10.1177/0361198120961382

Koo, T. K., Li, M. Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine 15* (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koslowsky, M., Kluger, A. N., Reich, M., 2013. Commuting stress: Causes, effects, and methods of coping. Springer Science & Business Media.

Krygsman, S., Dijst, M., Arentze, T., 2004. Multimodal public transport: An analysis of travel time elements and the interconnectivity ratio. *Transport Policy 11* (3), 265–275. https://doi.org/10.1016/j.tranpol.2003.12.001

Lai, W.-T., Chen, C.-F., 2011. Behavioral intentions of public transit passengers–the roles of service quality, perceived value, satisfaction and involvement. *Transport Policy 18* (2), 318–325. https://doi.org/10.1016/j.tranpol.2010.09.003

Li, Y.-w., 2003. Evaluating the urban commute experience: A time perception approach. *Journal of Public Transportation 6* (4), 41–67. https://doi.org/10.5038/2375-0901.6.4.3

Lin, T., Shalaby, A., Miller, E. J., 2016. Transit user behaviour in response to service disruption: State of knowledge. In: *Canadian transportation research forum 51st annual conference-North American transport challenges in an era of change//Les défis des transports en Amérique du Nord à une aire de changement*, pp. 1–8. https://doi.org/10.22004/ag.econ.319263

Lunke, E. B., 2020. Commuters' satisfaction with public transport. *Journal of Transport & Health 16*, 100842. https://doi.org/10.1016/j.jth.2020.100842

Mahapatra, S. C., Bellamkonda, R. S., 2023. Higher expectations of passengers do really sense: development and validation a multiple scale-FliQual for air transport service quality. *Journal of Retailing and Consumer Services 70*, 103162. https://doi.org/10.1016/j.jretconser.2022.103162

Mahudin, N. D. M., Cox, T., Griffiths, A., 2012. Measuring rail passenger crowding: Scale development and psychometric properties. *Transportation Research Part F: Traffic Psychology and Behaviour 15* (1), 38–51. https://doi.org/10.1016/j.trf.2011.11.006

Mapunda, M. A., 2021. Customers' Satisfaction on bus rapid transit services in tanzania: The servqual model perspective. In: *Sustainable education and development 9*. Springer, pp. 194–208. https://doi.org/10.1007/978-3-030-68836-3_18

Monsuur, F., Enoch, M., Quddus, M., Meek, S., 2021. Modelling the impact of rail delays on passenger satisfaction. *Transportation Research Part A: Policy and Practice 152*, 19–35. https://doi.org/10.1016/j.tra.2021.08.002

Mouwen, A., 2015. Drivers of customer satisfaction with public transport services. *Transportation Research Part A: Policy and Practice 78*, 1–20. https://doi.org/10.1016/j.tra.2015.05.005

Nichols, A., Ryan, J., Palmqvist, C.-W., 2024. The importance of recurring public transport delays for accessibility and mode choice. *Journal of Transport Geography 115*, 103796. https://doi.org/10.1016/j.jtrangeo.2024.103796

Oliveira, A. V. M., Oliveira, B. F., Vassallo, M. D., 2023. Airport service quality perception and flight delays: examining the influence of psychosituational latent traits of respondents in passenger satisfaction surveys. *Research in Transportation Economics 102*, 101371. https://doi.org/10.1016/j.retrec.2023.101371

Oppenheimer, D. M., Meyvis, T., Davidenko, N., 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45* (4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Osborne, J. W., 2014. Best practices in exploratory factor analysis. scotts valley. *Createspace independent publishing.* , 978–1500594343.

Papangelis, K., Corsar, D., Sripada, S., Beecroft, M., Nelson, J. D., Edwards, P., Velaga, N., Anable, J., 2013. Examining the effects of disruption on travel behaviour in rural areas. In: *Proceedings of the 13th world conference in transport research*, pp. 1–14.

Parasuraman, A., Zeithaml, V. A., Berry, L. L., 1988. Servqual: A multiple-item scale for measuring consumer perc. *Journal of Retailing 64* (1), 12.

Pett, M. A., Lackey, N. R., Sullivan, J. J., 2003. Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Sage.

Prasad, M. D., Shekhar, B. R., 2010. Development of railqual: A service quality scale for measuring indian railway passenger services. *Management Science and Engineering 4* (3), 87.

Pruyn, A., Smidts, A., 1998. Effects of waiting on the satisfaction with the service: Beyond objective time measures. *International Journal of Research in Marketing 15* (4), 321–334. https://doi.org/10.1016/S0167-8116(98)00008-1

Psarros, I., Kepaptsoglou, K., Karlaftis, M. G., 2011. An empirical investigation of passenger wait time perceptions using hazard-based duration models. *Journal of Public Transportation 14* (3), 109–122. https://doi.org/10.5038/2375-0901.14.3.6

Randheer, K., Al-Motawa, A. A., Vijay, P. J., 2011. Measuring commuters' perception on service quality using SERVQUAL in public transportation. *International Journal of Marketing Studies 3* (1), 21. https://doi.org/10.5539/ijms.v3n1p21
Reise, S. P., Waller, N. G., Comrey, A. L., 2000. Factor analysis and scale revision. *Psychological Assessment 12* (3), 287. https://doi.org/10.1037/1040-3590.12.3.287
Rezapour, M., Ferraro, F. R., 2021. Rail transport delay and its effects on the perceived importance of a real-time information. *Frontiers in Psychology 12*. https://doi.org/10.3389/fpsyg.2021.619308
Romero, C., Zamorano, C., Monzón, A., 2023. Exploring the role of public transport information sources on perceived service quality in suburban rail. *Travel Behaviour and Society 33*, 100642. https://doi.org/10.1016/j.tbs.2023.100642
Rüger, H., Pfaff, S., Weishaar, H., Wiernik, B. M., 2017. Does perceived stress mediate the relationship between commuting and health-related quality of life? *Transportation Research Part F: Traffic Psychology and Behaviour 50*, 100–108. https://doi.org/10.1016/j.trf.2017.07.005
Sam, E. F., Hamidu, O., Daniels, S., 2018. Servqual analysis of public bus transport services in kumasi metropolis, Ghana: Core user perspectives. *Case Studies on Transport Policy 6* (1), 25–31. https://doi.org/10.1016/j.cstp.2017.12.004
Sánchez Pérez, M., Carlos Gázquez Abad, J., María Marín Carrillo, G., Sánchez Fernández, R., 2007. Effects of service quality dimensions on behavioural purchase intentions: A study in public-sector transport. *Managing Service Quality: An International Journal 17* (2), 134–151. https://doi.org/10.1108/09604520710735164
Sarker, R. I., Kaplan, S., Mailer, M., Timmermans, H. J. P., 2019. Applying affective event theory to explain transit users' reactions to service disruptions. *Transportation Research Part A: Policy and Practice 130*, 593–605. https://doi.org/10.1016/j.tra.2019.09.059
Singleton, P. A., Clifton, K. J., 2021. Towards measures of affective and eudaimonic subjective well-being in the travel domain. *Transportation 48* (1), 303–336. https://doi.org/10.1007/s11116-019-10055-1
Spector, P. E., 1986. Perceived control by employees: A meta-analysis of studies concerning autonomy and participation at work. *Human Relations 39* (11), 1005–1016. https://doi.org/10.1177/001872678603901104
Sposato, R. G., Röderer, K., Cervinka, R., 2012. The influence of control and related variables on commuting stress. *Transportation Research Part F: Traffic Psychology and Behaviour 15* (5), 581–587. https://doi.org/10.1016/j.trf.2012.05.003
Stevens, J., et al., 2002. Applied multivariate statistics for the social sciences. Vol. 4. Lawrence Erlbaum Associates Mahwah, NJ.
Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science 12* (2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
Sweller, J., 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review 22*, 123–138. https://doi.org/10.1007/s10648-010-9128-5
Taylor, S., 1994. Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing 58* (2), 56–69. https://doi.org/10.1177/002224299405800205
Van Hagen, M., Galetzka, M., Pruyn, A. T., 2014. Waiting experience in railway environments. *Journal of Motivation, Emotion and Personality 2*, 41–55.
Van Lierop, D., Badami, M. G., El-Geneidy, A. M., 2018. What influences satisfaction and loyalty in public transport? a review of the literature. *Transport Reviews 38* (1), 52–72. https://doi.org/10.1080/01441647.2017.1298683
Vredenborg, M., van Kasteren, A., Masthoff, J., 2025. Personalization in public transport passenger information systems: A systematic review and framework. *ACM Computing Surveys*. https://doi.org/10.1145/3721478
Wen, C.-H., Lan, L. W., Cheng, H.-L., 2005. Structural equation modeling to determine passenger loyalty toward intercity bus services. *Transportation Research Record 1927* (1), 249–255. https://doi.org/10.1177/036119810519270012
Wener, R. E., Evans, G. W., Phillips, D., Nadler, N., 2003. Running for the 7: 45: The effects of public transit improvements on commuter stress. *Transportation 30*, 203–220. https://doi.org/10.1023/A:1022516221808
Wittmann, M., 2016. Felt time: The psychology of how we perceive time. MIT Press, Cambridge, MA. USA.
Worthington, R. L., Whittaker, T. A., 2006. Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist 34* (6), 806–838. https://doi.org/10.1177/0011000006288127
Woźniak, P. W., Hak, M., Kotova, E., Niess, J., Bentvelzen, M., Weingärtner, H., Schött, S. Y., Karolus, J., 2023. Quantifying meaningful interaction: Developing the eudaimonic technology experience scale. In: *Proceedings of the 2023 ACM designing interactive systems conference*, pp. 1904–1914. https://doi.org/10.1145/3563657.3596063
Woźniak, P. W., Karolus, J., Lang, F., Eckerth, C., Schöning, J., Rogers, Y., Niess, J., 2021. Creepy technology: What is it and how do you measure it? In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–13. https://doi.org/10.1145/3411764.3445299
Yap, M., Cats, O., 2021. Predicting disruptions and their passenger delay impacts for public transport stops. *Transportation 48* (4), 1703–1731. https://doi.org/10.1007/s11116-020-10109-9
Zografos, K. G., Androutsopoulos, K. N., Spitadakis, V., 2009. Design and assessment of an online passenger information system for integrated multimodal trip planning. *IEEE Transactions on Intelligent Transportation Systems 10* (2), 311–323. https://doi.org/10.1109/TITS.2009.2020198
Zwick, W. R., Velicer, W. F., 1986. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin 99* (3), 432. https://doi.org/10.1037/0033-2909.99.3.432