# Psychometric measurement and decision-making of accessibility in public transport for older persons with functional limitations

Birgitta Berglund[1, 2]

Mats E. Nilsson[1]

Catherine Sundling[3]

Ragne Emardson[4,5]

Leslie Pendrill[5]

[1]Department of Psychology, Stockholm University, SE-106 91 Stockholm, Sweden

[2]Institute of Environmental Medicine, Karolinska Institutet, P.O. Box 210, SE-171 77 Stockholm, Sweden

[3]School of Social Sciences, Södertörn University, Alfred Nobels allé 7, 141 89 Huddinge, Sweden

[4] Present address: Faculty of Textiles, Engineering and Business, University of Borås, 501 90, Borås, Sweden

[5]RISE Metrology, RI:SE Research Institutes of Sweden, Eklandagatan 86, 41261 Gothenburg, Sweden

Corresponding author E-mail address, full postal address

Leslie Pendrill, RI:SE Metrology, Eklandagatan 86, 41261 Gothenburg, Sweden; Email: leslie.pendrill@ri.se

*Extract*: “Mätningar Möjliggör Framtidens Järnväg För Alla,Vetenskaplig Rapport, Metodik För Att Mäta Tillgänglighet”, 2014
https://www.trafikverket.se/contentassets/773857bcf506430a880a79f76195a080/forskningsresultat/slutrapport_vetenskaplig_rad_42.pdf

Key words: accessibility, older persons, psychometric measurement, decision making, Rasch analysis, functional limitations

## Abstract

Vulnerable travellers face challenges in the transport environment potentially leading to decreased mobility. A research goal is to find out how to improve railway accessibility by reducing barriers for persons with functional limitations. A method for measuring accessibility has earlier been developed where travel barriers are assigned different weights based on persons' functional ability and travel behaviour. The more weight placed on a certain barrier, the less accessible and thus less probable a particular journey will be. In the present work, these travel-barrier weights are analysed psychometrically based on the replies to questions about ease or difficulty of accessing travel with various barriers in responses given by 162 older long-distance travellers. An invariant measure theory approach (Rasch) is employed that allows (i) transformation of ordinal questionnaire data onto a quantitative interval scale; and (ii) separate measures of barrier level of challenge and person capability. A principal component analysis revealed three main clusters: 1) mainly ergonomically related questions; 2) mainly informational/cognitive; while cluster 3) is a mix of these two. Correlations are investigated between perceived accessibility and functional ability, and between person capability and functional ability. Independent sources of measurement uncertainty (e.g. under-estimation of scores) are distinguished from separate estimates of task challenge and individual travel capability. These independent sources are accounted for in estimates of reliability and validity of the various measures.

## 1 Introduction

The proportion of older persons is increasing in many countries. As functional limitations become more common with age and many older persons have acquired more than one functional limitation [e.g. Sundling, Berglund, Nilsson, Emardson, & Pendrill, 2014], transport systems must be systematically adapted to meet the special and increasing needs of this group.

An overall research goal is to find out how to improve railway accessibility by reducing barriers for persons with functional limitations. In a previous study [Sundling, Berglund, Nilsson, Emardson, & Pendrill, 2013], a method for measuring accessibility has been developed: travel barriers are assigned different weights based on persons' functional ability and travel behaviour. The more weight placed on a certain barrier, the less accessible a particular journey will be and thus less probable to take place.

In the present paper, these travel barrier weights are analysed psychometrically where the data comes from responses to a questionnaire sent out earlier to a random sample of 1 000 older persons [Sundling, Berglund, Nilsson, Emardson, & Pendrill, 2014]. The present analysis is based on the replies to questions about the ease or difficulty of accessing travel with various barriers (ergonomic, informational, etc.) in long-distance train journeys. For the present paper, an invariant measure theory approach [Rasch, 1961] is employed which allows (i) transformation of ordinal questionnaire data onto a quantitative interval scale; and (ii) separate measures of barrier level of *challenge* and of

person *capability*.

## 2 Method and responses

### 2.1 Sample and procedure

In a previous study, questionnaire data were collected to examine older persons' (65-85 years old) perceived accessibility in public transport [see Sundling, Berglund, Nilsson, Emardson & Pendrill, 2014]. Questions addressed, apart from demography, the participants' *kinds of functional limitations/diseases* and *level of functional ability* as well as *travel behaviour*. A major part of the questionnaire addressed potential *barriers* to travel. The sample consisted of 574 persons (response rate 57%, 54% women and 46% men, with an arithmetic mean age of 73 years) living in the county of Stockholm, Sweden. A majority did not have reduced functional ability (57%) apart from 5% who had "very reduced" or "extremely reduced" functional ability. Vision impairment was the most common functional limitation/disease (22%), followed by hearing impairment (21%) and cardiovascular disease (17%). The proportion reporting having no functional limitations/diseases at all was 28%.

For the present paper, we have selected a part of the extensive questionnaire for our analyses [see Appendix]. Here, the main focus is on travelling by long-distance train. This part concerns perceived *barriers* in long-distance train traveling in combination with self-rated *functional ability*. The number of participants answering this part of the questionnaire are 162. An inclusion criterion is to have traveled by long-distance train during the last year.

### 2.2 Measures

In the present work we use (a) a five-category severity scale of self-rated functional ability ("Not reduced", …. "Extremely reduced") and (b) ten questions on potential barriers in long-distance train traveling. Nine of the barrier questions concern specific situations along the travel chain during the latest long-distance train journey and one question addresses perceived overall accessibility in train traveling. Scores for each question are on a five-category scale ("Totally disagree", …. "Totally agree" or "Very good", …. "Very bad").

A subset of questions, 38–46 and 51 on barrier challenges [see Appendix], were focussed on for the present invariant measure theory analysis [Section 4], where typical questions of informational and ergonomic character were:

- Q42. Was it easy or difficult to get information on board the long-distance train(s)?
- Q44. Was it easy or difficult to alight from the (long-distance) train(s)?

The participants were asked to rate how easy or difficult it was to use the travel environment during different parts of the *latest long-distance train journey*, from the planning of the trip until reaching the destination**.** Nine of our questions addressed such specific parts in the travel chain while one question addressed overall accessibility in train traveling.

### 2.3 Responses

An account and initial analysis of the responses have already been published [Sundling, *et al.* 2014], so only a brief recall will suffice here. Most of the respondents found travelling easy. Taken together the mean percentage of respondents perceiving the travel environment as ”easy” or ”good” was 69% while the number for those who thought it was ”difficult” or ”bad” was 7%.

#### *2.3.1 Informational barriers*

As many as 66% found it easy to take part of information at the stations. The 13%, who did not find it easy, wanted more staff or visual information. There were also complaints about unsystematic information or broken signs. Orientation within stations areas was easy according to 64%. But 11% found it difficult because of inadequate information and because of ongoing reconstruction. Information on board the train was perceived as good by 62% compared to the 73% who regarded information as good during the planning stage. Loudspeakers were the main cause of discontent; not functioning, the sound being difficult to hear, or too loud. Finally, personal service from staff during the journey was judged to be good by 68% and only 4% found it unsatisfactory, mainly because of absence of staff.

#### *2.3.2 Ergonomic barriers*

A majority of the respondents, 69%, found it easy to move within the station area(s): however, 10% found it relatively difficult or very difficult. Most respondents (74%) also found it easy to get on the train; difficulties mainly concerned climbing the high steps into the train with luggage. To move around inside the vehicle was perceived as easy by 72%. Complaints concerned mainly insufficient stability in the train; vibrations, lurching; etc. The lavatory was easy to use and/or to get to according to 65%, but some respondents had experienced lavatories that were closed, dirty, out-of-service or lumbered of luggage. Getting off the train was easy according to 79%. Again, luggage and also high steps were barriers mentioned as well as unhelpful staff.

## 3 Specification of aims, system & construct, and quality characteristics & metrics

As with measurements of all kinds that are made not for their own sake but as a basis for essential decisions, for instance, about whether or not a product conforms to specifications, the validity and reliability of the measurements have to be ensured from the start [e.g. Roach 2006; Pendrill, 2014]. This conformity assessment involves: a correct and proactive *definition of the aim* [see Section 3.1]; a *system definition* (user, product, task, environment) subject to assessment [Section 3.2]; the identification of key *quality characteristics* of the system and the *setting of specification limits* for each such characteristic [Section 3.3]. Thereafter, a *measurement system* fit for the task at hand should be developed and subjected to all the steps of *conformity assessment*. Finally, *decisions of conformity* can then be made of whether the aims are satisfied or not [Section 4.3]. Examples of earlier research in this area of *accessibility* include that of a Swedish research team who developed a supportive instrument, “The Travel Chain Enabler”, for assessing urban public bus transportation accessibility for persons with functional limitations [Iwarsson, Jensen & Ståhl, 2000; Jensen, Iwarsson & Ståhl, 2002]. See also the AIMFREE by Rimmer, Riley, Wang, and Rauworth [2004] for measuring accessibility to specific environments.

### 3.1 Aims

The overall aim is to provide the 'market', i.e. society, with improved 'services' of more accessible rail travel, by mustering available 'capital' – in terms of rail system products and infrastructure combined with persons' functional abilities [see Figure 1].

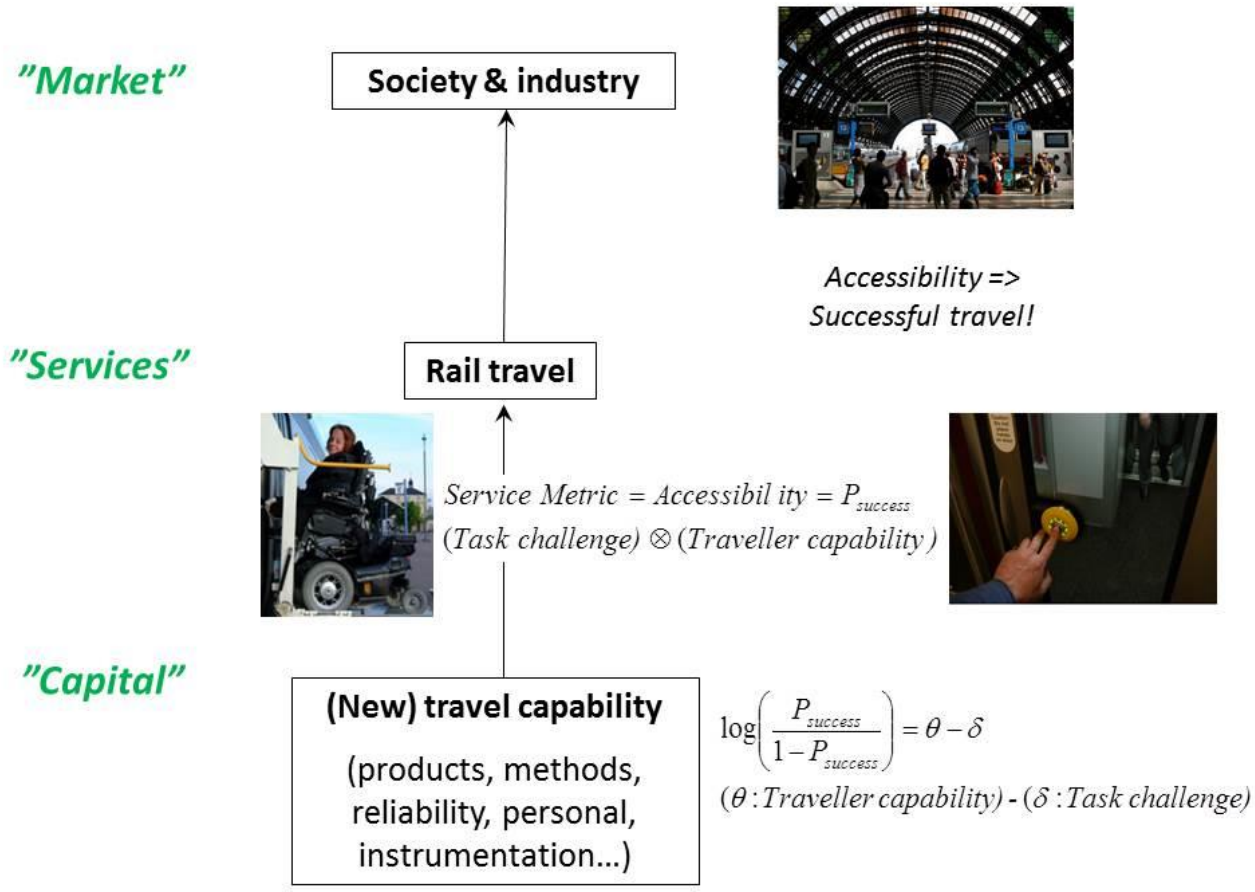


Figure 1 Accessibility = successful service of providing good rail travel.

### 3.2 System and Construct

The main construct of interest in the present study is the overall accessibility of the rail system as perceived by older travellers with functional limitations. That accessibility can be modelled as being determined by an aggregate of the pairwise, mutual interactions between the three principal elements of the studied system [Figure 2], namely: (i) the traveller; (ii) the rail system; and (iii) the person-environment tasks necessary for travel. These interactions are functions of the intrinsic attributes of each of these system elements (e.g. person capability, task challenge, product quality, etc), but can also be influenced by the overall environment.

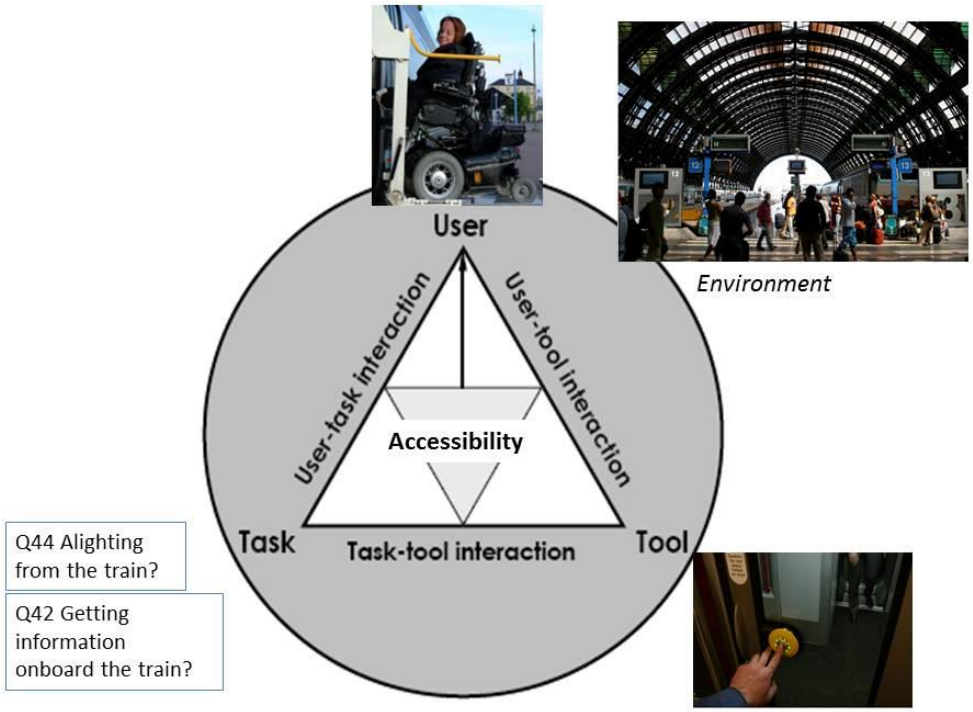


Figure 2 System and construct.

### 3.3 Quality characteristics and metrics

In the present study, amongst the various components of the system studied [Figure 2], two particular characteristics – travel-task challenge and traveller ('user') capability[1] - are systematically varied, while the other system aspects (certain parts of the environment) are kept as constant as possible. Accessibility is determined as a whole by the sums of interactions between the user, tools and tasks on all three sides of the system triangle shown in Figure 2. "Staircase" and "ticket counter" are examples of tools while "alighting from the train" and "getting information onboard" are examples of tasks.

Capability is here treated as an indirect measure of each person's item responses on the barrier challenges (10 items).

A mathematical model presented earlier [Emardson *et al.* 2012; Sundling, Berglund, Nilsson, Emardson, & Pendrill, 2013] for the aggregate accessibility ($A_{ij}$) for a person ($i$) making a complete journey ($j$) – from the initial planning, through travel, until arriving at the final destination (the whole trip) – is:

$$A_{ij}^{m} = \prod_{b} \left[ 1 - p_{jb} \psi_i(d_b) \right] \qquad (1)$$

as a product over the series of barriers, $b$, that have to be overcome during the journey of the individual person whose perceived effort is $\psi_i$ when facing a certain barrier at a distance $d_b$, together with the probability, $p_{jb}$, that a person, $i$, will face that particular barrier [Church & Marston 2003].

[1] Most of the questions posed in our survey do not address functional ability directly.

The *measured* perceived effort, $\psi_i$, for a person when encountering a barrier, $b$ ($p_b$ = 1), will be a function of the true value, $\phi$, and an error component $\varepsilon$:

$$\psi_i(d_b) = f\left[\phi_i(d_b), \varepsilon_{ib}\right] \quad (2)$$

The *true* value, $\phi$, of the perceived effort $\psi_i$ is a function for each person $i$ and barrier distance, $d_b$. In principle, the function might be different for different barriers and different persons. The true value is of course impossible to know, but its existence can still be assumed. A linear measurement model can be adopted so that the relationship between the measured perceived effort for each barrier and the true value can be written as:

$$\psi_i(d_b) = \phi_i(d_b) + \varepsilon_{ib} \quad (3)$$

The error component, $\varepsilon$, is a random variable which we assume is Normally distributed with a mean value, $\bar{\varepsilon}$, and variance, $\sigma_\varepsilon^2$.

The perceived effort function, $\phi$, is defined between 0 and 100%, and a key observation of the present work is that it can be converted into an accessibility score:

$$P_{success} = 100\% - \phi \quad (4)$$

of how successful deployment of the system is for each task at hand (i.e., the service of providing good rail travel [Figure 1]). Thus the function value $\phi$ =100% indicates a barrier such that the probability of cancelling the journey is 100% when facing such a barrier. A value of '0' indicates the opposite, i.e., the barrier causes no problem to the traveller and accessibility is complete, i.e.

$P_{success}$ = 100%.

In the present research, we will model the accessibility score for each individual traveller when negotiating a barrier as a product, in some way, of task challenge $\delta$ and traveller capability $\theta$ [Figure 1]:

$$Service\ Metric = Accessibility = P_{success} = (Task\ challenge) \otimes (Traveller\ capability) \quad (5)$$

A certain level of accessibility can be got for an easy task encountered by a less able traveller or for a difficult task scaled by a more able traveller. An understanding of how *perceived effort* depends on the intrinsic challenge of each barrier will enable future predictions of travel accessibility. The barrier should preferably be independent of abilities or attitudes of particular individuals and would enable accessibility for other people as well. This would facilitate the identification of which barriers limit accessibility the most and which barriers should be improved. As will be discussed below, this also opens up the possibility of establishing *metrological standards for accessibility* [Section 4.2.1].

Similarly, understanding how perceived effort when travelling depends on the intrinsic ability of each individual, will give opportunities to coach people how to negotiate barriers better. The perceived effort should preferably be independent of levels of challenge posed by specific barriers. This would help identifying what aspects of barrier challenge relate to human perception, so these challenges can be reduced.

A typical requirement (common to Generalised linear models and Logistic regression) is that the accessibility score, $P_{success}$, should preferably exceed 50% as a target specification limit.

## 4 Separate estimates of barrier challenge and traveller capability

To benefit fully from the information contained in perceived accessibility observations [Questions 38–46 & 51 on barrier challenges], it is essential to make separate estimates of *person* and *item* attributes. In measuring and analysing accessibility, according to the discussion in Section 3.3, it is conceivable that one could model the task-traveller interaction as some kind of arithmetic product, as indicated schematically in Eq. 5. But, to deal with the intrinsic nature of the perceived accessibility, we chose instead a classical-logistic regression approach [Theil, 1970] to the ordinal data typical of questionnaire responses, where the log-odds of success are modelled as linearly varying with a difference in a traveller attribute ("capability") and a task attribute ("level of challenge"):

$$\log\left(\frac{P_{success}}{1-P_{success}}\right) = \theta - \delta = (\theta : Traveller\ \ capability) - (\delta : Task\ challenge) \qquad (6)$$

One could also have included additional attributes for other system elements important for accessibility, e.g. train system "quality", how easily a product can facilitate accessibility or provide 'satisfaction', for instance.

The attribute values transformed from the ordinal raw scores, via Eq. 6, lie on a quantitative interval scale on which all of the usual statistical and metrological tools can be applied (in contrast to ordinal scales [Svensson, 2001]. The approach allows one to reveal explicitly typical ordinal scoring errors common in responses for the raw data not linear across the range [Massof, 2005, Figure 5]. Systematic investigation of Eq. 6 can be made for a fixed person attribute, $\theta$, for a range of item attributes, $\delta$, or vice versa, and similarly for the other system components, as required.

A significant advantage of this Rasch invariant-measure theory approach is that it provides:

- measures of the difficulties of *each* survey or test item not affected by the capabilities (or attitudes) of particular persons measured
- measures of the capabilities (or attitudes) of *each* person not affected by the difficulties of particular survey or test items [Fisher, 1997],

An earlier example of the Rasch approach in studies of persons with functional limitations is the work of Rimmer *et al*. [2004], who developed the AIMFREE set of psychometric measuring

instruments for the purpose, with 16 validated survey instruments measuring accessibility to recreational and fitness environments.

For our accessibility analyses the following nomenclature applies:

$$i = \begin{cases} 1, test\ \ person\ A \\ 2, test\ \ person\ B \\ 3, test\ \ person\ C \\ \quad ... \end{cases} \quad j = \begin{cases} 1, barrier\ a \\ 2, barrier\ b \\ 3, barrier\ c \\ \quad ... \end{cases} \quad k = \begin{cases} 1, question\ Q1 \\ 2, question\ Q2 \\ 3, question\ Q3 \\ \quad ... \end{cases}$$

In the case of classification into more than two ($k = 1, \ldots, K$) categories, the polytomous [Andrich 1978; Masters 1982) Rasch probability of response $v_{i,j}$ of person $i$ to item $j$ is given by:

$$v_{i,j,c} = \frac{e^{\left[c\cdot(\theta_i - \delta_j) - \sum_{k=1}^{c} \tau_{k,j}\right]}}{\sum_{c=0}^{K_j} e^{\left[c\cdot(\theta_i - \delta_j) - \sum_{k=1}^{c} \tau_{k,j}\right]}} \qquad (7)$$

where $\tau_k$ denotes the threshold for the $k^{\text{th}}$ category.

### 4.1 Results of Rasch analysis

The two diagrams of Figures 3 [i and ii] show two examples of Rasch curves representing the Rasch formula [Eq 7] fitted to the responses given by 162 long-distance travellers (out of the total sample of 574 respondents) to two questions, Q42 and Q44, about barrier challenges, which are found to lie at opposite ends of the classification scale for accessibility [see Figure 4].

Fitting of the Rasch formula [Eq. 7] in the 'log-odds' version [for the dichotomous case: $\log\left(\frac{P_{success}}{1-P_{success}}\right) = \theta - \delta$ ] to the experimental score is done with a 'logistic regression' [Wright, 1996], where least-squares fit residuals, the differences, $\Delta v_{i,j} = v_{i,j} - v'_{i,j}$, are minimised[2], point for point, between the measured raw scores of response, $v_{i,j} = Score(1,...,6) = 100\% - \psi$, and modelled responses, $v'_{i,j} = P_{success} = 100\% - \phi$, calculated by adjusting the latent variables θ and $\delta$ to best (least squares) fit the data for each of the 10 questions (Q38 to Q46 and Q51) on barrier challenges.

[2] WINSTEPS, http://www.winsteps.com/index.htm

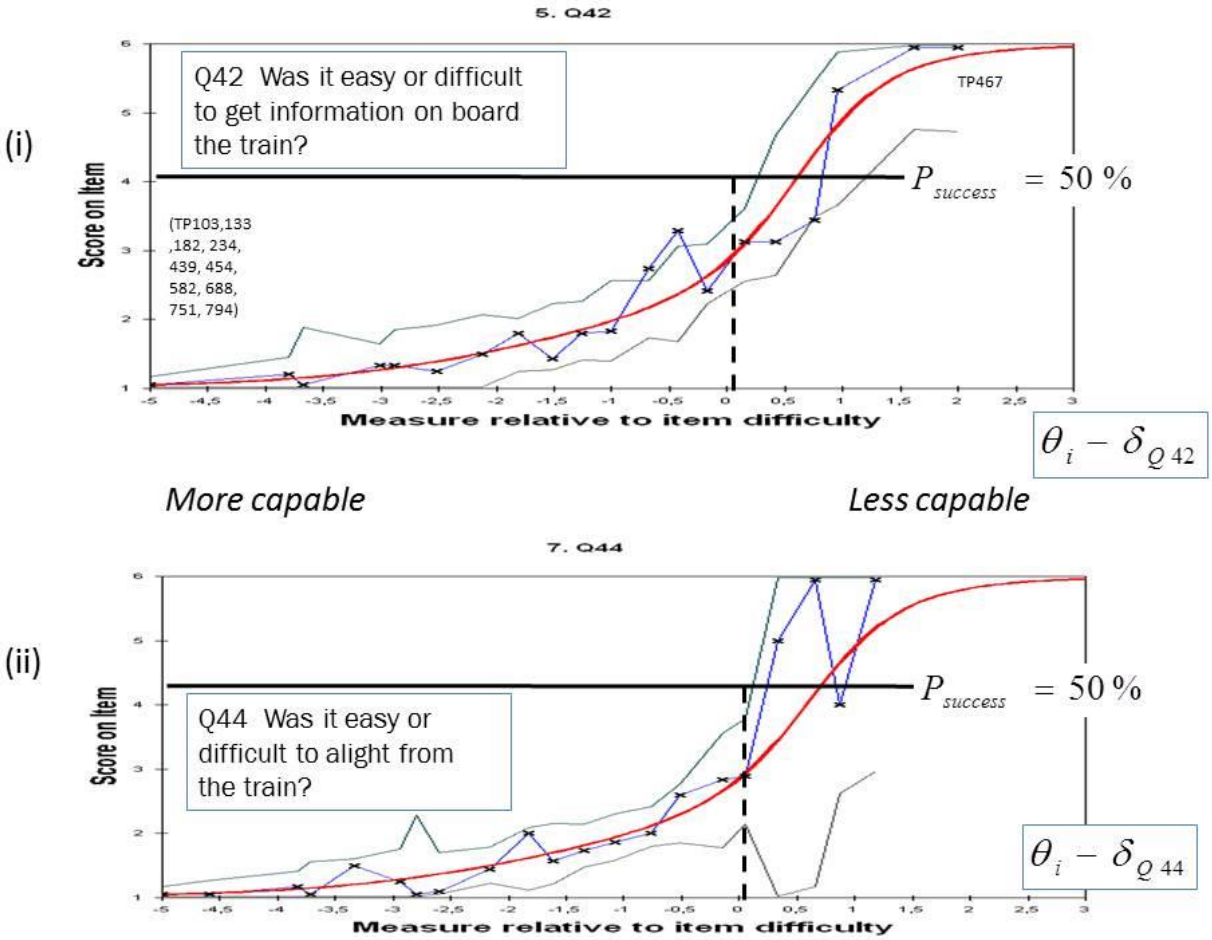


Figure 3 Rasch curves from logistic regression of Eq. 7 to data: Examples of responses of elderly (162 test persons) to two questions: (i) Q42 and (ii) Q44 over the range of person capability, $\theta_i$. Smooth curve: $v'_{i,j} = P_{success} = 100\% - \phi$; Noisy curve with crosses: $v_{i,j} = Score(1,...,6) = 100\% - \psi$, together with 95% confidence interval about the mean at each point . For Q42 for instance, Test Person TP467 scored as the least capable of the studied cohort for that task.

$P_{success}$ = 50% at the point where the person attribute (capability), $\theta_i$, equals the item attribute (difficulty), $\delta_j$.

For each fit, the Rasch analysis gives separate estimates of the 162 individual person attributes (person capability), $\theta_i$, and each of the 10 item attributes (the barrier challenges), $\delta_j$, as shown in the histograms of Figure 4.

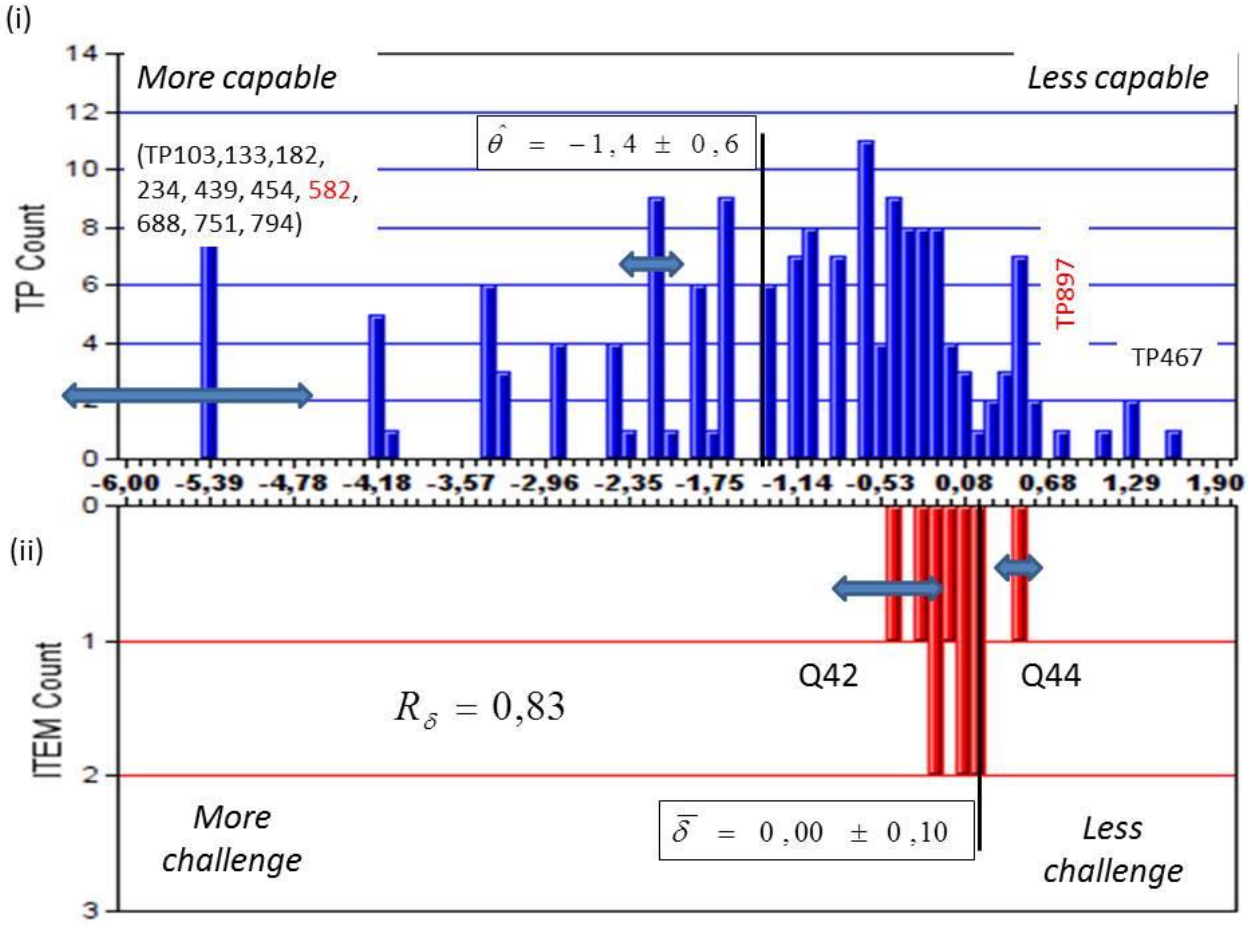


Figure 4 Rasch histograms for: (i) Test Persons' (TP) capability (for 162 persons) and (ii) barrier challenge (for 10 questions Q42 to Q48 and Q51, see Appendix). For instance, Test Person TP467 scored as the least capable of the studied cohort averaged over all tasks.

Apparent from the plots in Figures 3 and 4 are the location and dispersion of the person attributes and item attributes.

For instance, the individual person attribute values [Figure 4 part (i) upper histogram] indicate a considerable spread, skewed away from the corresponding distribution of the measured item attributes [Figure 4 part (ii) lower histogram] and heavily skewed towards the more capable test persons (TP), reflecting the spread of functional limitations of the persons studied.

For the items, broadly speaking, questions concerning ergonomic barriers [such as Q44; "to get off the train", Section 2.3.2] appear on the average to indicate less challenge than questions concerning informational (or cognitive) barriers [such as Q42; "to retrieve information on-board", Section 2.3.1].

The Rasch approach can also enable guidance to improving future studies. The considerable mismatch between person attributes and item attributes evident in Figure 4 could be reduced – thereby leading to better reliability – by posing additional questions of a more challenging nature, particularly for the more capable test persons. The present mismatch limits the reliability of the current estimates of person capability and barrier challenge [see further discussion in Section 4.3].

## 4.2 Data quality

In this Section, we clarify a set of concepts belonging to measurement data quality metrological traceability, incorrect scoring, as well as validity, construct alleys, and principal component analysis. No decisions [Section 4.3] about the significance of any apparent differences in different measures can be made without demonstrating sufficient measurement data quality.

### *4.2.1 Metrological traceability*

In general, the measured level of *barrier challenge* $\delta$ differs, because of limited reliability, from the 'true' value of $\delta'$, with an error $\varepsilon_\delta$:

$$\delta = \delta' + \varepsilon_\delta$$

"Invariant" measure theory, allowing the $\delta$ for a particular task (as created by a barrier to travel) to be estimated independently of who is encountering the particular challenge. This theory permits the identification of a metrological standard for barrier challenge. Once an agreed definition and realisation of the standard barrier has been achieved, the barrier can be used reproducibly as a *reference* in other travel situations. As in traditional metrology, this *traceability* provides all the advantages commensurate with objectively comparable measurement [Pendrill 2018].

For instance, having access to a *psychometric barrier-challenge standard* would allow an estimate of each person's capability $\theta$ to access travel for a range of barriers of different challenge $\delta$, to be metrologically calibrated by measuring a task of known challenge. This procedure determines the measurement error $\varepsilon_\theta$ in person capability: $\theta = \theta' + \varepsilon_\theta$

Inserting the corrected values for the item $\delta$ and person $\theta$ attributes in the Rasch expression [Eq. 6 or the polytomous variant Eq. 7] allows a more correct estimate of the accessibility score, $P_{success}$. It will in turn allow correction of the perceived effort function $\psi_i(d_b) - \mu_{ib}$ [as given by Eq. 3]. Finally, the *overall accessibility measure* ($A^m_{ij}$) for an individual's ($i$) complete journey ($j$), correctly calibrated would be obtained as:

$$A^m_{ij} = \prod_b \left[1 - p_{jb}\left(\psi_i(d_b) - \mu_{ib}\right)\right] \qquad (8)$$

This formula can be used to predict the expected accessibility of future journeys ($j$) characterised by the probability $p_{jb}$ of encountering each barrier ($b$). It is useful, for instance, if one wants to measure the impact of innovation – e.g. modernisation of ergonomic or informational aspects of various barriers – in terms of improvements in accessibility.

### *4.2.2 Incorrect scoring*

The Rasch analysis can reveal explicitly typical ordinal scoring errors common in responses for the raw data, which are not linear across the range [Massof, 2005]. Such incorrect scoring is often observed towards the extreme ends of the Likert scale [Likert, 1932; Massof, 2005], whereas scores mid-range are scaled approximately linearly.

As shown in Figure 5, a plot of the Rasch $\delta_j$ against the corresponding simple average, $\bar{v}_j = \frac{1}{N_{TP}} \cdot \sum_{i=1}^{N_{TP}} v_{i,j}$, of the raw barrier score data $v_{i,j} = Score(1,...,6)$ over all test persons for each of the 10 questions Q42 to Q48 and Q51 shows little evidence in the present case that the travellers tend to score incorrectly and there is a clear linear relation with high ($R^2$ close to 1) goodness of fit,

as expected for these mid-range scores [see also the reliability discussion in Section 4.3).

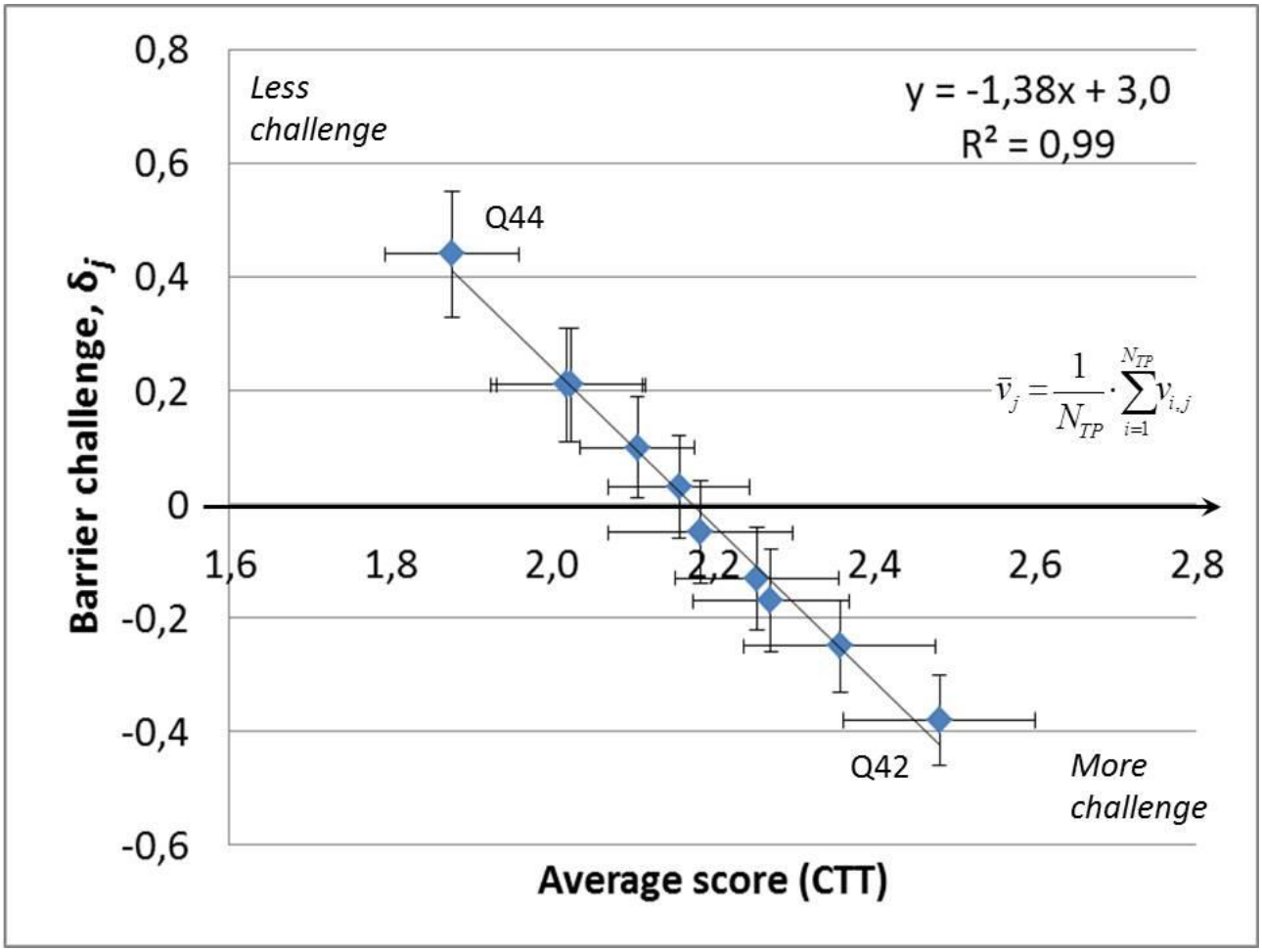


Figure 5 Test of Incorrect scoring of barrier challenge (10 items). Plot of invariant measure estimate of challenge [Rasch $\delta_j$ from Eq. 7) versus average challenge $\bar{v}_j$ with Classical Test-Theory (CTT) for each barrier *j*.

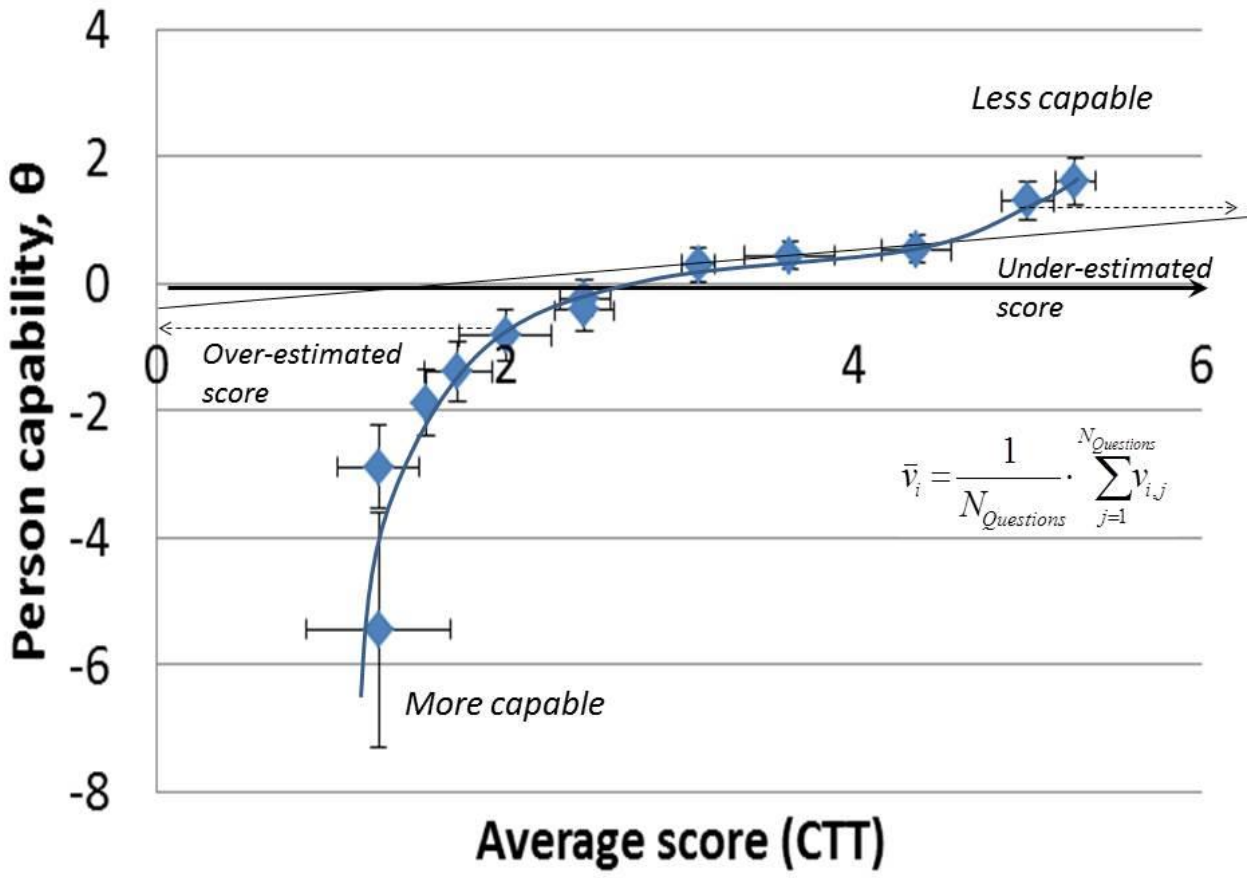


Figure 6 Test of incorrect scoring of person capability (162 travellers). Plot of invariant measure estimate of capability [Rasch $\theta_i$ from Eq. 7) versus average capability $\bar{v}_j$ with Classical Test Theory (CTT) for each person *I*.

In contrast, averaged over all barriers, individual person scores at the high capability (low score) as well as the low capability (high score) extremes of the Likert scale are strongly over-estimated respectively under-estimated [Figure 6]. Uncertainties are also relatively large for the Test Persons (TP) with high capability, owing to 'ceiling' effects as one approaches the end of the scale [Massof, 2005].

*4.2.3 Construct alleys*

A graphic measure of the goodness of fit in the logistic regression of measurements to the Rasch equations (6) and (7) is provided by so-called "construct alley" plots of Rasch $\delta_j$ (barrier challenge) against the Infit *z*-score [Massof, 2005].

The $X$ = chi$^2$ distribution of the fit residual $\chi = \Delta v_{i,j} = v_{i,j} - v'_{i,j}$ [Section 4.1] is known to tend slowly to a Normal distribution for large number of degrees of freedom $n$ [Fisher, 1924], and one has, therefore, over the years sought other distributions that converge faster. One proposed fit measure was derived by Wilson and Hilferty [1931):

$$INFIT - zstd = \frac{2 \cdot \left( \sqrt[3]{X} - \sqrt[3]{n - \frac{2}{3}} \right)}{\sigma}, \qquad (9)$$

in which $\sigma = \frac{\sqrt{2}}{3 \cdot \sqrt[6]{n - \frac{2}{3}}}$, and Eq. 9 has become a more or less standard choice in contemporary programs such as WINSTEPS ®, following the treatises of Wright and Stone [1979] and of Linacre [2002]. In such analyses, traditionally so-called 'construct alley' plots [Massof, 2005; WINSTEPS] are made [Figure 7] of item attribute ($\delta_j$ from Eq. 7, barrier challenge) versus Infit *z-standardized*-score [Eq. 5.4) in order to show how well each datum fits within the 'construct alley' of two standard deviations about zero. Evident from Figure 7 is that only a few of the item attribute values lie outside the statistically expected (95% confidence) alley interval, indicating a satisfactory fit.

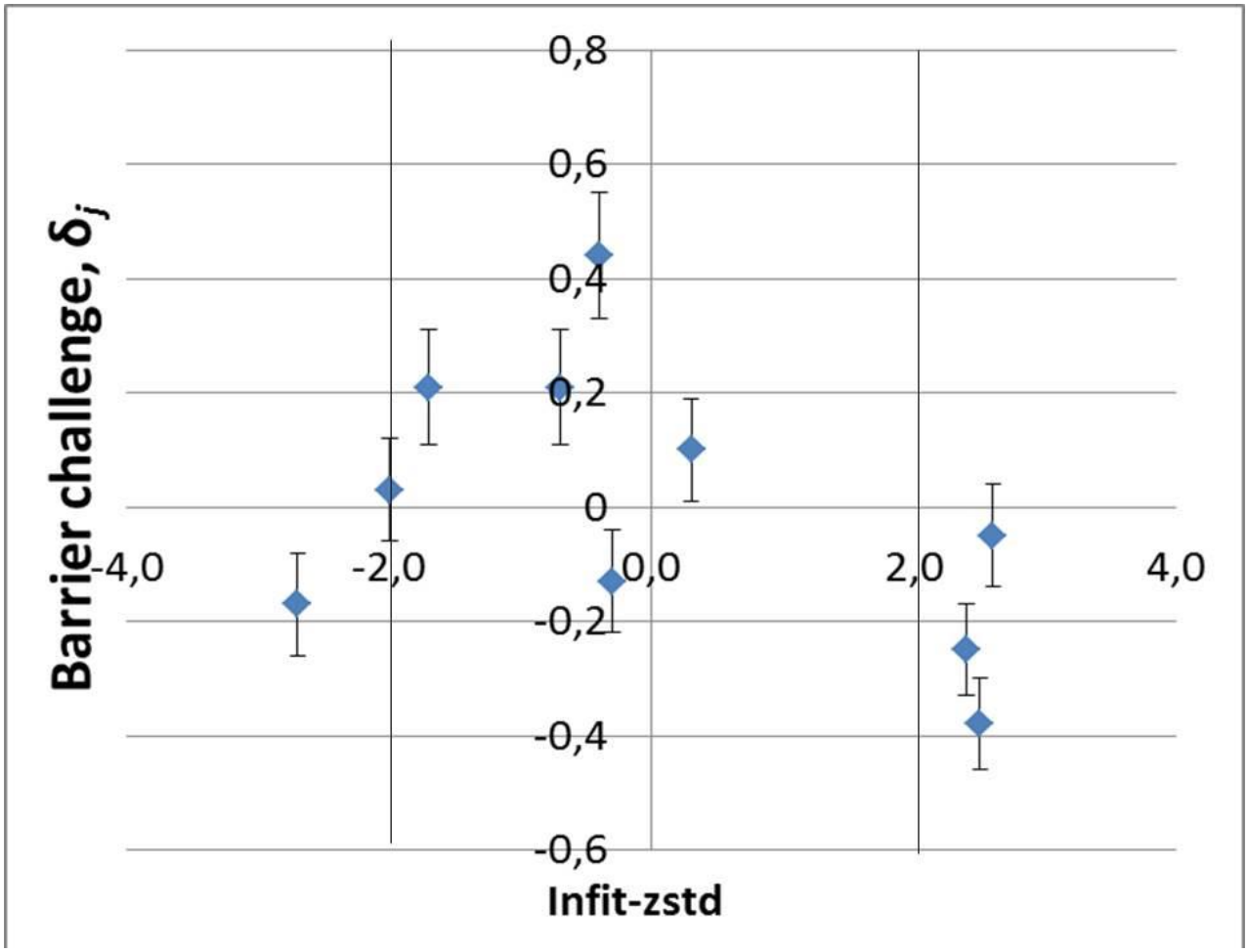


Figure 7. Construct alley plot of *item attribute* ($\delta_j$ from Eq. 7, barrier challenge) versus *Infit zstd-score* [Eq. 9).

*4.2.4 Validity*

To obtain valid capability-attribute estimates the intended quantity must actually be measured. Figure 8 shows a reasonable degree of correlation (goodness of linear fit $R^2$ close to 1) between the invariant measure estimates of capability (Rasch $\theta_i$ from Eq. 7, averaged by Q13 score) against the self-rated "functional ability" scores (Q13). The displayed standard measurement uncertainties include fit uncertainties as well as scatter of invariant measure estimates of capability within each Q13-score group.

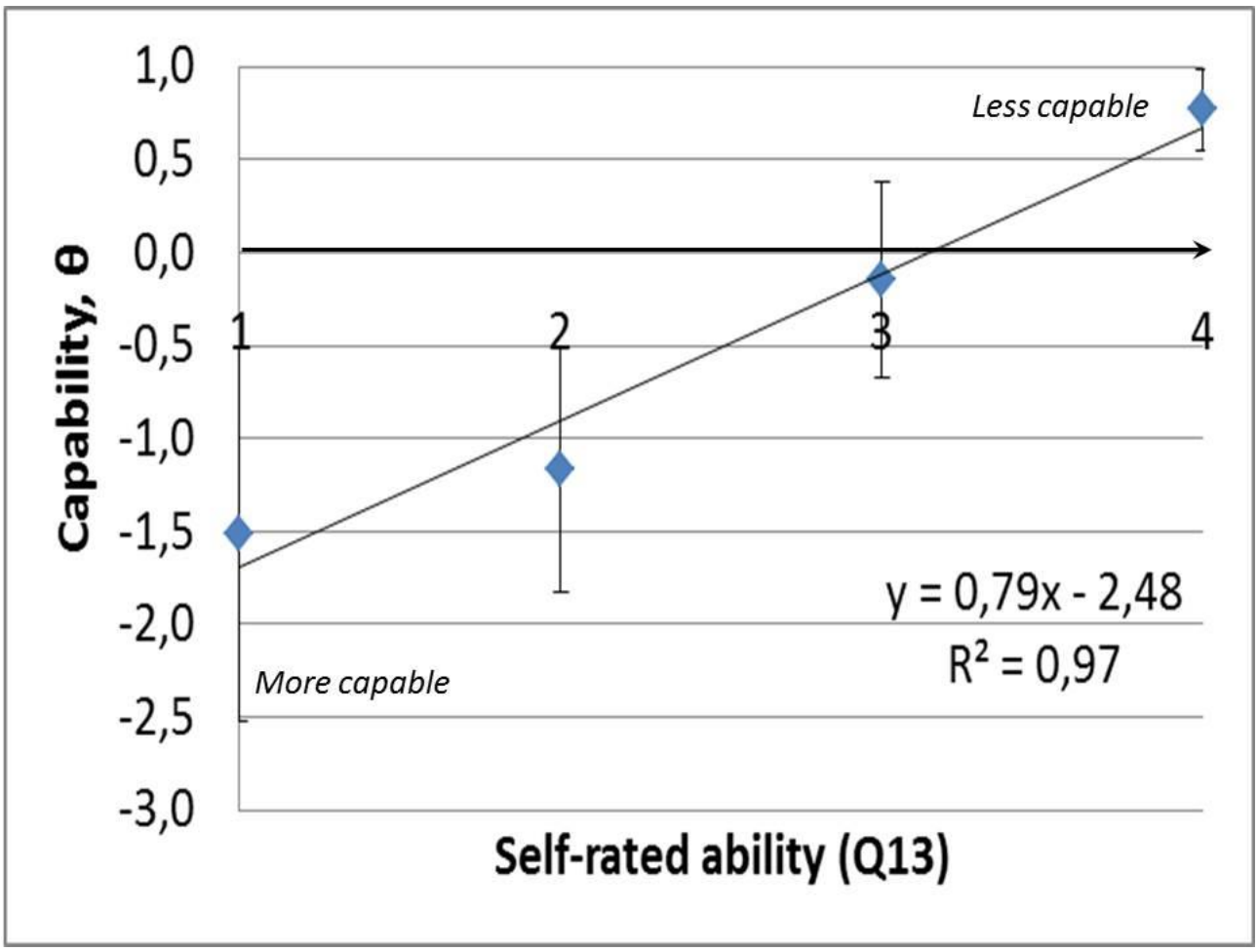


Figure 8 Test of validity of person capability estimates (162 travellers). Plot of invariant measure estimates of capability (Rasch $\theta_i$ from Eq. 7, averaged by Q13 score) versus self-rated ability (Q13). (1: *No functional limitation*,…, 4: *Very limited functionally*; Q13).

*4.2.5 Principal component analysis (PCA) in Rasch analysis*

A first-order assumption of a Rasch analysis is that a single item parameter, $\delta$, is the principal component for all the items (and similarly for the capability attributes, $\theta$, of the different persons). By default all items (or all persons) are in the "first factor" (or first PCA component in a correlation matrix of residuals) until proven otherwise.

A multivariate statistical method which tests this first-order assumption is the Rasch "PCA of residuals" which seeks *"unexpected" data patterns* that are not in accord with the Rasch measures [Wright, 1996].

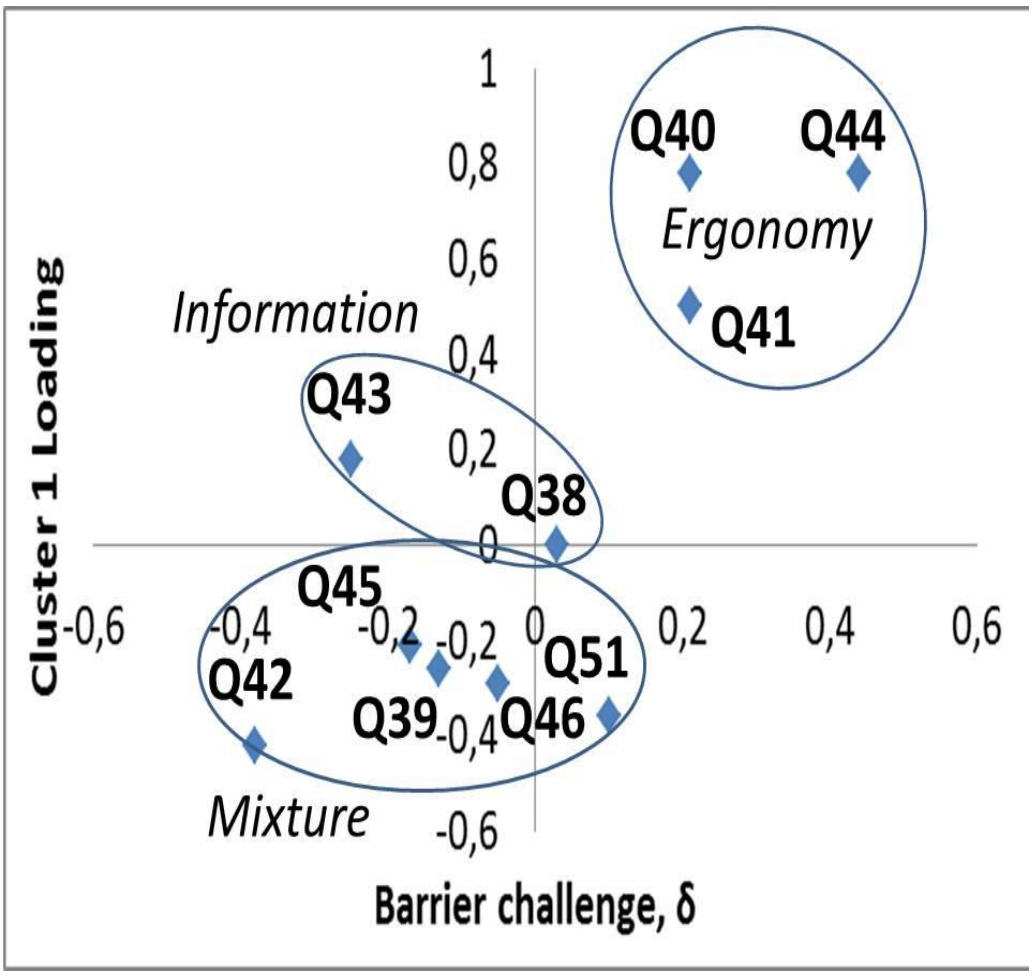


Figure 9 PCA loading plot for cluster 1

From a PCA of residuals [WINSTEPS] performed on the present data, one can note:

- Twice as much of the observed raw variance comes from persons rather than from items
- The variance (about 50%) unexplained by the initial Rasch analysis appears to be *associated with up to 3 contrasts* shown in the loading plot of Figure 9.
- Groups of items which share the same patterns of "*unexpectedness"*, probably also share a substantive attribute in common, a "secondary dimension" [Linacre, 2002]. Each cluster contains a set of items which seem to belong to a dimension orthogonal to the Rasch dimension. Cluster 1 is mainly ergonomically related questions; cluster 2 is mainly informational/cognitive; while cluster 3 is a mix of these two [Figure 9].

## 4.3 Decisions of conformity and significance testing

As in all conformity assessment, it is often a major challenge to separate 'true' intrinsic item variations from apparent dispersion arising from limited measurement quality.

### *4.3.1 Reliability*

Because limited measurement quality can lead to increased risks of incorrect decisions of conformity (e.g. accepting a non-conforming product or service), an important task is to evaluate correctly the actual measurement uncertainty.

Measurement reliability, according to Roach [2006], is a gauge of whether an outcome measure produces the same number each time a measurement instrument is administered. In using

persons as measurement instrument, it is necessary to separate repeated measurements with one individual from measurement with different persons. Repeated measurement can only be valid with the same individual because of the naturally existing inter-individual differences. There are several aspects to reliability; here we focus on *three of these*: self-reporting, internal consistency of accessibility, and rater performance:

1. *Self-reporting: Reliability estimated from wording & interpretation*

The Rasch analysis has not revealed significant differences among subjects in raw scoring related to interpretation of the wording of the questions (items) [see Figure 5].

Please note that the reliability is *estimated from inter-subject scale values*, thus including inter-individual differences, however, among the subjects the interpretation of each item is in great agreement and therefore supporting such a procedure.

The reliability of estimates of individual person scores averaged over all barriers [see Figure 6] has however been shown to be limited, especially at both extremes of the capability scale. As noted in connection with the results shown in Figure 4, the *mismatch* between the ranges of person capability and of barrier challenge *limits the reliability value*.

*2. Internal consistency of accessibility: Do all items in the outcome measure address the same underlying concept?*

Although there seems to be reasonable validity in the capability attribute estimates as demonstrated in Figure 8, a PCA of residuals revealed several principal components of variation beyond the first-order assumption in a Rasch analysis of a single item parameter, $\delta$, as shown in Figure 9.

*3. Rater performance: intra-rater consistency with repeats (reliability) and inter-rater consistency amongst different raters (validity)*

The measurement (standard) uncertainties calculated as reliability and validity are plotted in the various graphs of results in this article [Figures 5 – 8].

A gauge of measurement uncertainty in the Rasch attributes of capability θ and challenge δ is expressed as a standard error:

$$SE\left(\theta_i, \delta_j\right) = \sqrt{\frac{1}{\hat{P}_{success,i,j} \cdot \left(1 - \hat{P}_{success,i,j}\right)}} \qquad (10)$$

*4.3.2 Measurement/item separation and conformity decision-making*

A reliability coefficient ($R_\delta$) for the item attributes, δ, is calculated as:

$$R_\delta = \frac{True\ variance}{Observed\ variance} = \frac{\mathrm{var}(\delta)}{\mathrm{var}(\delta')} = \frac{\mathrm{var}(\delta') - \mathrm{var}(\varepsilon_\delta)}{\mathrm{var}(\delta')} \qquad (11)$$

If measurement uncertainty is estimated with Eq. 10, typical values of the reliability coefficient $R_\delta$ according to Eq. 11 in the present investigation, lie in the region of 0.8 [Figure 5]. At this value, about 75% of the observed variation is explained by item variations rather than limited measurement quality, which is considered acceptable [Linacre, 2002]. As is well known, the reliability coefficient in each attribute estimate for both person capability (θ) and barrier challenge (δ), is related in part to (a) the number of test persons [see discussion in Section 4.3.1] as well as (b) the number of items, for instance, according to the well-known Spearman-Brown prophecy

formula[3].

## 5 Discussion

In research on public-transport accessibility, the inter-variability is large among travellers with regard to capability. Therefore, all travellers do not have the same prerequisites for an accessible journey. Our research results show that differences among travellers have a significant effect on barrier challenge. A large variation in capability is matched by a narrow range of the pattern of item responses on barrier challenge. All individuals' ranking of items shows approximately the **same pattern,** indicating a good reliability. The questions are reliable independent of the individual's capability [Figure 4].

Among the subjects the interpretation of each barrier-challenge item is in great agreement. Because of the cross-sectional nature of our study, we lack longitudinal data. Therefore, we estimate "reliability" from inter-subject scale values rather than from intra-subject test-retest data.

As shown in Figure 4, the most challenging question (Q42 item) on "information" has a larger response variability than the least challenging question (Q44 item) "getting off the train". Potential explanations of the greater inconsistency found for information might be the unpredictable nature of informational barriers. Main complaints were e.g. inconsistent information or signs that were out of order as well as inability to obtain information at the stations [see Sundling, Berglund, Nilsson, Emardson, & Pendrill, 2015].

In the Rasch analysis, interindividual variability is excluded from the variance by standardisation. Our data indicates good within-individual variance [Figure 9]. Inter-individual variability is separated out from the limited measurement quality (inter-individual variability or measurement error).

There is a considerable distribution mismatch between the person attributes (person capability) and item attributes (barrier challenge) evident in Figure 4. There is a wide distribution of the level of individual capability (more to less capable) compared to the narrow distribution found for the item attributes (more to less challenge). Thus there is very low inter-individual variation in level of challenge among our 10 items. All item-attribute responses are located at the low-challenge side of the diagram. This means that all items turned out to constitute a low challenge for all individuals *regardless of their capability*. As a guide to improving future studies, this apparent mismatch could be reduced – thereby leading to better construct validity – by posing additional questions of a more challenging nature. On the other hand, because of the relatively larger sample of persons ($N$=162) compared to the number of items ($N$=10) in this study, one might expect that the distribution of person capability would be much wider than for the distribution for items referring to challenge. This is also the actual result as shown in Figure 4.

[3] https://en.wikipedia.org/wiki/Spearman%E2%80%93Brown_prediction_formula

# 6 Conclusions

The Rasch model is demonstrated to be an efficient tool for analysing how both person attributes and item attributes contribute to the overarching concept of accessibility in train traveling.

A set of reference barriers has to be selected to be used reproducibly in various travel situations. Such metrological traceability provides all the advantages commensurate with physical measurement. A barrier-challenge reference (or standard) would allow an estimate of each person's capability to travel involving a range of barriers of various challenges. By measuring a set of *reference barriers of known challenge*, the barrier challenge for other railway barriers can be metrologically calibrated. There are two approaches available for constructing such a measurement instrument, the psychometric or the psychophysical approach [Berglund, Rossi, Townsend, & Pendrill, 2012].

The Rasch method is a tool for revealing incorrect scoring. This would have been explicit in the extreme ends of the Likert scale (category scale), [Likert, 1932]. Conversely, our data show averages of the present scores on each question to be mostly *mid-range*, where the perceptive scale is an approximately linear estimation of each person's scores [Figure 5]. Therefore, in the present case, there is little evidence that the travellers tend to score incorrectly. However, uncertainties are large for certain test persons. These are traced to the low score – high capability (less challenge) end of the scale [Figure 4].

**Acknowledgements**
Financial support is gratefully acknowledged from the Swedish Transport Administration, the Swedish National Metrology Program, the Swedish Research Council for the Environment (FORMAS) and the Norrbacka-Eugeniastiftelsen in Stockholm, Sweden. The research results of the present project will help to achieve a more flexible and independent travel behaviour for all, thus contributing to the Swedish governmental assignment to its Transport Administration.

**Declarations of interest:** None

## References


Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika, 43,* 561-73.

Berglund, B., Rossi, G.B., Townsend, J.T., &Pendrill, L.R. (Eds.) (2012). *Measurement with Persons. Theory, Methods, and Implementation Areas*. New York, Psychology Press.

Church , R. L. &Marston, J. R. (2003). Measuring accessibility for people with a disability, *Geographical Analysis, 35,* 83 – 96, doi: 10.1353/geo.2002.0029

Emardson, R., Jarlemark, P., Pendrill, L. R., Sundling, C., Nilsson, M. E., & Berglund, B. (2012). Measurements of accessibility to rail transport systems. In F. Pavese, M. Bär, J-R. Filtz, A. B. Forbes, L. R. Pendrill, & K. Shirono (Eds.) *Advanced Mathematical and Computational Tools in Metrology and Testing IX*, (pp 136-142). World Scientific: Singapore

Fisher, R. A. (1924). The conditions under which chi-squared measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society, 87,* 442-450.

Fisher, Jr. W. P. (1997). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*(2), 87 – 113.

Iwarsson, S., Jensen, G., & Ståhl, A. (2000). Travel chain enabler: Development of a pilot instrument for assessment of urban public bus transport accessibility. *Technology and Disability, 12,* 3-12.

Jensen, S., Iwarsson, S., & Ståhl, A. (2002). Theoretical understanding and methodological challenges in accessibility assessments, focusing the environmental component: an example from travel chains in urban public bus transport. *Disability and Rehabilitation, 24*(5), 231-242.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *140*, 1–55.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Massof, R. W. (2005). Applications of stochastic measurement models to visual function rating scale questionnaires. *Opthalmic Epidemiology, 12,* 1 – 22. doi: 10.1080/09286580590932789, http://dx.doi.org/10.1080/09286580590932789

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, **47***, 149-174

Pendrill, L. R. (2014). Using measurement uncertainty in decision-making & conformity assessment. *Metrologia, 51*(4), S206. doi:10.1088/0026-1394/51/4/S206

Pendrill, L. R 2018, "Assuring measurement quality in person-centred healthcare", *Measurement Science & Technology*, Volume 29, Number 3, 034003 special issue Metrologie 2017, https://doi.org/10.1088/1361-6501/aa9cd2

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. .In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV, 321–334. Berkeley, California: University of California Press. Available free from Project Euclid.

Rimmer, J. H., Riley, B., Wang, E., & Rauworth, A. (2004). Development and validation of AIMSFREE: accessibility instruments measuring fitness and recreation environments. *Disability and Rehabilitation*, *26*(18), 1087-1095.

Roach, K. E. (2006). Measurement of Health Outcomes: Reliability, Validity and Responsiveness. *Journal of Prosthetics and Orthotics*, *18*(6), P8-P12.

Sundling, C., Berglund, B., Nilsson, M. E., Emardson, R., & Pendrill, L. R. (2013). New perspective on the accessibility of railway transport for the vulnerable traveller. *Joint IMEKO TC1-TC7-TC13 Symposium, Measurement across physical and behavioural sciences*, 4-6 September 2013, Genova, Palazzo Ducale (IT), *Journal of Physics: Conference Series* 459, 012021, doi: 10.1088/1742-6596/459/1/012021

Sundling, C., Berglund, B., Nilsson, M. E., Emardson, R., & Pendrill, L. R. (2015). Two models of accessibility to railway traveling for vulnerable, elderly persons. *Measurement, 72*, 96-101. http://dx.doi.org/10.1016/j.measurement.2015.02.053

Sundling, C., Berglund, B., Nilsson, M. E., Emardson, R., Pendrill, L. R. (2014). Overall Accessibility to Traveling by Rail for the Elderly with and without Functional Limitations: The Whole-Trip Perspective, *International Journal of Environmental Research and Public Health 11*11(12), 12938-12968. doi:10.3390/ijerph111212938.

Svensson, E. (2001). Guidelines To Statistical Evaluation Of Data From Rating Scales And Questionnaires. *Journal of Rehabilitation Medicine, 33*(1), 47–48.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology, 76,* 103–154.

E B Wilson and M M Hilferty 1931, "The Distribution of Chi-square", *Proc. NAS*, **17**, 684 - 8

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10(3),* 509 – 511. https://www.rasch.org/rmt/rmt103b.htm

B D Wright and M H Stone, 1979. *Best Test Design*, ISBN 0-941938-00-X. LC# 79-88489

## Appendix Extract of questions about barriers

*Swedish questions translated to English*

38. Taken together, was it easy or difficult to move around in the long-distance train station areas. Please consider the whole trip, including both departure and arrival stations. (*Sammantaget, var det lätt eller svårt att röra sig inom stationsområdena för fjärrtåg? Tänk på hela resans avgångs- och ankomststationer*).

39. Taken together, was it easy or difficult to retrieve information at the long-distance train stations? Consider the departure and arrival stations for long-distance trains you visited. (*Sammantaget, var det lätt eller svårt att ta del av informationen på fjärrtågsstationerna? Tänk på de avgångs- och ankomststationer för fjärrtåg du besökte*).

40. Was it easy or difficult to get on board the long-distance train(s)? (*Var det lätt eller svårt att ta sig ombord på fjärrtåget/tågen*)?

41. Was it easy or difficult to move around on board the long-distance train(s)? (*Var det lätt eller svårt röra sig ombord på fjärrtåget/tågen?*).

42. Was it easy or difficult to get information on board the long-distance train(s)? (*Var det lätt eller svårt att ta del av information ombord på fjärrtåget/tågen*)?

43. Was it easy or difficult to use or to get to the lavatory on board the long-distance train(s). (*Var det lätt eller svårt att ta sig till toaletten ombord på fjärrtåget/tågen?*)

44. Was it easy or difficult to alight from the long/distance train(s)? (*Var det lätt eller svårt att komma av fjärrtåget/tågen?*)

45. Was it easy or difficult to get an overview and find your way in the station areas? Consider long-distance train departure and arrival stations you visited. (*Var det lätt eller svårt att orientera sig och hitta inom stationsområdena? Tänk på de avgångs- och ankomststationer för fjärrtåg du besökte*).

46. How did you perceive the personal service from the staff during the trip, e.g. personal treatment, assistance when needed etc? (*Hur var den personliga servicen från personal under resan t ex bemötande, hjälp vid behov, etc*)?

51 Taken together, how accessible do you perceive train traveling to be? (*Sammantaget, hur tycker du att tillgängligheten är för dig vid tågresor?*)