Rasch analysis of the Listening Effort Questionnaire – Cochlear Implant (LEQ-CI)

Sarah E. Hughes[1], Alan Watkins[1], Frances Rapport[2], Isabelle Boisvert[3], Catherine M. McMahon[4], Hayley A. Hutchings[1]

[1]Patient and Population Health and Informatics, Swansea University Medical School, Singleton Park, Swansea, UK SA2 8PP

[2]Australian Institute of Health Innovation, Macquarie University, NSW, Australia

[3]Faculty of Medicine and Health, School of Health Sciences, University of Sydney, Sydney, Australia

[4]Audiology, H:EAR, Department of Linguistics, Faculty of Medicine, Health and Human Sciences, Macquarie University, NSW, Australia

Hughes, Sarah E; Watkins, Alan; Rapport, Frances; Boisvert, Isabelle; McMahon, Catherine M; Hutchings, Hayley A. Rasch Analysis of the Listening Effort Questionnaire—Cochlear Implant. *Ear and Hearing*: May 4, 2021 - Volume Publish Ahead of Print - Issue - doi: 10.1097/AUD.0000000000001059

Financial disclosures/conflicts of interest:
This study was funded by a Pathway to Portfolio Grant, Swansea Bay University Health Board. SEH is an independent consultant to Cochlear Ltd. For the remaining authors, none were declared.

All correspondence should be addressed to:
Sarah Hughes
Swansea University Medical School
Institute of Life Sciences 2
Singleton Park, Swansea, UK, SA2 8PP
E-mail: sarah.hughes@swansea.ac.uk

ABSTRACT

Objectives: Listening effort may be defined as the attentional and cognitive resources needed to understand an auditory message, modulated by motivation. Despite the use of hearing devices such as hearing aids (HAs) or cochlear implants (CIs), the requirement for high listening effort remains a challenge for individuals with hearing loss. The Listening Effort Questionnaire - Cochlear Implant (LEQ-CI) is a hearing-specific patient-reported outcome measure (PROM) that has been designed for use in the CI candidacy and rehabilitation process to assess perceived listening effort in everyday life in adults with severe-profound hearing loss. The LEQ-CI has been developed in line with international consensus-based standards for best practice in PROM construction. The aim of this study was to improve the measurement precision of the LEQ-CI and to assess its psychometric measurement properties.

Design: A field test was undertaken with 330 CI patients from five NHS auditory implant centres in the United Kingdom (UK). Participants were adults (≥18 years of age), had a severe-profound hearing loss, and met the UK candidacy criteria for cochlear implantation specified by the National Institute for Health and Care Excellence (NICE). Participants completed and returned an anonymised 29-item (each with a 5-point or 7-point response option), draft version of the LEQ-CI (LEQ-CI^{29}) and a demographic questionnaire. Rasch analysis was undertaken using Winsteps software and the partial credit model to assess rating scale function and item fit. Results informed refinements to produce a 21-item version (LEQ-CI^{21}) which underwent a further Rasch analysis.

Results: The sample was predominantly female: 60.3% (n = 191). Median age of participants was 66 (range 21 - 89) years, with 7.3% (n = 24) of respondents being CI candidates and 92.7% (n = 306) being CI recipients. Mean duration of implantation was 3.8 (SD 4.8) years. Initial Rasch analysis of the LEQ-CI^{29} revealed poor rating scale functioning. Collapsing the

5- and 7-point rating scales to 3-point and 4-point scales and removing eight items produced a 21-item PROM (LEQ-CI^{21}) which met the Rasch criteria for rating scale functioning. Rasch analysis of the LEQ-CI^{21} showed good fit to the Rasch model. No items showed misfit and dimensionality analysis supported the existence of a single Rasch dimension, defined as perceived listening effort in daily life. Person reliability was 0.91 and the person separation index was 3.28, establishing four levels of person ability. The item separation index was 9.69, confirming the item hierarchy. No items showed differential item functioning (DIF) for gender or age. The item difficulty range was -0.81 to 1.05, the person ability range for non-extreme persons was -3.54 to 2.49, and the mean person ability was -0.31.

Conclusions: Overall, the LEQ-CI^{21} was found to meet the Rasch model criteria for interval level measurement. The LEQ-CI^{21} is the first PROM to be developed specifically for the measurement of perceived listening effort and one of the first PROMs for use with CI patients to be developed using Rasch analysis. The LEQ-CI^{21} has the potential to be used as a research tool and in clinical practice to evaluate perceived listening effort in daily life. Further psychometric evaluation of the LEQ-CI^{21} is planned. (513words)

INTRODUCTION

Listening effort may be defined as the mental exertion required to attend to, and understand, an auditory message (McGarrigle et al. 2014). It is a significant problem for individuals with hearing loss insofar as they are required to exert greater listening effort than individuals with normal hearing despite the use of hearing devices such as hearing aids and/or cochlear implants (Ohlenforst & Zekveld 2017). Damage to the auditory pathway associated with sensorineural hearing loss results in the degradation of the quality of the auditory signal as well as its intensity. Furthermore, listening often takes place in less than optimal acoustic conditions requiring individuals to deploy additional cognitive resources to extract, decode and comprehend the incoming auditory message (Edwards 2016). Sustained effortful listening in adults with hearing loss is known to contribute to fatigue (Holman et al. 2019; Hornsby 2013; Hornsby et al. 2016; Hornsby & Kipp 2016), increased levels of workplace absence/sickness (Nachtegaal et al. 2009), poor mental health (Amieva et al. 2018), social withdrawal and isolation (Ramage-Morin 2016), and negatively impact quality of life (McRackan et al. 2017).

A hearing-specific, validated measure of listening effort could provide hearing healthcare professionals with a means to support interventions designed to reduce listening effort and mitigate its wider effects. Several methods have been developed to measure listening effort; however, deployment of these measures has been predominantly within research settings. These techniques include behavioural measures that rely on task performance as a proxy-measure of effort (cf. Gagné et al. 2017 for a review); physiological measures such as electroencephalography (EEG, Miles et al. 2017) and pupillometry (Zekveld et al. 2018); and self-report measures of perceived listening effort. Frequently used alongside behavioural or physiological measures, self-report measures have utilised visual analogue scales (VAS, Bräcker et al., 2019) or ordinal rating scales (Johnson et al. 2015) to rate intensity of effort during a specific listening task.

To complicate matters, the relationship between these various measures of listening effort is not yet well understood and there is a generalised lack of agreement in the empirical literature among findings using these different types of measures (Hornsby 2013; Miles et al. 2017). This lack of consensus has led to the suggestion that these various methods may be measuring different aspects of the listening effort construct (Strand et al. 2018; Strand et al. 2020). For example, Alhanbali et al. (2018) used factor analysis to explore the relationship among behavioural, physiological and self-report measures, revealing four underlying dimensions. It seems likely that a range of measures, with method selection depending on the aims of measurement, will be required if listening effort is to be evaluated appropriately.

Patient-reported outcome measures (PROMs) offer promise as a viable clinical tool for clinicians and researchers wishing to gain insight into perceived listening effort in daily life as experienced by individuals with hearing loss. PROMs are self-report measures, typically validated questionnaires, that measure aspects of a person's health (such as symptoms, functioning, and quality of life) known only to the individual (Devlin & Appleby 2010; FDA 2020). Originally developed for clinical trials research, these tools are now embedded in routine clinical practice across a range of health conditions (Field et al. 2019). There is a long tradition of self-report in audiology and, commensurate with this tradition, there are a number of hearing-specific PROMs (Granberg et al. 2014). Most of these measures have been developed for use in the population of adults with mild-moderate hearing loss (MMHL) who use hearing aids. Furthermore, most hearing-specific PROMs have been developed and validated using traditional psychometric methods which limits their use as clinical measures with individual patients. Notably, no hearing-specific PROMs have been identified that measure perceived listening effort in daily life for the population of adults who use CIs. A systematic review undertaken by the authors identified a number of PROMs containing items considered to measure perceived listening effort in hearing loss. However, no dedicated measures that have been developed with substantial input from CI recipients, or validated in a large, representative sample, were found (Hughes et al. 2017a; Hughes et al. 2017b).

International guidance and consensus-based standards are available to guide PROM development and validation (Mokkink et al. 2018; Patrick et al. 2011a; Patrick et al. 2011b). The PROM development process includes content generation through direct engagement with members of the target population; pretesting and qualitative refinement; and psychometric evaluation to establish the new PROM's measurement properties (FDA 2009; FDA 2020). Psychometric evaluation concerns the measurement of the target construct which is often largely unobservable (e.g., effort, pain or fatigue). It aims to establish whether the conceptualisation of the variable of interest (i.e., as defined by the PROM's content) has been operationalised successfully (Hobart and Cano 2009). There are a number of psychometric methods that may be used to evaluate a PROM's measurement characteristics, with Classical Test Theory (CTT) and Item Response Theory (IRT) being two commonly used methods. CTT is the most widely used of these methods and has dominated the field of psychometrics for the last century (Streiner et al. 2015). Also known as weak true score theory, CTT examines how measurement error affects rating scale scores. It proposes that an observed score may be broken down into a true score and an error score, and although these values are unknown, they may be estimated (Cano et al. 2011). CTT methods construct PROMs that generate ordinal (rather than interval-level) data. The data are considered to represent ordered counts, and the difference between adjacent categories on a response scale cannot be assumed to have the same meaning (e.g., the distance between "sometimes" and "frequently" may not be the same as "frequently" to "always") (Stevens 1946) . CTT evaluation is based on the instrument as a whole and uses evidence primarily from correlations and descriptive statistics (Hobart & Cano 2009).

Rasch measurement theory, a variant of IRT, is a modern psychometric method that is used increasingly alongside CTT to develop and validate PROMs (Aryadoust et al. 2019). Developed by Danish mathematician Georg Rasch, the Rasch model is a mathematical ideal which specifies a set of criteria for the construction of interval level measures. The model states that the probability that a person will affirm an item is a logistic function of the difference between a person's trait level (this is expressed as person ability in Rasch terms) and the amount of the trait expressed by the item (expressed as item difficulty in Rasch terms), and is only a function of that difference (Rasch 1960). Therefore, the higher a person's ability relative to item difficulty, the higher the probability of the person endorsing that item. On this basis, a summed raw score may be used to derive an estimate of the target trait (i.e., perceived listening effort) on an interval scale. Moreover, parameter separation, a key characteristic of the Rasch model, means item difficulty is independent of the sample and that a person's level of trait (i.e., ability) is independent of items (Bond & Fox 2015).

Rasch analysis is the term used to describe the formal evaluation of a PROM against Rasch's mathematical measurement model. If data are found to conform to the model criteria, then PROM developers can theoretically be confident that interval-level scales (measured using the Rasch logit or log odds unit) have been constructed; that is, a respondent's answers can be used to determine their precise location on a continuum that ranges from low to high levels of perceived listening effort. Unlike CTT, the Rasch model enables estimates suitable for individual-person analyses as well as group comparisons (Hobart & Cano 2009). These two properties are of particular importance if a PROM's intended use is with individual patients in the clinic. Because psychometric evaluation focuses on item-level rather than scale-level assessment, a Rasch analysis also provides opportunity for PROM developers to refine individual items and response scales, addressing potential sources of error and enabling measurement precision to be improved (Boone 2016).

To satisfy the Rasch model, several criteria must be met. Firstly, an instrument's items must demonstrate unidimensionality and local independence. For a scale to be unidimensional, the items must only measure a single construct. Local independence means items should only correlate through the latent trait that the test is measuring (i.e., a response to an item must not influence responses to other items) (Lord et al. 1968). Thirdly, items must fit the Rasch model expectations (i.e., item fit), as assessed through the examination of fit statistics. Lastly, measures must be invariant such that the relative location of any two persons on the construct continuum should be independent of the items used to make that comparison (Hobart & Cano 2009).[1]

The opportunity for item refinement and construction of an instrument capable of interval-level measurement means Rasch analysis could be argued to be a sensible choice for undertaking a first evaluation of a new PROM's psychometric measurement properties. Building upon developmental work reported elsewhere (Hughes et al. 2018), this paper presents the quantitative refinement and initial validation of the Listening Effort Questionnaire-Cochlear Implant (LEQ-CI), a new PROM measuring perceived listening effort in daily life (Hughes et al. 2019). The specific study aims were: 1) to use Rasch analysis to refine the LEQ-CI's rating scales and response categories; and 2) to assess the fit of the LEQ-CI to the Rasch model. The overall goal is to make available to clinicians and researchers a carefully developed and validated PROM of perceived listening effort that is appropriate for use both in research and clinical practice.

[1] For readers interested in undertaking further study of psychometrics including Classical Test Theory and Rasch measurement theory, we suggest Bond et al. (2020), Boone et al. (2014), and Streiner et al. (2015) as introductory texts.

MATERIALS AND METHODS

Participants

Linacre (1994) specifies a minimum sample size of 250 participants to increase precision and robustness of estimates when conducting a Rasch analysis. Altogether, a total of 511 cochlear implant patients from five regional National Health Service (NHS) auditory implant centres in the UK were approached. Patients were invited to participate if they met the following inclusion criteria: 1) they were adults aged 18 years or older; 2) they had a diagnosis of severe-profound sensorineural hearing loss (SNHL); 3) they were either a CI candidate or CI recipient according to the UK candidacy criteria (National Institute for Health and Care Excellence (NICE) 2009); 4) they were able to read/write in competent English; 5) they had no history of medical conditions that would preclude their ability to self-complete the LEQ-CI; and 6) they were able to return the completed questionnaire by post. Both candidates and recipients were invited to participate to increase the likelihood that persons representing the full continuum of the latent trait (i.e., low to high levels of perceived listening effort) were included in the study sample. Participating CI teams identified potential participants and sent each participant an invitation letter, patient information sheet, and a booklet comprised of a demographic questionnaire and the draft LEQ-CI[29]. A postage-paid reply envelope was included for returning the completed booklet to the study team.

Ethical considerations

The study was approved by the North East - Newcastle and North Tyneside 2 Research Ethics Committee (Ref: 18/NE/0320) and the Swansea University-Swansea Bay University Health Board’s Joint Study Review Committee.

The Listening Effort Questionnaire – Cochlear Implant (LEQ-CI)

The LEQ-CI was developed to measure perceived listening effort in daily life in adult CI patients. It is intended for use as both a candidacy and outcome measure in routine clinical practice as well as in clinical research studies. The LEQ-CI's underlying conceptual framework (see Figure 1) was developed from a mixed methods qualitative study that utilised in-depth focus groups and a follow-up postal survey with cochlear implant patients (Hughes et al. 2018). Findings from these first-hand accounts suggested that perceived listening effort in daily life is a complex construct that includes the mental energy needed to attend to and process an auditory signal while adapting and compensating for hearing loss. Effort is subject to a cost-benefit analysis and is mediated by a number of motivational factors, particularly that of social connectedness and pleasure. These findings also supported the heuristic Framework for Understanding Effortful Listening (FUEL). The FUEL proposes that listening effort depends not only on hearing difficulties and task demands, but also on the listener's motivation to expend mental effort in the challenging situations of everyday life (Hughes et al. 2017b; Pichora-Fuller et al. 2016).

An item pool was generated from the focus group accounts; a review of the literature; and extant PROMs measuring relevant concepts within the conceptual framework (e.g., social connectedness, attention, and memory) (Hughes et al. 2018). Item candidates were selected for inclusion in the draft LEQ-CI using the PROMIS® qualitative item review procedures (DeWalt et al. 2007). The draft instrument (including respondents' instructions, items and rating scales) was appraised for relevance, comprehensiveness, comprehensibility and acceptability by an online survey of subject matter experts and also by CI patients in a series of cognitive interviews (Hughes et al. 2019a). Qualitative findings from the online survey and cognitive interviews were used to refine the LEQ-CI prior to undertaking further refinement or performing an initial validation using Rasch analysis.

Qualitative refinement produced a draft LEQ-CI comprised of 29 items (LEQ-CI[29]) covering four domains (see Table 1): 1) the effort of attending (5 items); 2) the effort of processing (9 items); 3) the effort associated with adapting and compensating for a hearing loss (4 items); and 4) motivation (11 items). Items were measured using 5-point or 7-point ordinal scales assessing either frequency (i.e., 1 = never; 2 = rarely; 3 = occasionally; 4 = sometimes; 5 = frequently; 6 = usually; 7 = always) or intensity (i.e., 1 = not at all; 2 = slightly; 3 = moderately; 4 = quite a bit; 5 = extremely), with higher scores indicating greater levels of perceived listening effort. All relevant items were included in the draft instrument to maximise coverage of the target construct (Bond & Fox 2015). A broad response scale was utilised to enable greater differentiation in the judgements being made (Krosnick & Presser 2010) and to enable the use of Rasch analysis for optimisation of the rating scale response categories (Bond & Fox 2015). The inclusion of a relatively large number of items reflected the expectation that some items could be removed to optimise model fit without sacrificing construct coverage. Readability of the full participant questionnaire booklet (including both the LEQ-CI[29] and demographic questionnaire) was scored using the web-based application Readable (www.readable.io). The booklet received a Flesch-Kincaid Grade Level score of 4.8 and SMOG Index of 7.8 (Flesch 1948; McLaughlin 1969). These results suggested the wording of the instructions, items and response scales of the booklet would be understood by a respondent with literacy skills equivalent to a 4th grade reading level based on the American education system. A SMOG score 7.8 suggested the LEQ-CI[29] would be understood by 93% of the UK population (NHS Digital Service Manual 2019). Estimated reading time was approximately seven minutes.

Data collection

Participating CI clinical teams sent a copy of the study information leaflet, the LEQ-CI[29], a demographic questionnaire, and a reply-paid envelope to CI patients who met the study eligibility criteria. Patients who wished to participate completed the questionnaires and returned these directly to the research team in the reply-paid envelope. Paper-based questionnaires were used to optimise responding in an older CI patient population and to promote inclusivity by limiting digital exclusion (Rowen et al. 2019; Smith et al. 2019). To ensure participant anonymity, each booklet was coded with a unique identifier and no identifiable information was collected. The clinical team at each CI centre held a master list linking the identifiers to patients for purposes of data chasing (i.e., a reminder letter sent two weeks after the initial mail out) and to pay each participant an honorarium in the form of a £10 GBP retail gift card as a gesture of appreciation for their time and effort. All participant follow-up was undertaken solely by the clinical team. To maintain anonymity, consent was presumed when participants completed and returned the questionnaire booklet directly to the research team (UK Data Service 2020).

Missing data

Any LEQ-CI[29] items not answered by respondents were considered missing data. The full dataset, including cases with missing data, was included for analysis using Winsteps analysis software (Version 4.4.7, www.winsteps.com). Unlike CTT which requires a complete dataset, it is standard practice in Rasch analysis to include cases with missing data. Winsteps’s use of Joint Maximum Likelihood Estimation (JMLE) enables parameter estimation for each missing case via a likelihood function based on the available data (Linacre 2020a; Waterbury 2019).

Data analysis plan

Data analysis was undertaken in two stages: 1) a description of sample characteristics, response rates, data quality and missing data; and 2) psychometric evaluation using Rasch analysis, performed with Winsteps using JMLE and the partial credit model (Masters 1982). The partial credit model is an unconstrained model suitable for polytomous data where the response structure is permitted to vary across items (Bond and Fox 2015). The specific steps of the Rasch analysis are described below.

Assessment of rating scale functioning - The rating scale structure of a PROM must fulfil several criteria in order to satisfy the Rasch model requirements (Linacre 2002a). Firstly, all items must orient with the latent variable. Any reverse-scored items should be re-scored, as the rating scale categories may otherwise function differently. Secondly, a greater level of perceived listening effort by a person should equate to higher scores on the LEQ-CI. As such, it is expected that the higher the level of effort, the higher the response category that will be endorsed, with response category thresholds advancing monotonically across the measurement continuum. For each item, a category probability curve shows the probability of observing each ordered category according to the Rasch model. If category thresholds do not progress in a linear manner (i.e., along a continuum partitioned into equal, contiguous intervals), or if the distance between two categories is judged to be inadequate, thresholds are considered disordered. Disordering occurs when respondents select a response option that is inconsistent with their ability and implies that a rating scale's categories may be confusing or difficult to use (Boone et al. 2014; Pallant & Tennant 2007). Category disordering may be resolved by collapsing one or more rating scale categories with adjacent categories (Boone & Noltemeyer 2017). Thirdly, each response category should be endorsed by a minimum of ten persons per response category as low category endorsement could mean the step calibration is imprecisely estimated and potentially unstable (Linacre 2002a). A uniform distribution of observations across response categories is optimal for step calibration. Lastly, category fit statistics (i.e., as indices of how well the rating scale category structure meets the Rasch model requirements) must fall within an acceptable range. Infit and outfit are the indices used to assess model fit. Infit denotes inlier-sensitive or information-weighted fit, whilst outfit refers to outlier-sensitive fit. Fit statistics are reported as mean square values (MNSQ), or as standardised (ZSTD) values. MNSQ is the mean of the squared residuals for an item and is sample independent, whilst ZSTD is a transformation of the mean square value with a sample size correction. Outfit mean square values less than 2.0 logits (MNSQ < 2.0 logits) were considered to be evidence of acceptable category fit (Bond et al. 2020).

Assessment of item and person fit to the model - Fit statistics show the extent that item performance and person ability differ from the Rasch modelled expectations for fundamental measurement. Mean-square values were reported using Winsteps, with values between 0.5 and 1.5 logits indicating acceptable fit (Linacre 2002b; Wright & Linacre 1994). Item fit was confirmed by inspecting the item characteristic curves (ICCs) for each item. The ICC is an ogive-shaped plot of the probabilities of responding to an item for any value of the underlying trait (Bond & Fox 2015). The empirical and model ICCs should present a close alignment of the observed and predicted scores for an item.

Person-fit measurement aims to identify individuals in the sample whose response patterns deviate from the Rasch-modelled expectations (Boone et al. 2014). Person misfit has the potential to affect item fit and, as a consequence, a scale's internal construct validity. Person fit was assessed in the same manner as item fit. Examination of infit and outfit statistics with mean square values outside the 0.5 – 1.5 logit range were considered evidence of misfit (Linacre 2002b). Underfit, defined as high mean square values (MNSQ > 1.5), is indicative of noise or randomness in the data with potential to influence item calibration;

therefore, it is standard practice to remove underfitting persons from the sample when evaluating items' fit to the Rasch model (Bond & Fox 2015).

Assessment of unidimensionality and local independence - Unidimensionality (i.e., the set of items under assessment represents a single construct) and local independence of items (i.e., the entire correlation between the items is captured by the latent trait) are core requirements of the Rasch model (Hagquist et al. 2009). Unidimensionality was assessed in Winsteps using a principal component analysis of the residuals (PCAR) (Linacre 2018). The following criteria were used to determine whether the LEQ-CI met the requirement of unidimensionality: 1) an eigenvalue less than 3.0 on the first residual contrast (Fan & Bond 2019; Linacre 2018); 2) the percentage variance explained by the first contrast of 5% or less (Smith et al. 2007); and 3) disattenuated correlations between the person measures for the item clusters on the first residual contrast greater than 0.70 (Linacre 2018). Local independence was appraised by examining the item fit statistics for overfit and correlations between items of the standardised residuals. Overfit and a substantial correlation of the standardised residuals for two items was considered evidence of local dependence (Fan & Bond 2019). A range of critical values for residual correlations have been reported in the literature (Christensen et al. 2017). For this study, residual correlation values of $r < 0.4$ were considered evidence of low dependency (Bond et al. 2020; Linacre 2020a).

Assessment of differential item functioning (DIF) - Differential item functioning (DIF) analysis investigated whether the LEQ-CI's items measure perceived listening effort in daily life in the same way for different sub-groups of the sample when both groups have equal levels of effort. The presence of DIF implies a lack of measurement invariance that can impact model fit (Tennant & Conaghan 2007). In the case of the LEQ-CI, DIF was explored for gender (i.e., male and female) and age (i.e., adults aged less than 70 years and adults, aged 70 years and older, as an approximation of the median age of the sample). Group comparisons (e.g., males v females) were assessed for each item, examining probability of DIF and effect size. Items found to have p-values less than 0.05 and an effect size (i.e., DIF contrast value) of 0.64 or greater were considered to have sizable DIF (Linacre 2020).

Targeting of the scale - Targeting assesses the ability of an instrument's items to measure the full range of persons in a sample. Items that are used to represent the latent construct (i.e., perceived listening effort) should collectively form a "difficulty" hierarchy, with difficulty defined as the amount of trait represented by an item. The item hierarchy should range from the item representing the least effort to the item measuring maximum effort. In a Rasch analysis, item difficulty and person ability is measured using log-odds units or logits.[2] The scale is always centred on zero logits, representing the item of average difficulty for the scale (Tennant & Conaghan 2007). Item difficulties and person abilities are plotted on the same logit scale and, for well-targeted instruments (i.e., one that is neither too easy nor too difficult), the distribution of persons should closely match the distribution of items. The mean person location and mean item location should correspond, with both location scores close to zero. Targeting of the LEQ-CI was assessed by inspecting: 1) the person-item distribution map 2) the summary statistics of item and person measures; and 3) the location range for items and persons on the Rasch scale.

[2] The terms "item difficulty" and "person ability" are used in this manuscript to maintain consistency with Rasch terminology. Item difficulty refers to the level of perceived listening effort (i.e., the latent trait) that an item measures. Person ability is an estimate of the underlying trait or attribute being measured in an individual. In the case of the LEQ-CI, high ability is indicative of a low level of perceived listening effort (i.e., a person who reports listening to be relatively effortless).

Assessment of reliability - In Rasch analysis, reliability is defined as “reproducibility of relative measure location”, or the probability that persons or items with high measures actually do have higher measures than items or persons with low measures (Wright & Masters 1982). Reliability is sample size dependent, and measured in Winsteps using separation indices. Separation indices indicate the number of distinct levels of functioning that can be distinguished; ability in the case of persons; and difficulty in the case of items (Duncan et al. 2003; Wright & Masters 1982). As a heuristic, a person separation index of 1.50 may be considered to represent an acceptable level of separation, whereas an index of 2.00 to represent a good level of separation and index of 3.00 represent to an excellent level of separation (Wright & Masters 1982). A second statistic, the person reliability index, indicates the replicability of person ordering if the same sample were given another set of parallel items measuring the same construct. Higher values are indicative of better reliability with values exceeding thresholds of 0.7, 0.8, and 0.9 indicating “acceptable”, “good”, and “excellent” replicability (Bond & Fox 2015). Item separation and reliability is reported in the same manner as person separation and reliability. The item separation index is used to verify the item hierarchy (i.e., ordering of items according to difficulty) as confirmation of the construct validity of an instrument (Wright & Masters 1982). Low item separation (< 3) imply that the person sample is not large enough to confirm the item difficulty hierarchy (Linacre 2011). High item reliability suggests item ordering would be replicated if the LEQ-CI was given to a new but identical person sample (Wright & Masters 1982).

RESULTS

From a total of 511 questionnaires distributed, 330 cochlear implant patients completed the LEQ-CI[29], a response rate of 64.6%. The respondents represented a wide age range (mean = 62.5 SD = 15.05, range = 21 – 89) that was negatively skewed with one third of respondents aged 70 years or older (n = 125, 37.9%). There was a higher proportion of females (n = 199, 60.3%) to males, and the majority of respondents used one CI only (n = 193, 58.5%) or one CI and a hearing aid (n = 106, 32.1%). No notable differences between the candidate and recipient sub-samples were found for age, gender, or age at diagnosis of hearing loss which suggested the subgroups were broadly similar in their demographic characteristics. The sample characteristics were consistent with demographic trends for this patient group (Amin et al. 2020). The demographic characteristics for the sample are reported in Table 2.

Data quality

All data were extracted from the questionnaire booklet and entered manually into REDCap (Version 8.10.2), clinical data management software hosted by Swansea Trials Unit, Swansea University. Double entry was completed by the lead author (SEH) for 10% (n=33) of the returned questionnaires. Eight data entry errors were detected (0.58%) across 1386 datapoints.

The complete dataset for all 330 participants was exported from REDCap to an Excel spreadsheet and uploaded to Winsteps. Prior to upload into Winsteps, the Excel spreadsheet was checked for formatting, and all reverse-scored items were transformed.

Missing data

Responses were missing for 73 items (0.007%) from the LEQ-CI[29] dataset of 9,570 (29 items x 330 respondents) responses. Eight respondents failed to complete one or more pages of the LEQ-CI[29] accounting for 72.6% (n = 53) of the missing responses. On a per item basis, the median number of missing responses was 2 (range = 0 – 6 missing responses, corresponding to missing response rates of 0% - 1.8%). No further discernible response patterns were identified from visual inspection of the data.

Assessment of rating scale functioning and instrument refinement

The first step of the Rasch analysis involved assessment of rating scale of functioning for the LEQ-CI[29]'s items, following the guidelines proposed by Linacre (2002a). The majority of the LEQ-CI[29]'s items showed multiple problems with category functioning when 5-point or 7-point rating scale structures were used. Categories failed to advance monotonically for six items; disordered category fit statistics were reported for nine items; and six items had response categories with fewer than 10 respondents. The category probability curves showed threshold disordering and a lack modal peaks for all response options. To improve rating scale function, adjacent response categories were collapsed to construct either a 3-point or a 4-point rating scale. Restructuring of the rating scale categories proceeded iteratively on a per item basis with the aim of optimising the response structure. With each iteration, the revised rating scale was reviewed to check for uniform category endorsement, to establish whether categories advanced monotonically and to confirm that outfit MNSQ values were less than 2.0 logits as evidence of model fit. The category probability curves were inspected, checking for modal peaks and advancing categories (Bond & Fox 2015; Linacre 2002a; Wright 1996).

Despite refinements to the rating scale categories, eight items continued to exhibit problems with scale functioning. No discernible patterns in terms of these items' content, difficulty, or response category structure, were identified. These items were removed to yield a 21-item version with all items satisfying the essential criteria for rating scale functioning (Linacre 2002a). The LEQ-CI[21]'s rating scales were found to be oriented with the latent variable (i.e., increasing category score = increasing effort requirement), average measures were found to advance monotonically, and outfit mean-square values were less than 2.0 logits. Table 3 presents a summary of the category structure for the refined rating scales of the LEQ-CI[21]. Most category probability curves showed good rating scale functioning; however, six items had category probability curves with ordered thresholds that failed to show clear modal peaks for intermediate categories. The lack of modal peaks was attributed to a lack of uniformity in category endorsement and was not considered to be detrimental to rating scale functioning (see Figure, Supplemental Digital Content 1, showing the category probability curves for the 29-item and 21-item versions of the LEQ-CI). Following rating scale refinement, a further Rasch analysis was undertaken to establish whether the 21-item version of the LEQ-CI met the Rasch model criteria for successful interval-level measurement. The results of this analysis are reported below.

Assessment of item and person fit to the model

High, positive person fit residuals (MNSQ values > +2.0) are indicative of an abnormal, random response pattern. Forty-six persons (13.9%) with MNSQ-values > +2.0 were identified as underfitting in the model and these persons were temporarily removed from the dataset (Bond et al 2020) to ascertain their contribution to distortion of model fit. Rasch analyses were conducted for the full data set (n = 330) and with underfitting persons removed (n = 284). No notable differences between the analyses were identified; therefore, the results of the Rasch analysis conducted with the full sample were reported (see Table, Supplementary Digital Content 2, which presents the Rasch analyses for the full sample (n=330) and with underfitting persons removed (n = 284)). All 21 items showed appropriate fit to the model, with item fit statistics (mean-square values) falling within the 0.5 - 1.5 logits range for productive measurement (see Table 4) (Linacre 2002b). The empirical range was 0.68 - 1.52. Empirical item characteristic curves (ICCs) showed good fit to the theoretical

ICCs with data points falling within the 95% confidence bands (see Figure, Supplemental Digital Content 3, which shows the ICCs for the LEQ-CI[21]).

Assessment of unidimensionality and local independence

The LEQ-CI was found to meet the Rasch model requirements of unidimensionality and local independence. The results of the principal component analysis of the residuals (PCAR) supported the existence of only one Rasch dimension. The observed variance was 56.5%, and the unexplained variance of the first residual contrast was 5.1% (Eigenvalue = 2.47). Disattenuated correlations of item clusters (i.e., a group of items based on a cluster analysis of the PCAR loadings) were above the recommended minimum of 0.7 (empirical range 0.89 – 0.98), and suggested the clusters measure a single construct. A second PCAR was performed using Winsteps-generated Rasch simulated data that was equivalent to the empirical data file (i.e., in terms of person measures, item difficulties, and rating scale structure). The simulated data were used to verify the empirical data, establishing whether the observed eigenvalue of 2.47 was due to a second dimension or the result of an artefact (noise) in the data structure. The unexplained variance for the Rasch-modelled data was 3.3% (Eigenvalue = 1.34), which corresponded closely to the empirical data and provided further evidence of unidimensionality. None of the items were overfitting and standardised residual correlation values were less than 0.33 for the empirical dataset, indicative of local independence.

Assessment of differential item functioning (DIF)

DIF analyses were undertaken for gender and age (see Tables, Supplemental Digital Content 4, which presents the results of the DIF analyses for the LEQ-CI[21]). No evidence of DIF was identified. Although two items were found to have p-values indicating statistical significance ($p < 0.05$), the DIF contrast values (i.e. effect sizes) did not exceed the 0.64 threshold for DIF in terms of effect size (Linacre 2020b).

Targeting of the scale

The person-item distribution map showed good targeting of items to persons (see Figure 2). The mean person ability was -0.32, suggesting persons had a slightly lower requirement for effort relative to the mean item difficulty (which is always “0”). The full sample ability ranged from -4.58 to 3.84 logits, with non-extreme person locations ($n = 327$) ranging from -3.54 to 2.49 logits. To illustrate the meaning of the scale as a continuum of perceived listening effort, Item 21 (e103) was located at a logit of 1.05 on the scale. At this location, Item 21 provides content coverage for a person with moderately high level of perceived listening effort (see Figure 2 and Table 5). Conversely. Item 3 (nq4) was located at a logit of -0.81, providing content coverage for persons with a relatively low level of perceived listening effort (see Figure 2 and Table 5). The person with the lowest level of perceived listening effort was located at -4.58 logits, whilst the person with the highest level of perceived listening effort was located at 3.84 logits. Overall, the items were found to provide good coverage for the majority of the sample (item measure range -0.81 to 1.05); however, a lack of items at the extreme ends of the effort continuum suggested that precision may be reduced for the measurement of persons with extremely high or extremely low levels of perceived listening effort.

Assessment of reliability

The LEQ-CI^{21} person separation reliability was 0.91 and the person separation index was 3.28, suggesting excellent reliability and an excellent level of separation distinguishing approximately four levels of person ability defined as “low”; “below average”; “above average”; and “high” levels of perceived listening effort (Duncan et al. 2003; Wright & Masters 1982). The item separation index was 9.69, indicating the sample size was sufficient to confirm the ordering of items on the listening effort continuum (i.e., construct validity).

Raw score to Rasch score conversion

The raw score for the LEQ-CI[21] was calculated by summing the item responses for all 21-items. Rasch scores were calculated in Winsteps and rescaled to 0-100 values as an index of perceived listening effort (see Table, Supplemental Digital Content 5 which presents the conversion table).

DISCUSSION

The present study is an essential step in developing a clinically useful and validated measure of perceived listening effort in daily life. The LEQ-CI[21] is the first hearing-related PROM to be developed specifically for the measurement of perceived listening effort in adults with hearing loss, and one of only a few hearing-related PROMs to be developed and validated using modern psychometric techniques. The results of this initial validation study using Rasch analysis showed that the 21-item version of the LEQ-CI satisfies the Rasch model requirements for interval-level measurement and, therefore, has the potential to be used as a clinical tool for monitoring individual changes in perceived listening effort and as a research tool for group comparisons using parametric statistical tests (Browne & Cano 2019).

In the context of listening effort assessment, using Rasch analysis to develop the LEQ-CI offered several advantages. Firstly, the Rasch model creates a “fixed ruler” that represents an effort continuum. The LEQ-CI’s raw scores, when expressed as a summed total in logits or converted to a 0-100 scale, may be used as an index of a person’s perceived level of listening effort (Wright & Masters 1982). This transformation renders the meaning of LEQ-CI’s total scores clearer and more easily interpretable. Secondly, performance outcome measures of listening effort, such as tasks carried out under controlled conditions (e.g., dual task paradigm), could be co-calibrated with the LEQ-CI[21] using the same interval scale, thus enabling the performance-based measure to be interpreted using the content of the LEQ-CI (Regnault et al. 2020). Co-calibration is of particular relevance to the study of listening effort as the relationships among reported findings from behavioural, physiological and self-report measures are, as yet, not well understood (Alhanbali et al. 2019; Strand et al. 2018).

Using Rasch analysis to refine and validate the LEQ-CI also facilitates its use across different cultural contexts, meaning it could potentially be a suitable tool for international research (Riff et al. 2017; Tennant et al. 2004). If translated into other languages or used within different cultural contexts, evaluation of measurement invariance through DIF analysis could enable adjustments to the LEQ-CI[21] (i.e., removal of items with DIF) to create a measure that is invariant, thus ensuring the equivalence of any translated versions as well as the cultural appropriateness of included items and their response scales. A culturally and linguistic invariant measure supports the pooling of data, enabling CI research to be undertaken with large datasets, a phenomenon that is, as yet, relatively uncommon in the CI literature (Boisvert et al. 2020).

Lastly, the Rasch model would, in turn, support the development of short forms and multiple versions of an instrument, due to psychometric analysis occurring at item rather than at scale level. One limitation of CTT is its focus on an instrument's total score such that evidence of validation relates to the instrument in its entirety, meaning it cannot be applied to individual items (McKenna et al. 2019). Information on item fit; item difficulty; and item discrimination (derived through application of the Rasch model) can enable a small set of items to be selected that will optimise precision, whilst ensuring full coverage of persons across the Rasch logit scale. The study findings, therefore, could support the use of the LEQ-CI[21] in the development of individualised measures of perceived listening effort, such as computer adaptive tests (CAT) where the aim is to minimise respondent burden without sacrificing measurement accuracy (Kane et al. 2020; Petersen et al. 2018).

Evaluation of the LEQ-CI using Rasch analysis offers a theoretical contribution to the conceptualisation of perceived listening effort in daily life by providing evidence in support of the Framework for Understanding Effortful Listening (FUEL, Pichora-Fuller, 2016). Building upon the existing body of work on mental effort and motivation, the FUEL proposes motivational factors to have a modulating influence on effort deployment (Pichora-Fuller et al. 2016). The finding that all of the domains represented in the LEQ-CI's conceptual framework (including motivation) are components of a single dimension, defined as perceived listening effort in daily life, contributes new evidence to the growing body of literature on the involvement of motivational factors when listening is effortful (Koelewijn et al. 2018; Picou & Ricketts 2014; Zekveld et al. 2019). More specifically, the LEQ-CI[21]'s assessment of constructs such as pleasure (Matthen 2016); anxiety (Monzani et al. 2008); social connectedness (Lee & Robbins 1995); and effort-reward imbalance (Siegrist 1996) provides evidence in support of the contribution of social, emotional and psychological constructs to perceived listening effort (Bruya & Tang 2018; Pichora-Fuller 2016).

Limitations

This study has begun to provide insights into the theoretical understanding of perceived listening effort and its measurement. However, there are some limitations to the study which must be considered when evaluating its potential contribution. Firstly, refinement of the LEQ-CI[29], including item removal and the collapsing of rating scale categories, was undertaken *a posteriori* based on empirical data, with fit to the Rasch model confirmed for the refined 21-item instrument using the same dataset. To partly address this limitation, simulated Rasch-modelled data based on the empirical dataset were used. Further evaluation of the LEQ-CI[21] is now needed to verify fit to the Rasch model in a new sample of adult cochlear implant patients and, in particular, CI candidates.

In addition, a change to the UK candidacy criteria for cochlear implantation was implemented during data collection phase of the study, meaning patients who met the updated criteria were not included in the study sample. Further research is needed to confirm the fit of the LEQ-CI[21] to the Rasch model in a sample of adults meeting the revised UK CI candidacy criteria, with particular attention given to whether the instrument is invariant in this group when compared with the existing UK CI population.

Lastly, although targeting of items and person was good overall, incomplete item coverage for persons at the extreme ends of the listening effort continuum represents another limitation of the LEQ-CI[21]. Further work is now therefore needed to develop a better understanding of how effort is perceived when an individual is listening at the limits of their ability, under highly taxing acoustic conditions, or just prior to quitting a listening task assessed as too arduous. Future work could also include a qualitative exploration of the experiences of perceived listening effort from the perspective of adults with mild-moderate hearing loss. These findings could then be used for purposes of developing additional items suitable for measuring persons with an extremely low or an extremely high level of perceived listening effort. Further studies to explore the invariance of the LEQ-CI[21] across the full range of hearing loss and different hearing devices is also needed.

CONCLUSIONS

The LEQ-CI[21] is the first PROM developed specifically to measure perceived listening effort in daily life for adult cochlear implant candidates and recipients. It is one of only a few hearing-specific PROMs to be developed using modern psychometric methods and is the product of a body of work undertaken in line with international consensus-based standards on PROM development. The current study began the task of establishing the LEQ-CI[21] as a reliable and valid, interval-level measure of perceived listening effort in daily life suitable for use in research and in clinical practice. The use of Rasch analysis helped to engage the LEQ-CI[21] with theoretical frameworks in the published literature. Further validation work is planned that will add to this body of evidence through assessment of the LEQ-CI[21]'s construct validity, reliability, and responsiveness using traditional psychometric approaches, as well as to study its implementation in clinical practice.(7218 words)

ACKNOWLEDGMENTS

The authors would like to thank the Agnes Allen, Scottish Cochlear Implant Programme; Wanda Alesky, UCL Auditory Implant Programme; Louise Craddock, Midlands Auditory Implant Programme; Unai Martinez de Estabariz, Manchester Auditory Implant Programme, and the South Wales Cochlear Implant Programme (Bridgend) for their support with participant recruitment and data collection. The authors are grateful to Laura Boddy, University Hospitals Birmingham NHS Trust, for her help with participant identification and to Tim Pickles, Centre for Trials Research, Cardiff University, for support with Rasch analysis. Finally, the authors thank the participants for taking the time to be part of this research.

Portions of this article were presented at the British Cochlear Implant Group (BCIG) Academic Meeting, Nottingham, UK, March 2020.

All authors contributed to the design of the study. SEH collected and analysed the data in the UK and drafted the manuscript. All authors discussed the results and implications, provided critical revisions and approved the final version of the manuscript.

For further information about the LEQ-CI please contact: PROMS@swansea.ac.uk.

REFERENCES

Alhanbali, S., Dawes, P., Millman, R.E., et al. (2018). Measures of listening effort are multidimensional. *Ear Hear*, 40(5), 1084-1097.

Amieva, H., Ouvrard, C., Meillon, C., et al. (2018). Death, depression, disability, and dementia associated with self-reported hearing problems: A 25-year study. *J. Gerontol*, 73, 1383–1389.

Amin, N., Wong, G., Nunn, T., et al. (2020). The Outcomes of Cochlear Implantation in Elderly Patients: A Single United Kingdom Center Experience. *Ear Nose Throat J*, 0145561320910662.

Aryadoust, V., Tan, H.A.H., Ng, L.Y. (2019). A scientometric review of Rasch measurement: The rise and progress of a specialty. *Front Psychol*, 10:2197.

Boisvert, I., Reis, M., Au, A., et al. (2020). Cochlear implantation outcomes in adults: A scoping review. *PLoS One*, 15, e0232421.

Bond, T.G., Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* 3rd edition. London: Routledge.

Bond, T.G., Yan, Z., Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* 4th edition. London: Routledge.

Boone, W.J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sci Educ*, 15:rm4.

Boone, W.J., Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Educ*, 4, 1–13.

Boone, W.J., Yale, M.S., Staver, J.R. (2014). *Rasch Analysis in the Human Sciences.* London: Springer.

Bräcker, T., Hellmiss, S., Batsoulis, C., et al. (2019). Introducing real-life listening features into the clinical test environment: Part II: Measuring the hearing performance and evaluating the listening effort of individuals with a hearing implant. *Cochlear Implants Int*, 20, 165–175.

Browne, J.P., Cano, S.J. (2019). A Rasch measurement theory approach to improve the interpretation of patient-reported outcomes. *Med Care*, 57, S18-S23.

Bruya, B., Tang, Y.Y. (2018). Is attention really effort? Revisiting Daniel Kahneman’s influential 1973 book *Attention and Effort*. *Front Psychol*, 9, 1133.

Cano, S.J., Hobart, J.C. (2011). The problem with health measurement. *Patient Prefer Adherence*, 5, 279–290.

Christensen, K.B., Makransky, G., Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas*, 41, 178–194.

Devlin, N.J., Appleby, J. (2010). *Getting the Most Out of Proms: Putting Health Outcomes at the Heart of NHS Decision-making*, London: The King's Fund.

DeWalt, D.A., Rothrock, N., Yount, S., et al. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Med Care*, 45, 12–21.

Duncan, P.W., Bode, R.K., Lai, S.M., et al. (2003). Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. *Arch Phys Med Rehabil*, 84, 950–963.

Edwards, B. (2016). A model of auditory-cognitive processing and relevance to clinical applicability. *Ear Hear*, 37, 85S–91S.

Fan, J., Bond, T.G. (2019). Applying Rasch Measurement in Language Assessment: Unidimensionality and Local Independence. In V. Aryadoust and M. Raquel, eds. *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*. Abdingdon: Routledge.

FDA (2020). *Principles for Selecting, Developing, Modifying, and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation: Draft Guidance for Industry and Food and Drug Administration Staff, And Other Stakeholders*, Rockville, MD: U.S. Food and Drug Administration.

FDA (2009). Guidance for Industry Use in Medical Product Development to Support Labelling Claims Guidance for Industry. *Clin Fed Regist*, 1–39.

Field, J., Holmes, M.M., Newell, D. (2019). PROMs data: can it be used to make decisions for individual patients? A narrative review. *Patient Relat Outcome Meas*, 10, 233–241.

Flesch, R. (1948). A new readability yardstick. *J Appl Psychol*, 32, 221–233.

Gagné, J., Besser, J., Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends Hear*, 21, 1–25.

Granberg, S., Dahlström, J., Möller, C., et al. (2014). The ICF Core Sets for hearing loss - researcher perspective. Part I: Systematic review of outcome measures identified in audiological research. *Int J Audiol*, 53, 65–76.

Hagquist, C., Bruce, M., Gustavsson, J.P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *Int J Nurs Stud*, 46, 380–393.

Hobart, J., Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technol Assess*, 13, 1-177.

Holman, J.A., Drummond, A., Hughes, S.E., et al. (2019). Hearing impairment and daily-life fatigue: A qualitative study. *Int J Audiol*, 58, 408–416.

Hornsby, B.W.Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear Hear*, 34, 523–534.

Hornsby, B.W.Y., Kipp, A.M. (2016). Subjective ratings of fatigue and vigor in adults with hearing loss are driven by perceived hearing difficulties not degree of hearing loss. *Ear Hear.*, 37, e1-10.

Hornsby, B.W.Y., Naylor, G., Bess, F.H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear Hear*, 37, 136S-144S.

Hughes, S.E., Boisvert, I., Rapport, F., et al. (2019a). Measuring the experience of listening effort in cochlear implant candidates and recipients: Development and pretesting of the LEQ-CI. In *Poster presentation*. Southampton: BCIG Academic Meeting 2019.

Hughes, S.E., Hutchings, H.A., Dobbs, T.D., et al. (2017a). A systematic review and narrative synthesis of the measurement properties of patient-reported outcome measures (PROMs) used to assess listening effort in hearing loss. In *British Society of Audiology*. Harrogate.

Hughes, S.E., Hutchings, H.A., Rapport, F., et al. (2017b). Qualitative data supporting the FUEL: Perceived listening effort in cochlear implantation. In *Cognitive Hearing Science for Communication*. Linköping, Sweden.

Hughes, S.E., Hutchings, H.A., Rapport, F.L., et al. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A Grounded Theory to establish content validity for a new patient reported outcome measure. *Ear Hear*, 39, 922-934.

Hughes, S.E., Rapport, F., Watkins, A., et al. (2019b). Study protocol for the validation of a new patient-reported outcome measure (PROM) of listening effort in cochlear implantation: The Listening Effort Questionnaire-Cochlear Implant (LEQ-CI). *BMJ Open*, 9, 1–8.

Hughes, S.E., Rapport, F.L., Boisvert, I., et al. (2017c). Patient-reported outcome measures (PROMs) for assessing perceived listening effort in hearing loss: Protocol for a systematic review. *BMJ Open*, 7:e014995.

Johnson, J., Xu, J., Cox, R., et al. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *Am J Audiol*, 24, 419–431.

Kane, L.T., Namdari, S., Plummer, O.R., et al. (2020). Use of computerized adaptive testing to develop more concise patient-reported outcome measures. *JB JS Open Access*, 5, e0052.

Koelewijn, T., Zekveld, A.A., Lunner, T., et al. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hear Res*, 367, 106–112.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In J.D. Wright and P.V. Marsden, eds. *Handbook of Survey Research*, 2nd edition. San Diego, CA: Elsevier.

Lee, R.M., Robbins, S.B. (1995). Measuring belongingness: The Social Connectedness and the Social Assurance scales. *J Couns Psychol*, 42, 232–241.

Linacre, J.M. (1994). Sample size and item calibration or person measure stability. *Rasch Meas Trans*, 7, 328.

Linacre, J. (2002a). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *J Appl Meas*, 3, 85–106.

Linacre, J.M. (2002b). What do infit and outfit, mean-square and standardized mean. *Rasch Meas Trans*, 16, 878.

Linacre, J.M. (2018). Detecting multidimensionality in Rasch data using Winsteps Table 23. Available at: https://youtu.be/sna19QemE50 [Accessed November 29, 2020].

Linacre, J.M. (2020a). *A User's Guide to WINSTEPS MINISTEPS Rasch-Model Computer Programs*, Available at: http://homes.jcu.edu.au/$\sim$edtgb/%5Cnpapers3://publication/uuid/D56B724A-62FF-4D00-84E1-ECC888298B70.

Linacre, J.M. (2020b). Table 30.1 Differential item functioning DIR pairwise. Available at: https://www.winsteps.com/winman/table30_1.htm [Accessed November 29, 2020]..

Lord, F.M., Novick, M.R., Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores.* Oxford: Addison-Wesley.

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

Matthen, M. (2016). Effort and displeasure in people who are hard of hearing. *Ear Hear*, 37, 28S–34S.

McGarrigle, R., Munro, K.J., Dawes, P., et al. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper". *Int J Audiol*, 53, 433–40.

McLaughlin, G.H. (1969). SMOG grading: A new readability formula. *J Read*, 12, 639–646.

McKenna, S.P., Heaney, A., Wilburn, J. et al. (2019). Measurement of patient-reported outcomes. 1: The search for the Holy Grail. *J Med Econ*, 22, 516-522.

McRackan, T.R., Velozo, C.A., Holcomb, M.A., et al. (2017). Use of adult patient focus groups to develop the initial item bank for a cochlear implant quality-of-life instrument. *JAMA Otolaryngol Head Neck Surg*, 143, 975–982.

Miles, K., McMahon, C., Boisvert, I., et al. (2017). Objective assessment of listening effort: coregistration of pupillometry and EEG. *Trends Hear*, 21, 1-13.

Mokkink, L.B., Prinsen, C.A.C., Patrick, D.L., et al. (2018). *COSMIN methodology for systematic reviews of Patient - Reported Outcome Measures (PROMs),*

Monzani, D., Galeazzi, G., Genovese, E., et al. (2008). Psychological profile and social behaviour of working adults with mild or moderate hearing loss. *Acta Otorhinolaryngol Ital*, 28, 61–66.

Nachtegaal, J., Kuik, D.J., Anema, J.R., et al. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *Int J Audiol*, 48, 684–91.

National Institute for Health and Clinical Excellence (NICE) (2009). *NICE technology appraisal guidance 166: Cochlear implants for children and adults with severe to profound deafness*. Available at: guidance.nice.org.uk/TA166/pdf/English [Accessed March 24, 2020].

NHS Digital Service Manual (2019). Use a readability tool to prioritise content - NHS digital service manual. Available at: https://service-manual.nhs.uk [Accessed July 22, 2020].

Ohlenforst, B., Zekveld, A. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. *Ear Hear,* 38, 267-281.

Pallant, J.F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol,* 46, 1–18.

Patrick, D.L., Burke, L.B., Gwaltney, C.J., et al. (2011a). Content validity - Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2 - Assessing respondent understanding. *Value Health*, 14, 978–988.

Patrick, D.L., Burke, L.B., Gwaltney, C.J., et al. (2011b). Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1–eliciting concepts for a new PRO instrument. *Value Health*, 14, 967–77.

Petersen, M.A., Aaronson, N.K., Arraras, J.I., et al. (2018). The EORTC CAT Core—The computer adaptive version of the EORTC QLQ-C30 questionnaire. *Eur J Cancer*, 100, 8–16.

Pichora-Fuller, M.K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear Hear*, 37, 92S–100S.

Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., et al. (2016). Hearing impairment and cognitive energy. *Ear Hear*, 37, 5S–27S.

Picou, E.M., Ricketts, T. a (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *Int J Audiol*, 53, 1–9.

Ramage-Morin, P.L. (2016). *Hearing Difficulties and Feelings of Social Isolation Among Canadians Aged 45 or Older*, Available at: http://www.statcan.gc.ca/pub/82-003-x/2016011/article/14671-eng.pdf [Accessed April 12, 2020].

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielson and Lydiche.

Regnault, A., Meunier, J., Ciesluk, A., et al. (2020). Providing meaningful interpretation of a performance outcome measure by co-calibration with a patient-reported outcome: illustration with widely used multiple sclerosis measures. In Virtual ISOQOL 2020.

Riff, K.W.Y.W., Tsangaris, E., Goodacre, T., et al. (2017). International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). *BMJ Open*, 7, e015467.

Rowen, D., Carlton, J., Elliott, J. (2019). PROM validation using paper-based or online surveys: Data collection methods affect the sociodemographic and health profile of the sample. *Value Health*, 22, 845–850.

Siegrist, J. (1996). Adverse health effects of high-effort / low-reward conditions. *J Occup Health Psychol*, 1, 27–41.

Smith, A.B., Wright, P., Selby, P.J., et al. (2007). A Rasch and factor analysis of the Functional Assessment of Cancer Therapy-General (FACT-G). *Health Qual. Life Outcomes*, 5, 1–10.

Smith, M.G., Witte, M., Rocha, S., et al. (2019). Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. *BMC Med Res Methodol*, 19, 230.

Stevens, S.S. (1946). On the Theory of Scales of Measurement. *Science*, 103, 677–680.

Strand, J.F., Brown, V.A., Merchant, M.B., et al. (2018). Measuring listening effort: convergent validity, sensitivity, and links with cognitive and personality measures. *J Speech Lang Hear Res*, 61, 1463–1486.

Strand, J.F., Dillman-Hasso, N., Villanueva, J., et al. (2020). Understanding speech amid the jingle and jangle: Recommendations for improving measurement practices in listening effort research. *Pre-Print*. Available at: https://doi.org/10.31234/osf.io/3e7mf [Accessed November 29, 2020].

Streiner, D.L., Norman, G.R., Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to Their Development and Use*, 5th Edition. Oxford: Oxford University Press.

Tennant, A., & Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum*, 57, 1358-1362.

Tennant, A., Penta, M., Tesio, L., et al. (2004). Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales Through Differential Item

Functioning Within the Framework of the Rasch Model: The PRO-ESOR Project. *Med Care*, 42, I37-I38.

UK Data Service (2020). Consent for Data Sharing. Available at: https://ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/surveys.aspx [Accessed November 5, 2020].

Waterbury, G.T. (2019). Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation. *J Appl Meas*, 20, 154–166.

Wright, B.D. (1996). Comparing Rasch measurement and factor analysis. *Struct Equ Model*, 3, 3–24.

Wright, B.D., Masters, N.G. (1982). *Rating Scale Analysis*, Chicago: MESA Press.

Wright, B.D., Stone, M.H. (1979). *Best Test Design*, Chicago: MESA Press.

Wright, Linacre (1994). Reasonable mean-square fit values. *Rasch Meas Trans*, 8, 370.

Zekveld, A.A., Koelewijn, T., Kramer, S.E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends Hear*, 22, 1-25.

Zekveld, A.A., van Scheepen, J.A.M., Versfeld, N.J., et al. (2019). Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. *Hear Res*, 381, 107768.

FIGURE LEGEND

Figure 1. The conceptual framework for measurement of perceived listening effort in daily life for the adult cochlear implant population. This diagrammatic representation shows the domains and concepts, as well as the expected relationships between concepts (i.e., items), measured by the LEQ-CI[21]

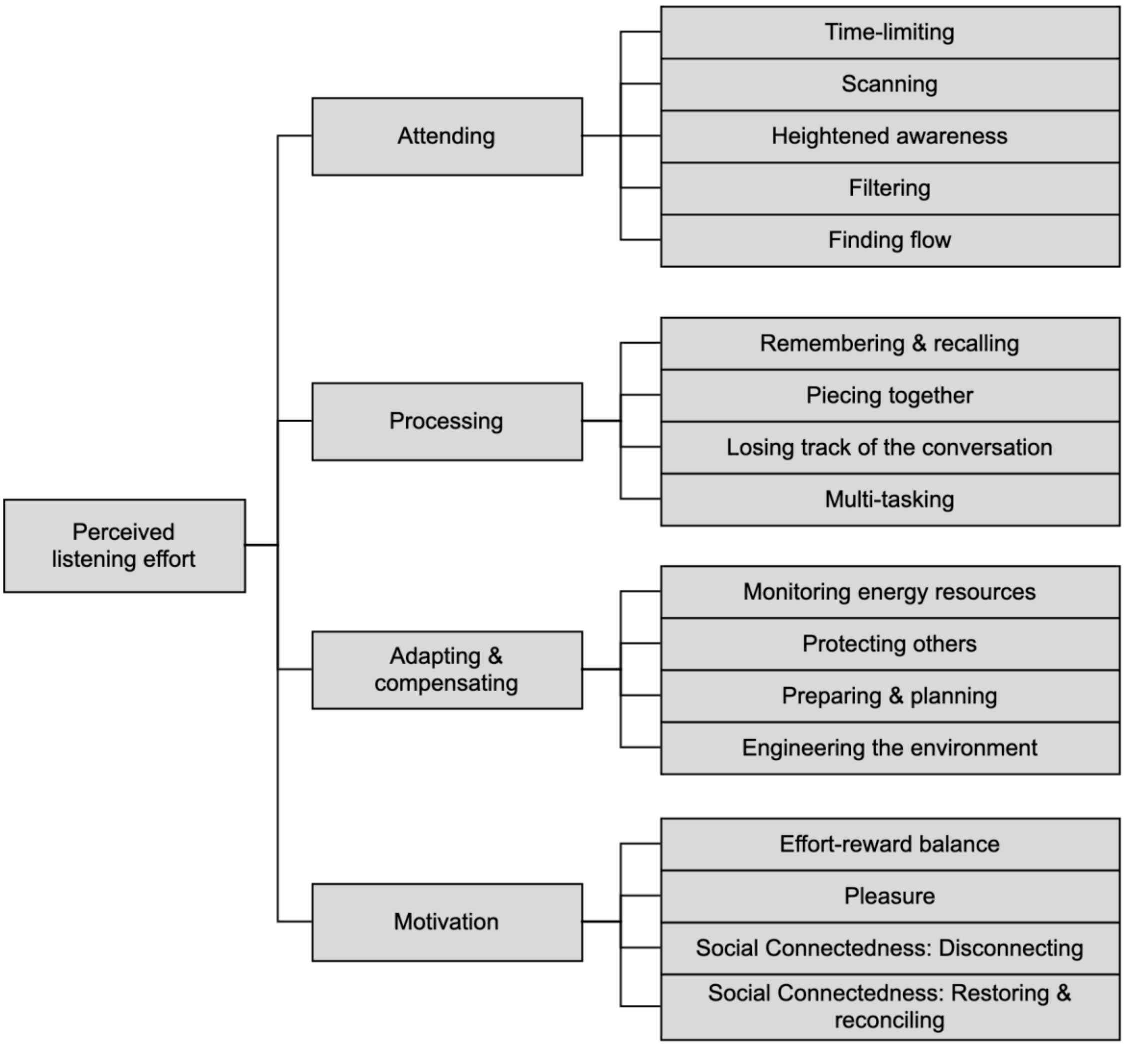


Figure 2. The person-item map for the LEQ-CI[21] (n = 330) showing the distribution of persons (upper half) and items (lower half). The x-axis displays the perceived listening effort continuum on a logit scale and the y-axis displays frequency counts.

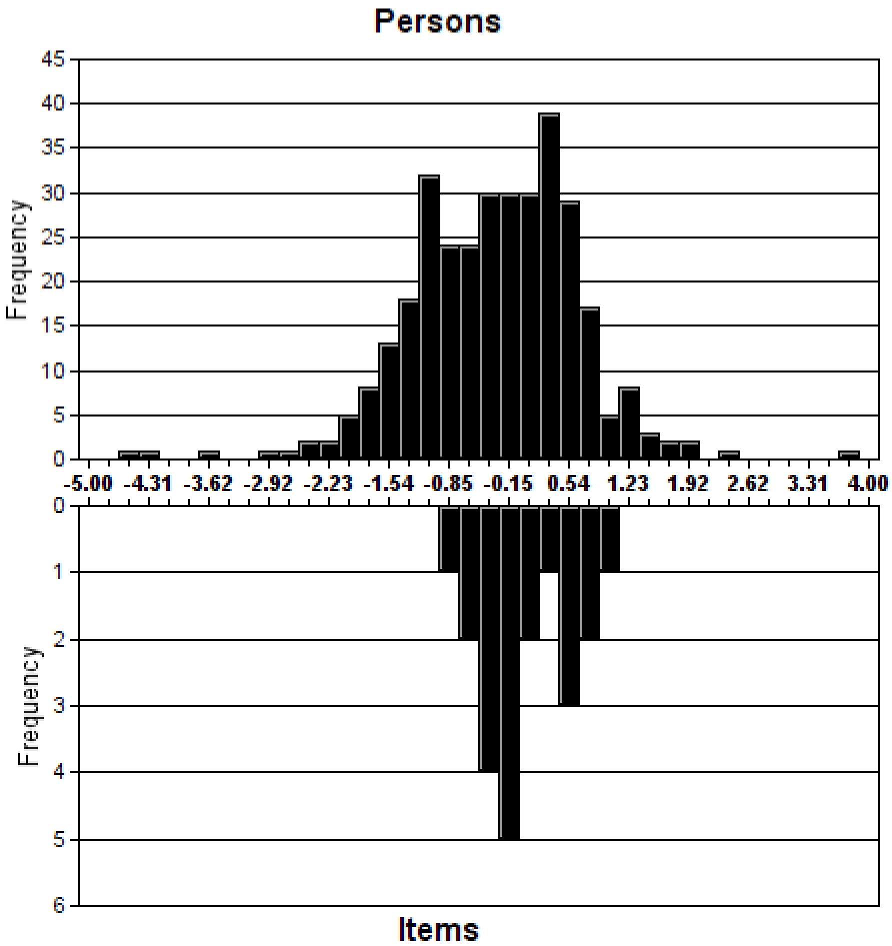


Table 1: Item stems per domain for the LEQ-CI showing all 29 items included in the draft instrument. An asterisk (*) denotes an item removed due to misfitting the Rasch model. Remaining items comprise the LEQ-CI[21].

Item No.	Item Label	Domain	Item stem
1.	e398	Attending	Have to stay alert
2.	e319	Attending	Strain to hear sounds around you
3.	nq4	Attending	Strain to hear speech
4.	e320	Attending	Figure out which sound to focus on
5.	e31	Attending	Listen as long as needed*
6.	n2-	Processing	Anticipate what someone says*
7.	e35	Adapting & compensating	Ask others to repeat
8.	n14	Processing	Figure out what said - odd word or phrase*
9.	e50	Processing	Forget what other person just said
10.	n16	Processing	Take longer to understand
11.	e323	Processing	Able to listen to someone talk while doing something else
12.	e71	Processing	Listen and plan reply*
13.	e365	Attending	Listening just seems to happen*
14.	e310	Motivation	Stop listening because too much effort
15.	e34	Processing	Understand group talking - no background noise*
16.	nq3	Processing	Understand group talking - background noise
17.	e121	Adapting & compensating	Do things to make listener easier*
18.	n11	Adapting & compensating	Make others feel at ease
19.	e344	Motivation	Effort bother you
20.	e213	Motivation	Avoid situations because of effort needed to listen
21.	e103	Motivation	Give up on listening
22.	e335	Adapting & compensating	Plan day around effort*
23.	e105	Adapting & compensating	Run out of energy for listening
24.	e214	Adapting & compensating	Can be yourself
25.	e118	Motivation	Feel tense
26.	e112	Motivation	Choose to be alone because of listening effort

Item No.	Item Label	Domain	Item stem
27.	e326	Motivation	Stop doing things want to do
28.	e363	Motivation	Find pleasure in listening
29.	e329	Motivation	Feel connected with others

*Item removed as misfitting the Rasch model

Table 2: Participants' demographic characteristics

	Total (N = 330)
Patient age (years)	
Mean (SD)	62.50 (15.1)
Median (range)	66 (21 – 89)
Missing	0
Gender	
Male	131 (39.7%)
Female	199 (60.3%)
Prefer to self-describe	0 (0%)
Prefer not to say	0 (0%)
Missing	0 (0%)
Employment status	
Employed, working full time	76 (23.0%)
Employed, working part-time	40 (12.1%)
Not employed	11 (33.3%)
Retired	177 (53.6%)
Full-time homemaker	13 (3.9%)
Volunteer worker	2 (0.6%)
Student	2 (0.6%)
Other	8 (2.4%)
Unknown	1 (0.3%)
Highest level of education	
Not applicable	8 (2.4%)
Primary school	1 (0.3%)
Secondary school	96 (29.1%)
GCSE/O-level qualification	52 (15.8%)
A-level/BTEC or equivalent post-16 education	70 (21.2%)
Undergraduate degree	21 (6.4%)
Graduate degree	50 (15.2%)
Doctorate degree	9 (2.7%)
Other	20 (6.1%)
Unknown	3 (0.9%)
UK Region	
London	71 (21.5%)
Midlands	62 (18.8%)
Northwest	97 (29.4%)
Scotland	91 (27.6%)
Wales*	9 (2.72%)
Age at HL onset	
Mean (SD)	28.6 (20.1)
Median (range)	28.0 (0 – 78.0)

Missing	10 (3.0%)
Age at receipt of first hearing device	
Mean (SD)	35.0 (20.9)
Median (range)	39.5 (0 – 84.0)
Missing	12 (0.04%)
Hearing devices used	
One hearing aid	7 (2.1%)
Two hearing aids	14 (4.2%)
One cochlear implant	193 (58.5%)
One cochlear implant + one hearing aid	106 (32.1%)
Two cochlear implants	7 (2.1%)
None	3 (0.9 %)
Other	0 (0%)
Missing	0 (0%)
Hours of hearing device use	
Mean (SD)	14.1 (2.9)
Median (range)	15.0 (0 – 24.0)
Missing	5 (1.5%)
Years implanted	
Mean (SD)	3.8 (4.8)
Median (Range)	2.0 (0 – 26.0)
Missing	26 (1.0%)

*Fewer patients from the three Welsh CI centres were eligible to participate due to previous involvement as participants in LEQ-CI content validity studies. Nine responses were received from the 11 Welsh patients eligible to participate, a response rate of 81.8%.

Table 3: Summary of the category fit statistics for each of the LEQ-CI[21] rating scales. Category values are the raw category values that correspond to the category label, average measures can be seen to advance montonically, Outfit MnSq values < 2.0 are indicative of model fit.

Item Group	Category Value	Average Measure (logits)	OUTFIT MnSq (logits)	Andrich Thresholds (logits)	Category Label
1[a]	1	-1.25	1.13	None	Never
	3	-0.47	0.87	-1.96	Occasionally
	5	0.14	0.94	-0.05	Frequently
	7	0.57	1.54	2.01	Always
2[b]	1	-0.67	1.12	None	Not at all
	2	-0.27	0.68	-0.77	Slightly
	3	0.22	0.75	-0.09	Moderately
	5	0.74	1.06	0.86	Extremely
3[c]	1	-0.74	0.89	None	Never
	3	0.01	0.44	-0.18	1 - 4 days
	5	0.42	0.97	0.18	5 - 7 days
4[d]	1	-0.99	0.90	None	None of the time
	3	-0.19	0.96	-1.80	Some of the time
	4	0.33	0.95	1.80	Most of the time
5[e]	1	-1.16	0.98	None	Not at all tense
	3	-0.44	0.77	-1.82	Moderately tense
	5	0.36	0.80	1.82	Extremely tense

Items are labelled according to their entry number in Winsteps and are grouped according to their rating scale structure:
[a]Items = Item 1, 4, 7, 9-11, 14, 16, 20, 21, 26, 27
[b]Items = Item 2, 3, 18, 19, 28, 29
[c]Items = 23
[d]Items = Item 24
[e]Items = Item 25

Table 4: Item fit statistics showing final item locations for the LEQ-CI[21] (n = 330)

Item ID	Measure	Infit Statistics		Outfit Statistics	
		MNSQ	ZSTD	MNSQ	ZSTD
e103	1.05	1.03	0.43	0.90	-0.55
e326	0.84	0.97	-0.31	0.86	-0.97
e112	0.64	0.82	-2.32	0.72	-2.32
e50	0.63	1.14	1.65	1.08	0.65
e310	0.60	0.87	-1.72	0.86	-1.18
e213	0.48	1.03	0.39	0.93	-0.66
e105	0.32	0.91	-1.11	0.7	-1.08
e320	0.09	1.18	2.27	1.21	2.18
e329	-0.02	0.73	-3.87	0.81	-2.19
n16	-0.11	0.80	-2.79	0.93	-0.81
e363	-0.14	1.02	0.27	1.10	1.04
e214	-0.20	1.03	0.41	1.02	0.19
e319	-0.23	1.22	2.84	1.25	2.41
e35	-0.25	0.66	-5.12	0.7	-3.33
e323	-0.30	1.43	5.07	1.47	5.01
e398	-0.43	1.48	5.61	1.52	5.51
n11	-0.45	1.27	3.41	1.37	3.28
e344	-0.47	0.72	-4.21	0.68	-3.50
nq3	-0.61	1.38	4.55	1.36	3.94
e118	-0.67	0.91	-1.25	0.88	-1.40
nq4	-0.81	0.70	-4.34	0.83	-1.30
Mean	0.00	1.01	0.0	1.01	0.20
SD	0.52	0.24	3.1	0.25	2.5

The MNSQ acceptable range for productive measurement = 0.5 – 1.5 logits.
MNSQ = mean square
ZSTD = Z-standardised statistics

Table 5: Extreme scores for items and persons from the sample (n = 330).

Extreme Items/Persons	Total Score	Measure	Meaning of Item/Person Location
Item 21 (e103)	n/a	1.05	Low trait level (more likely to be endorsed by persons with moderately high levels of perceived listening effort)
Item 3 (nq4)	n/a	-0.81	High trait level (more likely to be endorsed by persons with low levels of perceived listening effort)
Person 276 10/05/2021 11:26:00	124	3.84	Person with highest level of perceived listening effort
Person 35	72	0.01	Person with average levels of perceived listening effort
Person 203	21	-4.58	Person with the lowest level of perceived listening effort

LIST OF ABBREVIATIONS

CAT	Computer adaptive test
CI	Cochlear implant
CTT	Classical Test Theory
DIF	Differential item functioning
EEG	Electroencephalography
FUEL	Framework for Understanding Effortful Listening
HA	Hearing aid
ICC	Item characteristic curve
IRT	Item response theory
JMLE	Joint Maximum Likelihood Estimation
LEQ-CI	Listening Effort Questionnaire – Cochlear Implant
MMHL	Mild-moderate hearing loss
MNSQ	Mean square
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
PCAR	Principal component analysis of the residuals
PROM	Patient-reported outcome measure
PROMIS®	Patient-Reported Outcomes Measurement Information System
PSI	Person separation index
RMT	Rasch measurement theory
SNHL	Sensorineural hearing loss
UK	United Kingdom
VAS	Visual analogue scales
ZSTD	z-standardised

LIST OF SUPPLEMENTAL DIGITAL CONTENT

Supplemental Digital Content 1. Category Probability Curves for the draft 29-item version and the refined 21-item version of the LEQ-CI. docx

Supplemental Digital Content 2. Table comparing the findings of Rasch analyses undertaken with the full sample (n = 330) and with underfitting persons removed (n = 284). docx

Supplemental Digital Content 3. Item Characteristic Curves for all items of the LEQ-CI^{21}. pptx

Supplemental Digital Content 4. Tables presenting DIF analyses for gender (males v females) and age (adults < 70 years and adults aged 70 years or older). docx

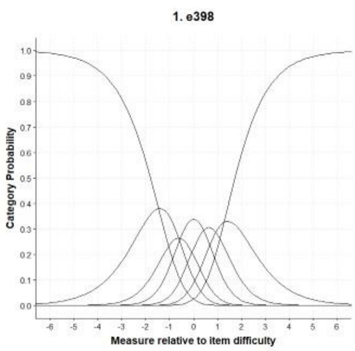
Supplemental Digital Content 5. Raw score to 0-100 Rasch score conversion table for use with the LEQ-CI^{21}. docx

Supplemental Digital Content 1.

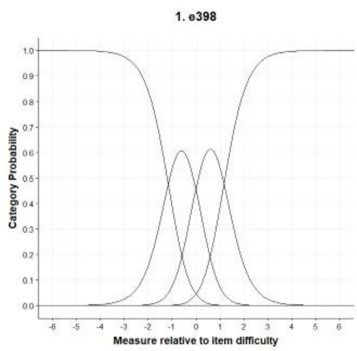
Category Probability Curves for the LEQ -CI[29] and LEQ-CI[21]

Rasch model category probability curves for each item of the original 29-item and the revised 21-item versions of the LEQ-CI. The x-axis represents the attribute in logits. The y-axis represents the probability of a response category being selected. The curves represent the likelihood that a respondent with a particular amount of the latent trait will select a category. The y-axis represents the expected probability of endorsement of any given category when a person responds to the item. The x-axis represents the person ability relative to the item difficulty, with origins set to 0. The scale is measured in logits (log odds units). Distances to the right of zero indicate higher and higher levels of effort. Locations to the left of zero indicate lower and lower levels of effort. Categories should advance in ascending order from left to right along the x-axis. Categories should each have a distinct peak indicating that it is the most probable (modal) category at that point on the latent variable. The crossover points between two curves are the equal-probability points or thresholds. Thresholds advance in line with categories, thresholds that fail to advance monotonically are considered disordered. The initial item category structure shows threshold disordering and a lack of modal peaks. The category probability curves have not been displayed for any items that have been removed from the LEQ-CI due to model misfit

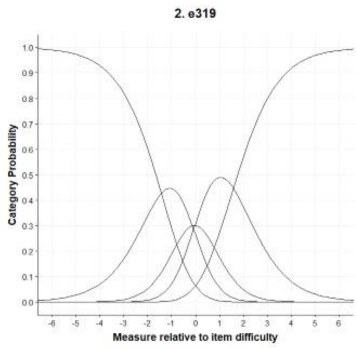
LEQ-CI[29]



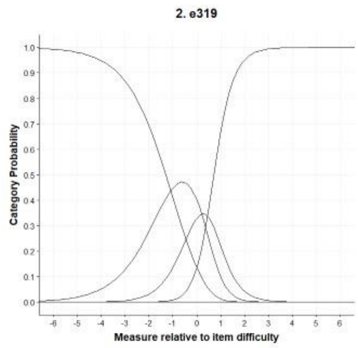
LEQ-CI[21]



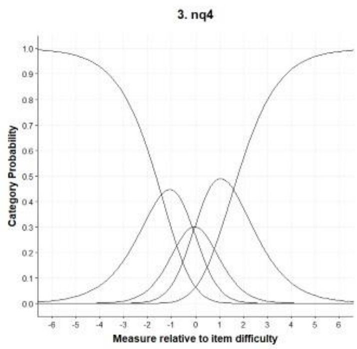
LEQ-CI^{29}



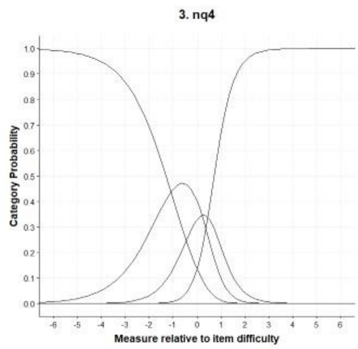
LEQ-CI^{21}



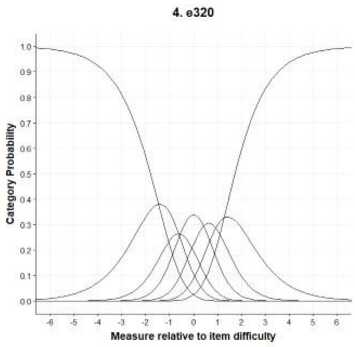
LEQ-CI^{29}



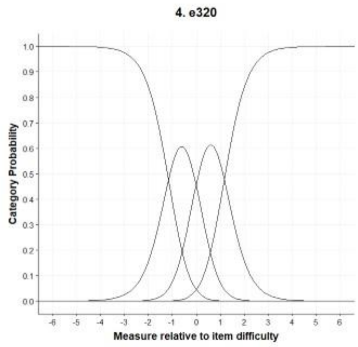
LEQ-CI^{21}



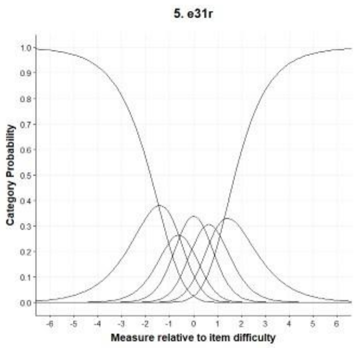
LEQ-CI[29]



LEQ-CI[21]



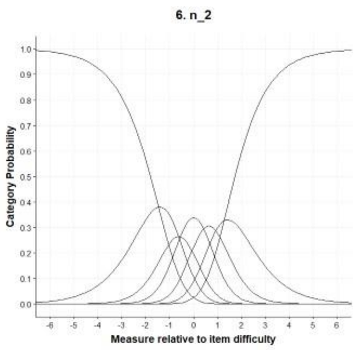
LEQ-CI[29]



LEQ-CI[21]

- Item removed

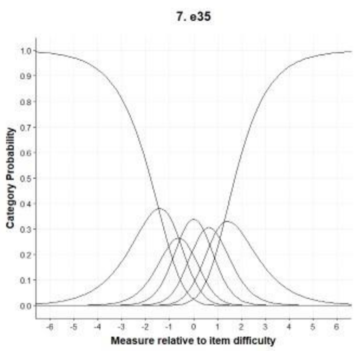
LEQ-CI[29]



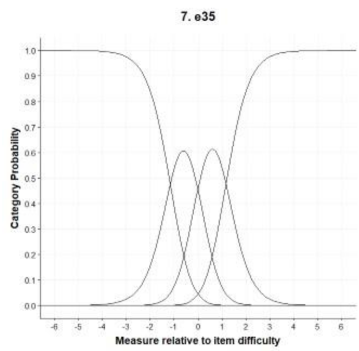
LEQ-CI[21]

- Item removed

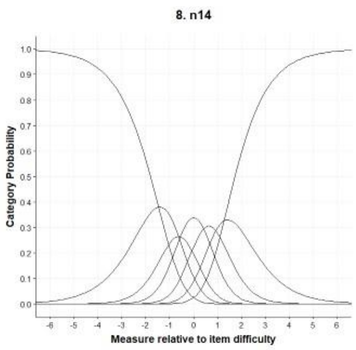
LEQ-CI[29]



LEQ-CI[21]



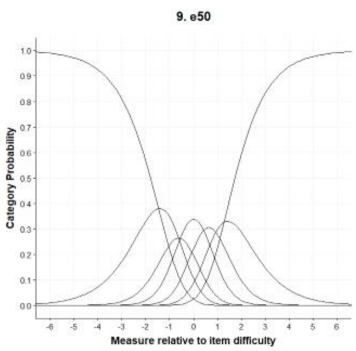
LEQ-CI[29]



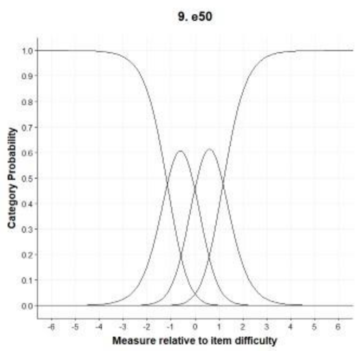
LEQ-CI[21]

- Item removed

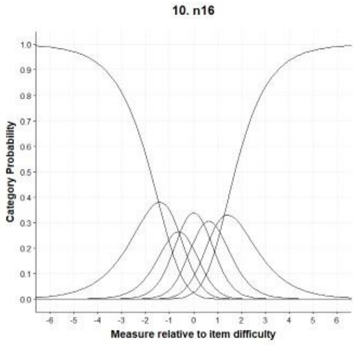
LEQ-CI[29]

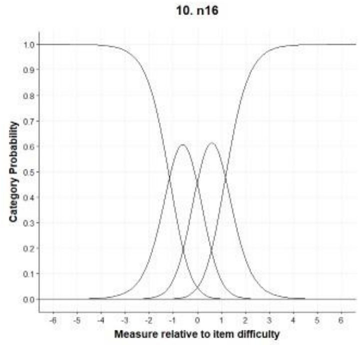


LEQ-CI[21]

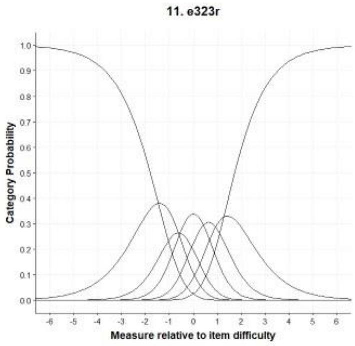


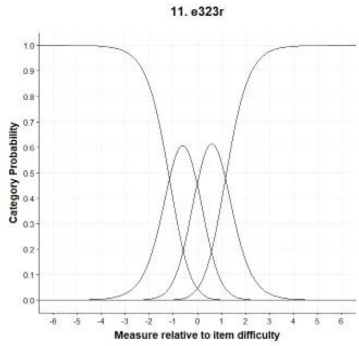
LEQ-CI[29] LEQ-CI[21]



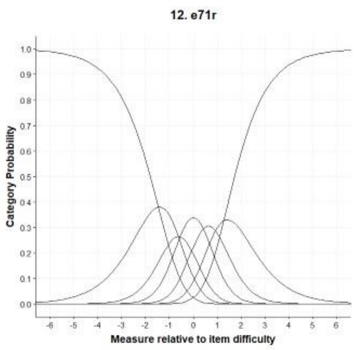


LEQ-CI[29] LEQ-CI[21]

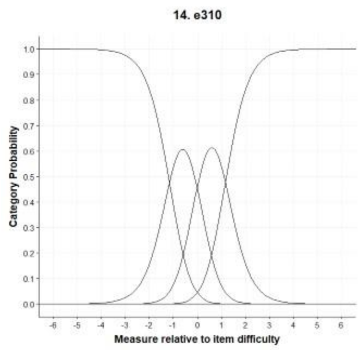




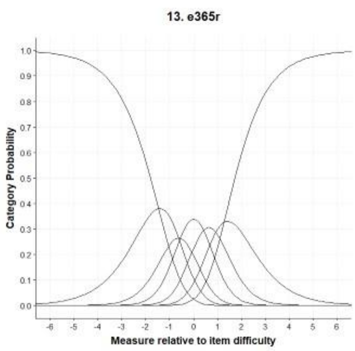
LEQ-CI[29]



LEQ-CI[21]

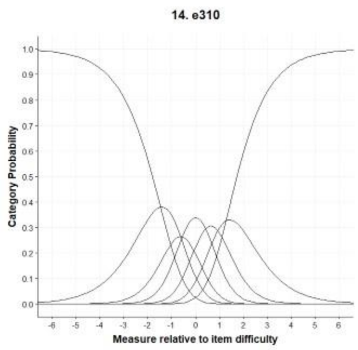


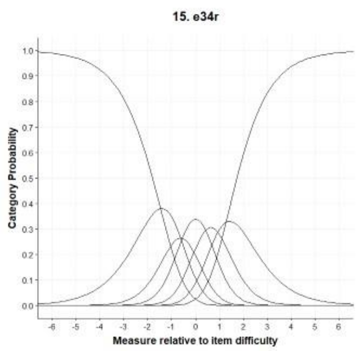
LEQ-CI[29]



LEQ-CI[21]

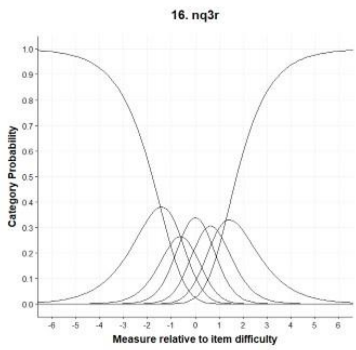
- Item removed

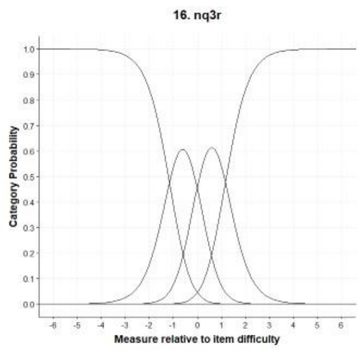
LEQ-CI^{29}	LEQ-CI^{21}
	• Item removed

LEQ-CI^{29}	LEQ-CI^{21}
	• Item removed

LEQ-CI[29]

LEQ-CI[21]

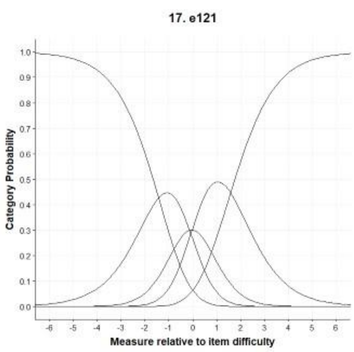




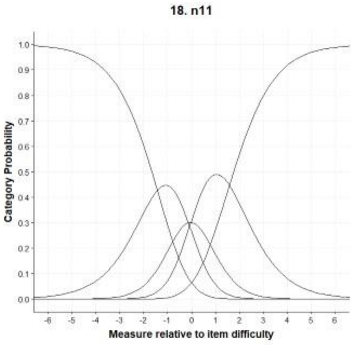
LEQ-CI[29]

LEQ-CI[21]

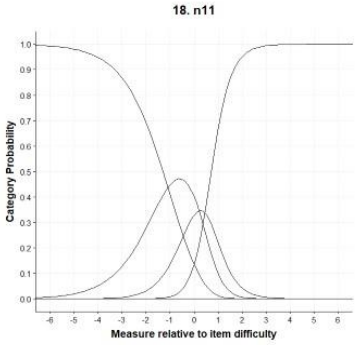
- Item removed



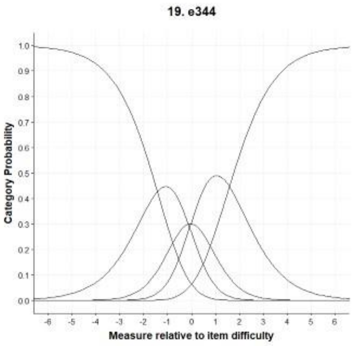
LEQ-CI^{29}



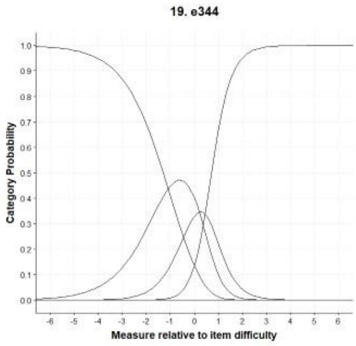
LEQ-CI^{21}



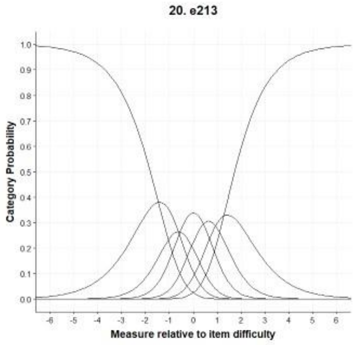
LEQ-CI^{29}



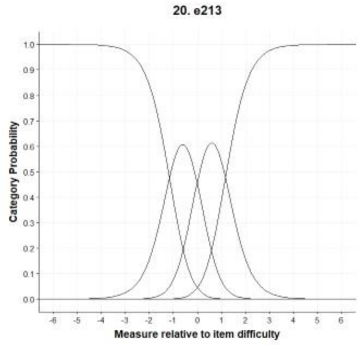
LEQ-CI^{21}



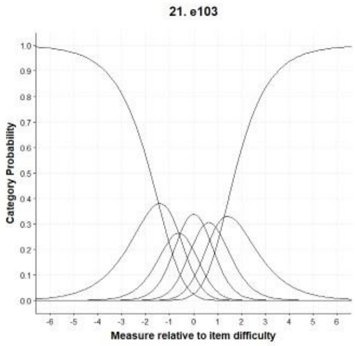
LEQ-CI[29]



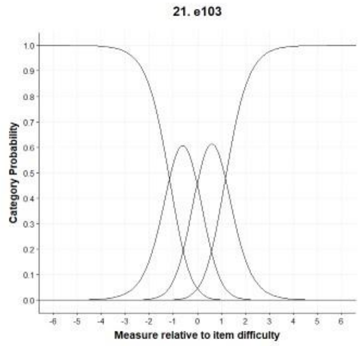
LEQ-CI[21]



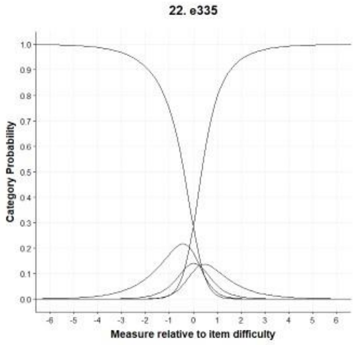
LEQ-CI[29]



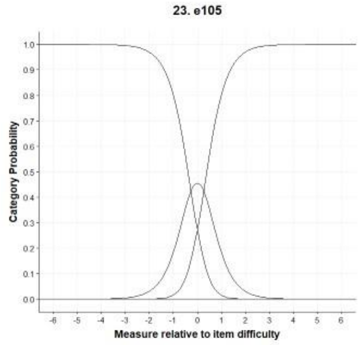
LEQ-CI[21]



LEQ-CI[29]

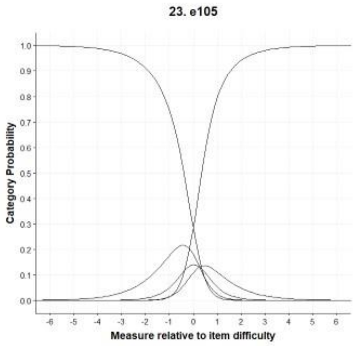


LEQ-CI[21]



44

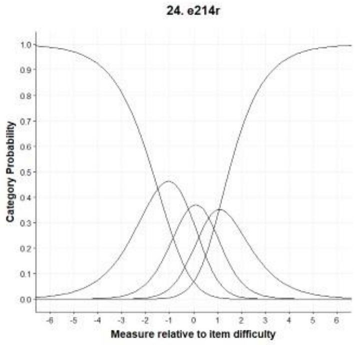
LEQ-CI[29]



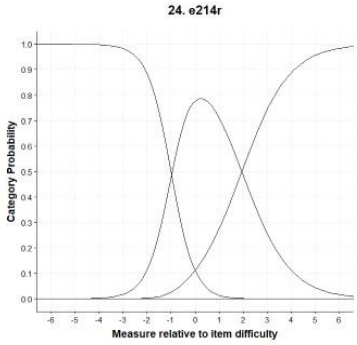
LEQ-CI[21]

45

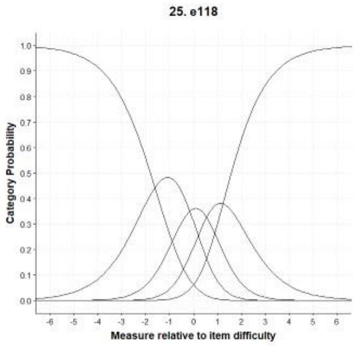
LEQ-CI[29]



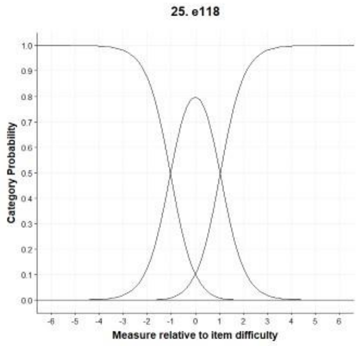
LEQ-CI[21]



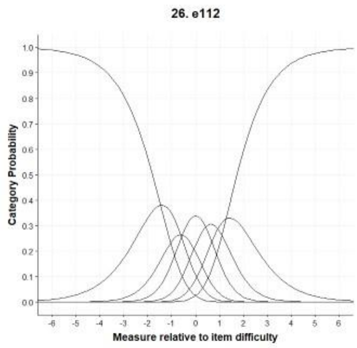
LEQ-CI[29]



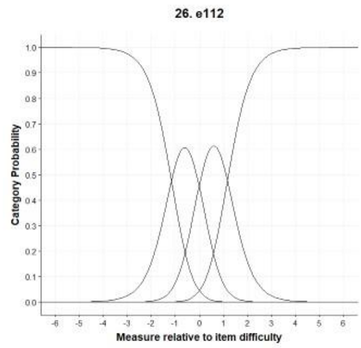
LEQ-CI[21]



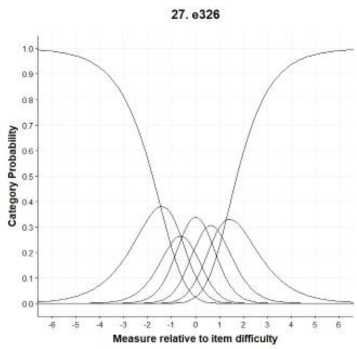
LEQ-CI[29]



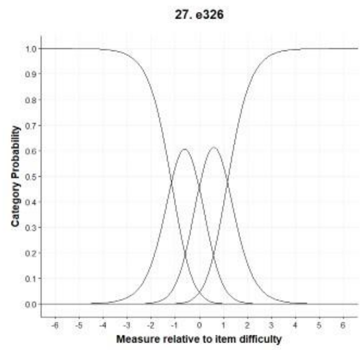
LEQ-CI[21]



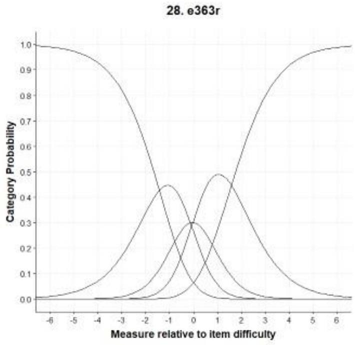
LEQ-CI[29]

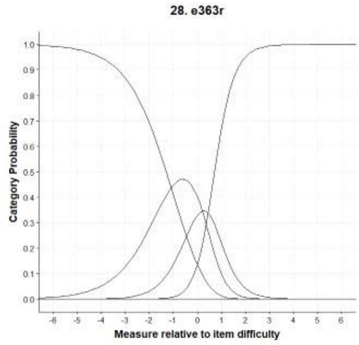


LEQ-CI[21]

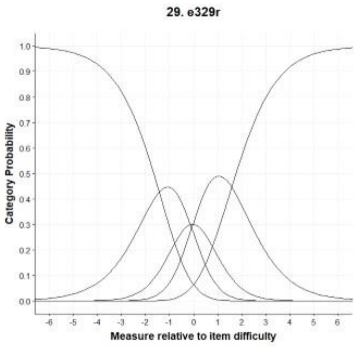


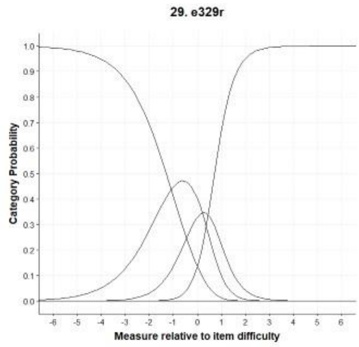
LEQ-CI^{29} LEQ-CI^{21}





LEQ-CI^{29} LEQ-CI^{21}





Supplemental Digital Content 2: Table comparing the Rasch analyses undertaken with the full sample (n = 330) and with underfitting persons removed (n = 284)

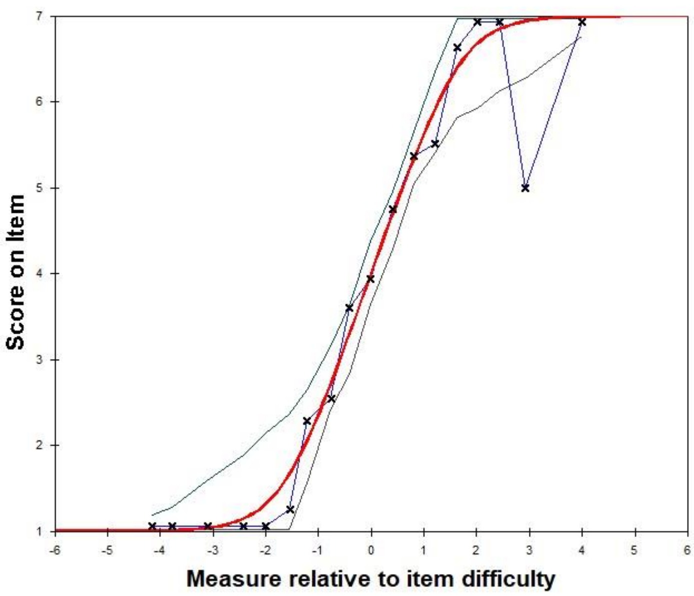
	Full sample (n = 330)	Underfitting persons removed (n = 284)
Rasch variance explained by measures	56.5%	60.8%
Eigenvalue for PCA of 1st residual	2.48	2.33
Range disattenuated correlations of 1st residual	0.89 - 0.98	0.92 – 1.00
Mean person location	-0.31	-0.39
Person separation	3.28	3.62
Person reliability	0.91	0.94
Item separation	9.69	10.07

Supplemental Digital Content 3.

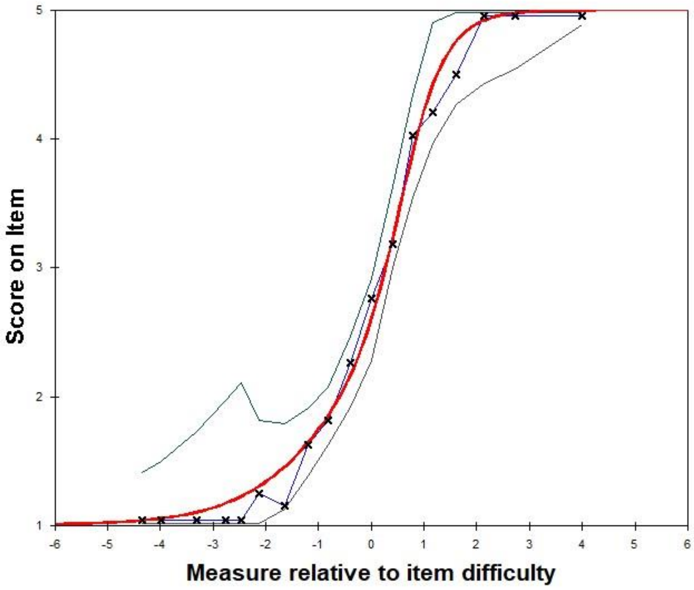
Empirical + theoretical item characteristic curves (ICCs) for the LEQ -CI[21]

The x-axis representsthe average ability of personsrelative to the item score (y-axis). The solid red line is an ogive shapewhich shows the expected relationshipbetween a person measureand the raw rating they would receive for that item. Areas of the curve that are steeperrepresentregionswhere the item is morediscriminating The blue line representsthe empirical(data-descriptive) item characteristiccurve. The black "x" representsthe average observedmeasurefor an interval on the latent variable which is perceived listeningeffort. The grey lines represent95% confidenceintervals. The blue line is expected to approximatethe red line (i.e., the empiricaldata fit the model) An "x" plotted outside of the confidence intervals is evidence of divergence from the Rasch-modelledexpected pattern of responses"

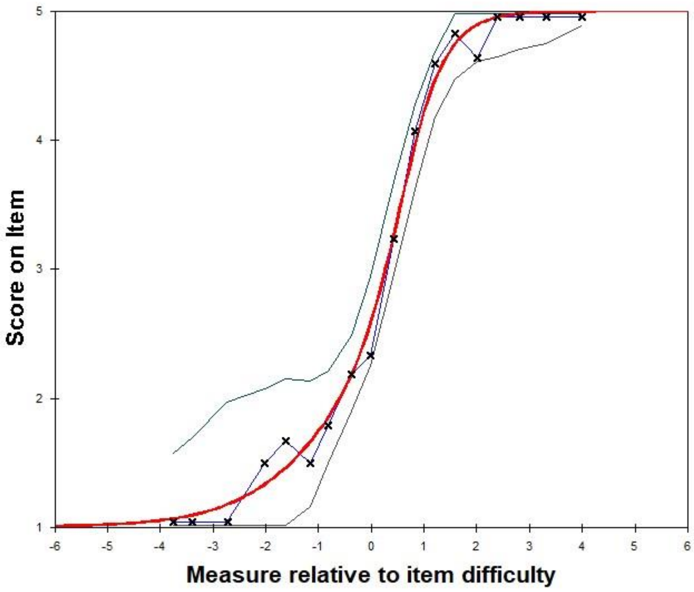
1. e398

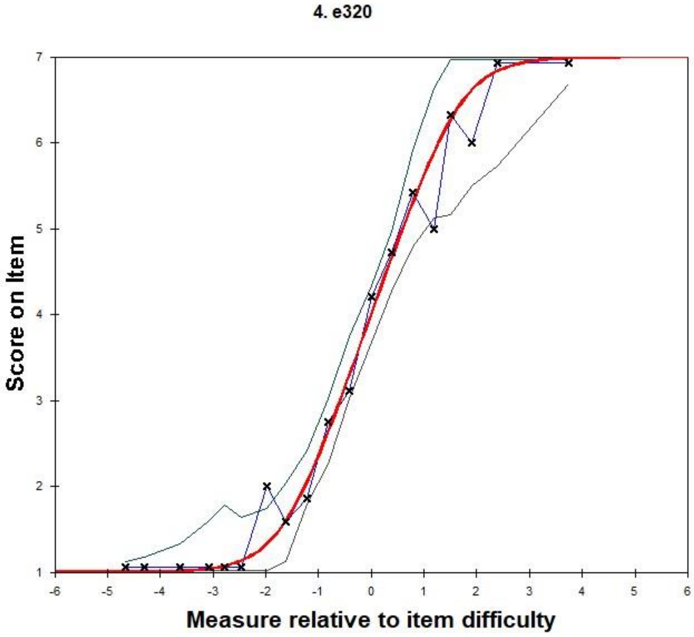


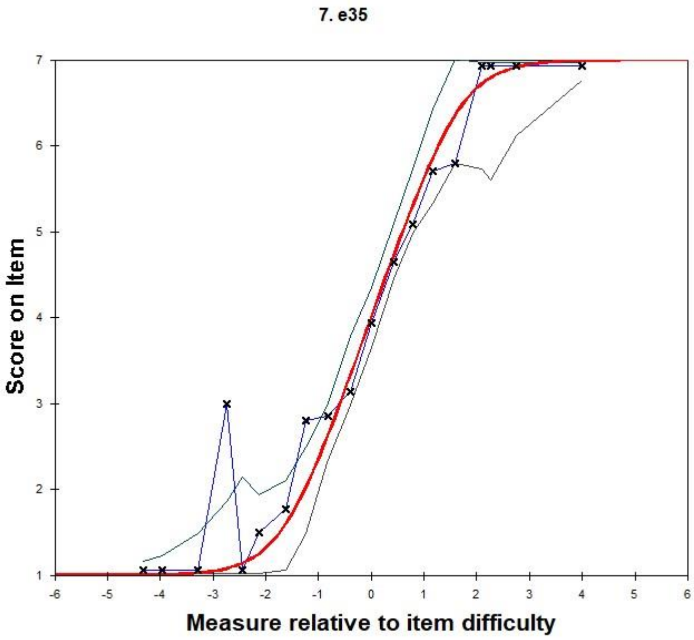
2. e319

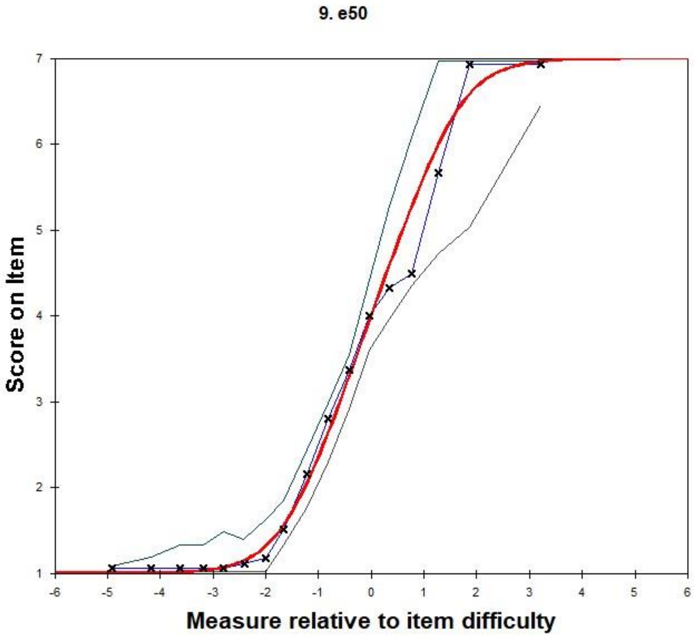

Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

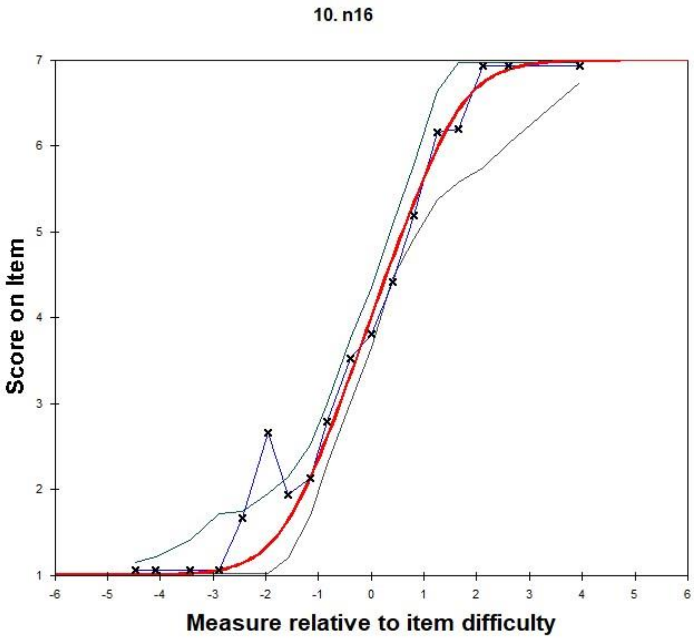
3. nq4

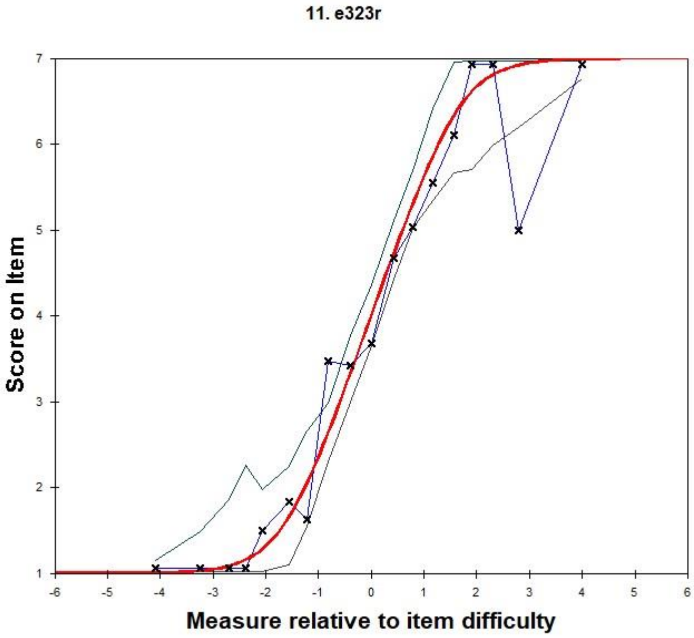

Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

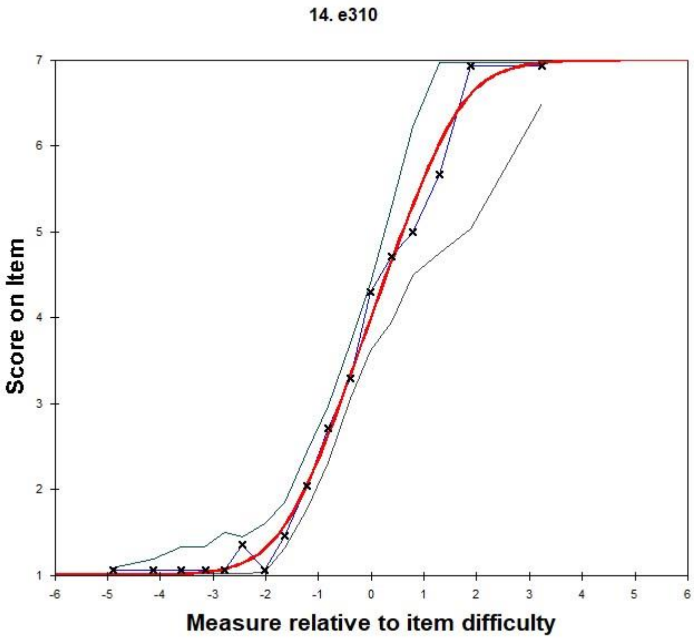

4. e320
Score on Item
Measure relative to item difficulty
7
6
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


7. e35
Score on Item
Measure relative to item difficulty
7
6
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

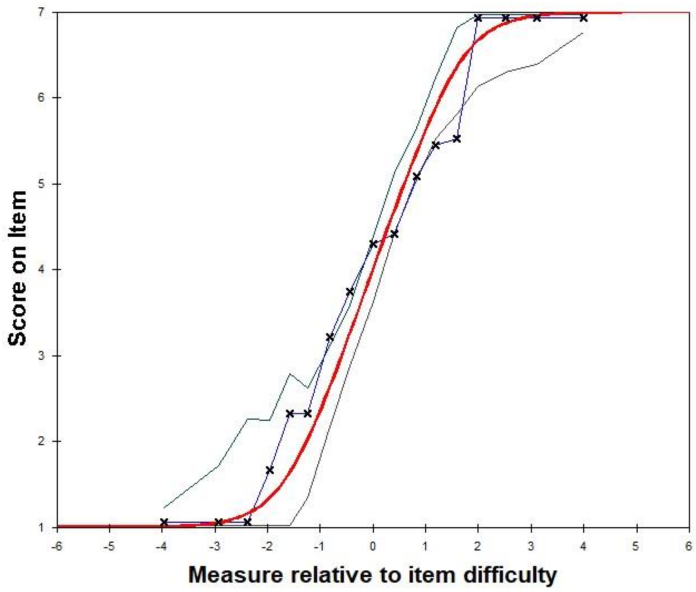

9. e50
Score on Item
Measure relative to item difficulty


10. n16
Score on Item
Measure relative to item difficulty

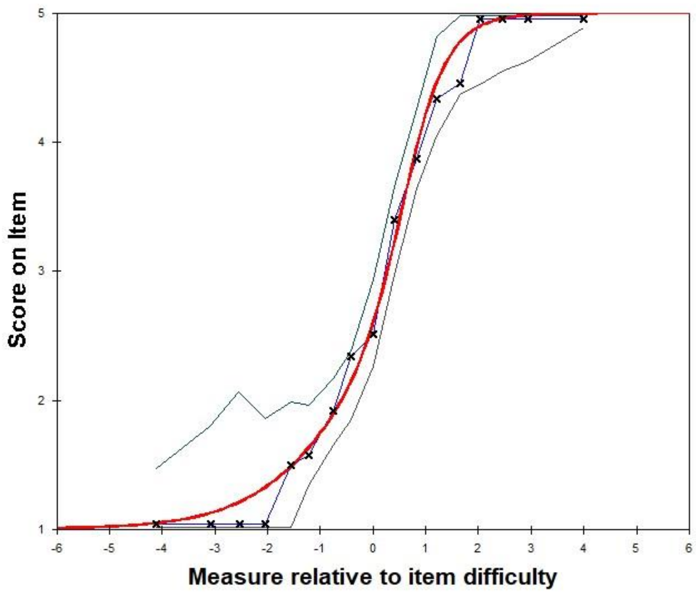

11. e323r
Score on Item
Measure relative to item difficulty

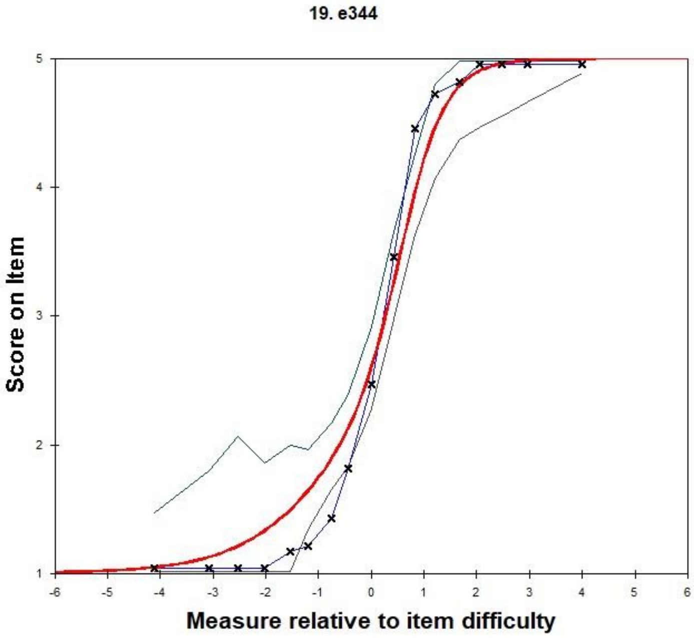

14. e310
Score on Item
Measure relative to item difficulty

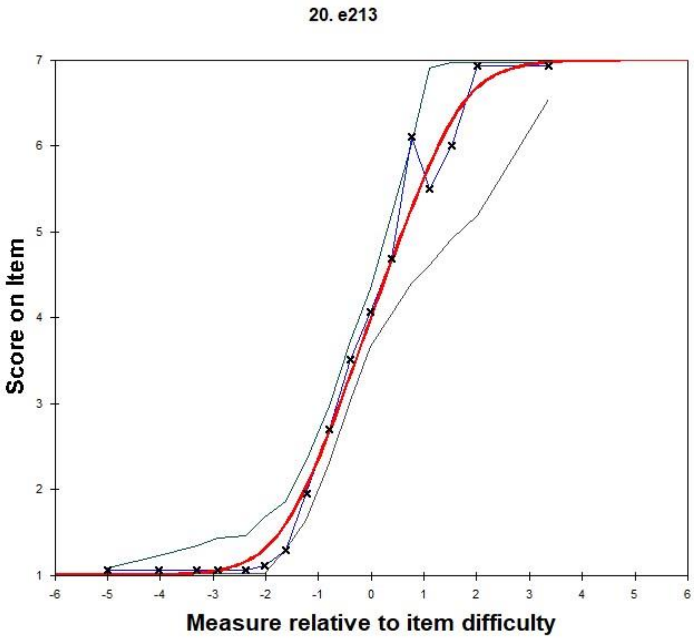
16. nq3r



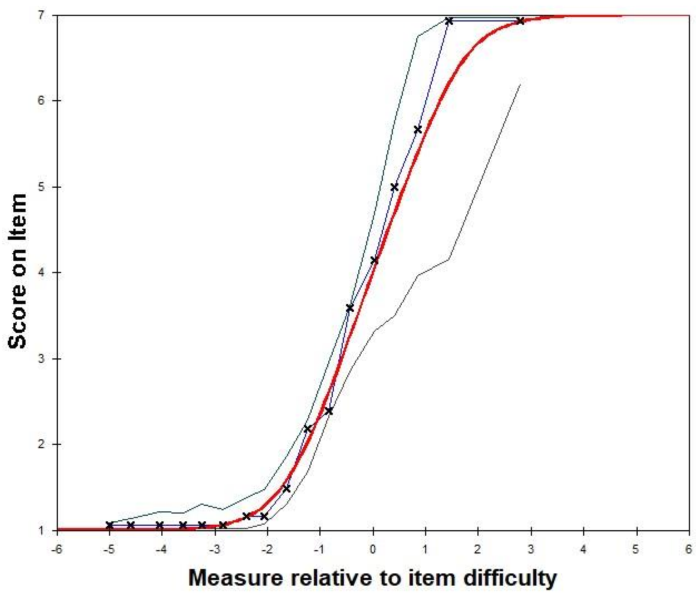
18. n11



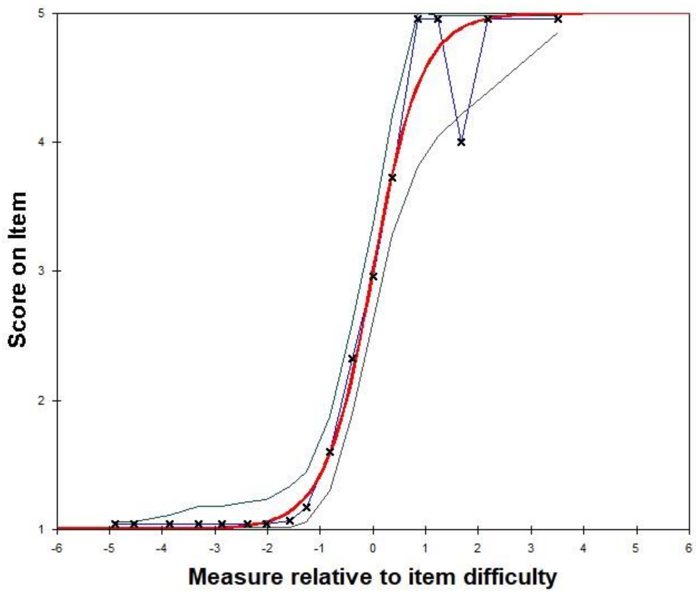

19. e344
Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

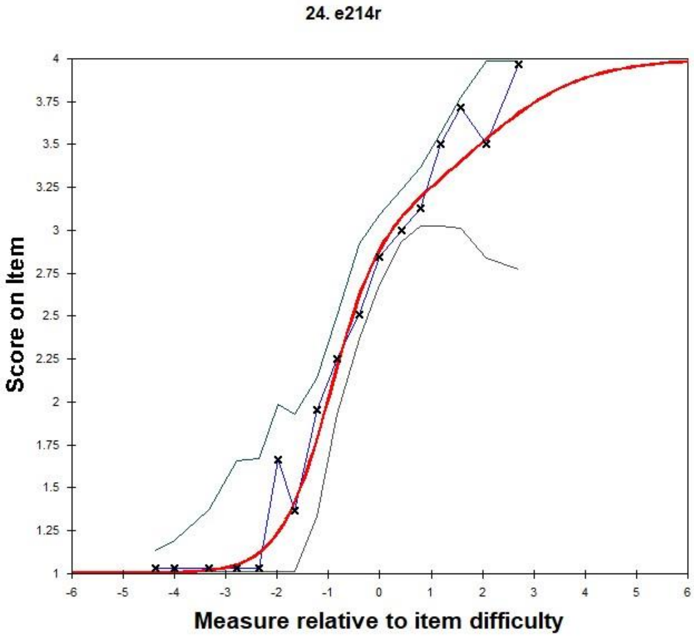

20. e213
Score on Item
Measure relative to item difficulty
7
6
5
4
3
2
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

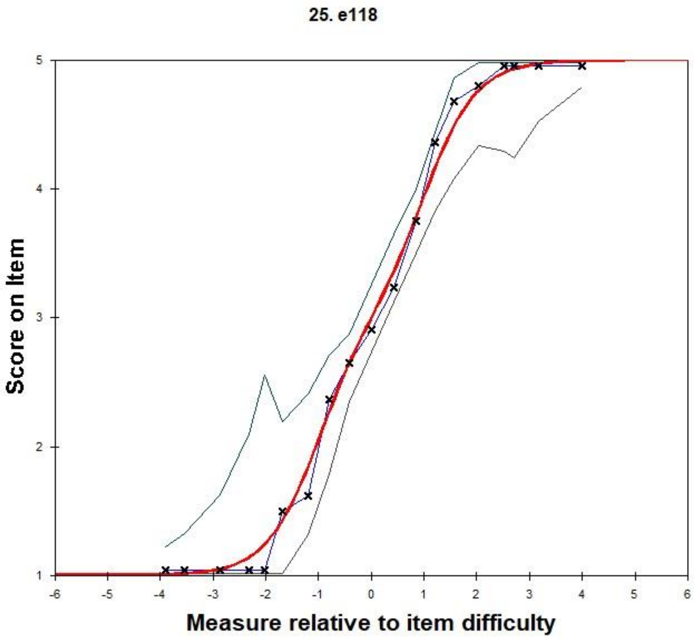
21. e103

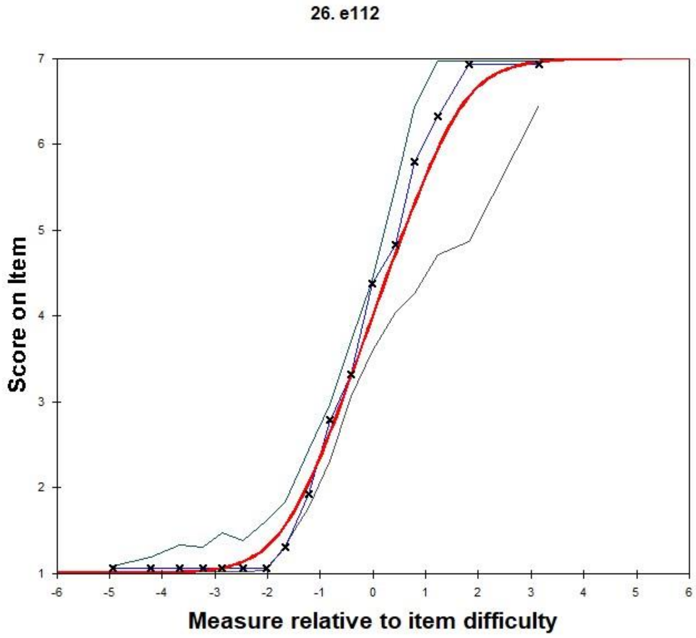


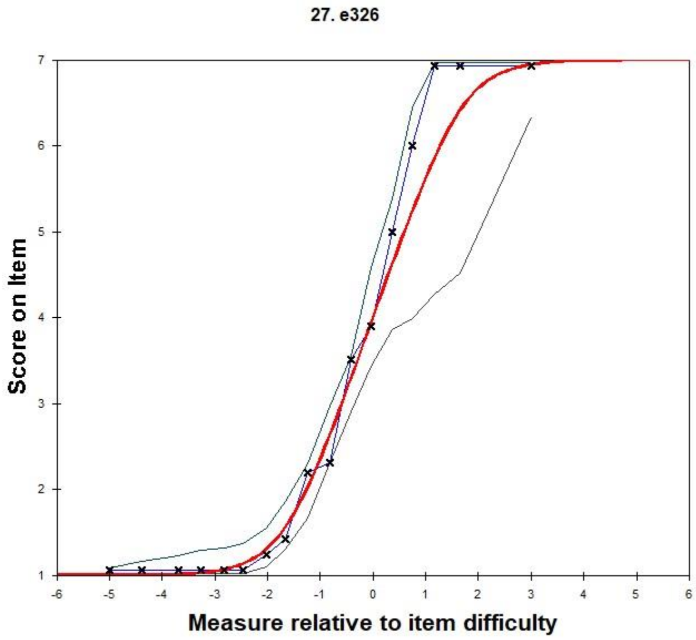
23. e105

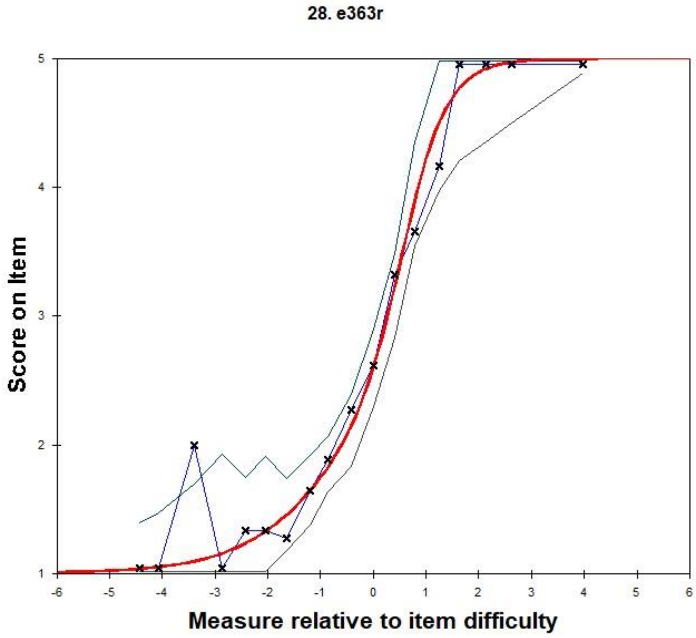


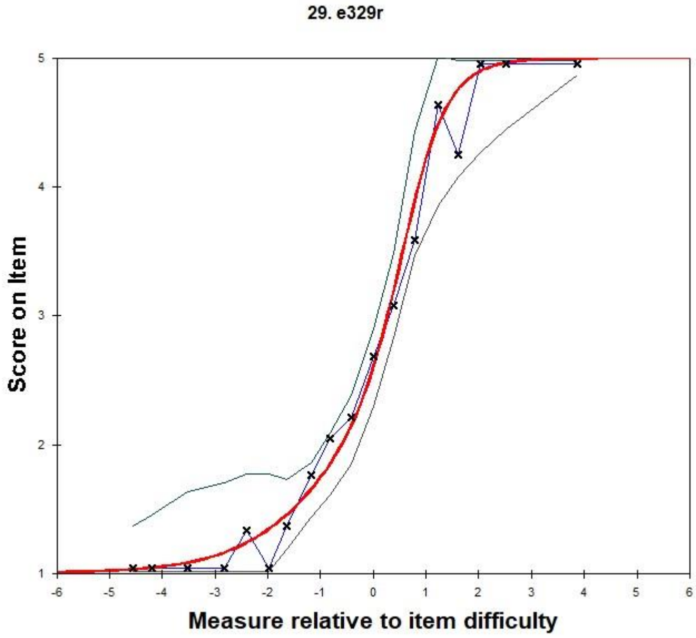

24. e214r
Score on Item
Measure relative to item difficulty
4
3.75
3.5
3.25
3
2.75
2.5
2.25
2
1.75
1.5
1.25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


25. e118
Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


26. e112
Score on Item
Measure relative to item difficulty


27. e326
Score on Item
Measure relative to item difficulty


28. e363r
Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


29. e329r
Score on Item
Measure relative to item difficulty
5
4
3
2
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

Supplemental Digital Content 4: Tables showing Winsteps generated DIF analysis (DIF) for gender and age for the LEQ-CI[21]
Table 4.1: Results of DIF analysis for gender

Item ID	Person Group	Observed-Expected Average	DIF Measure	DIF S.E.	Person Group	Observed-Expected Average	DIF Measure	DIF S.E.	DIF CONTRAST	Joint S.E.	Prob
e398	M	-0.04	-0.40	0.07	F	0.02	-0.43	0.06	0.03	0.09	0.9408
e319	M	0.22	-0.42	0.08	F	-0.14	-0.09	0.07	-0.33	0.11	0.0045
nq4	M	0.02	-0.81	0.08	F	-0.01	-0.81	0.07	0.00	0.11	0.8316
e320	M	0.21	-0.05	0.07	F	-0.14	0.19	0.06	-0.24	0.09	0.0063
e35	M	0.14	-0.34	0.07	F	-0.09	-0.19	0.06	-0.15	0.09	0.0256
e50	M	0.00	0.63	0.08	F	0.00	0.63	0.06	0.00	0.10	0.5162
n16	M	0.03	-0.13	0.07	F	-0.02	-0.11	0.06	-0.02	0.09	0.3804
e323	M	-0.08	-0.25	0.07	F	0.05	-0.33	0.06	0.08	0.09	0.4454
e310	M	-0.02	0.60	0.08	F	0.02	0.60	0.06	0.00	0.10	0.6047
nq3	M	-0.04	-0.58	0.07	F	0.02	-0.61	0.06	0.03	0.09	0.3308
n11	M	-0.12	-0.35	0.08	F	0.08	-0.53	0.07	0.18	0.11	0.0408
e344	M	0.01	-0.47	0.08	F	-0.01	-0.47	0.07	0.00	0.11	0.8532
e213	M	0.00	0.48	0.08	F	0.00	0.48	0.06	0.00	0.10	0.6526
e103	M	-0.05	1.10	0.08	F	0.04	1.02	0.07	0.08	0.11	0.8380
e105	M	-0.19	0.48	0.08	F	0.13	0.21	0.06	0.27	0.10	0.0268
e214	M	0.05	-0.29	0.12	F	-0.03	-0.15	0.09	-0.14	0.15	0.8090
e118	M	-0.13	-0.53	0.09	F	0.09	-0.77	0.08	0.24	0.12	0.0730
e112	M	-0.04	0.71	0.08	F	0.03	0.65	0.06	0.06	0.10	0.4402
e326	M	-0.06	0.89	0.08	F	0.05	0.80	0.07	0.09	0.10	0.3312
e363	M	0.07	-0.20	0.08	F	-0.04	-0.09	0.07	-0.10	0.11	0.3648
e329	M	0.04	-0.06	0.08	F	-0.03	0.00	0.07	-0.07	0.11	0.9951

M = males; F = females
DIF present if DIF CONSTRAST > 0.64 and p ≤ 0.05

Supplemental Digital Content 5: Raw score to 0-100 Rasch score conversion table for use with the LEQ-CI[21]

Raw Score	Rasch Score
21	0
22	12.06
23	18.37
24	21.9
25	24.37
26	26.29
27	27.87
28	29.23
29	30.43
30	31.51
31	32.51
32	33.42
33	34.28
34	35.09
35	35.85
36	36.58
37	37.28
38	37.95
39	38.6
40	39.22
41	39.82
42	40.41
43	40.98
44	41.53
45	42.07
46	42.59
47	43.1
48	43.6
49	44.09
50	44.57
51	45.04
52	45.5
53	45.95
54	46.39
55	46.82
56	47.25
57	47.67
58	48.08
59	48.48
60	48.89
61	49.28
62	49.67
63	50.06
64	50.44
65	50.81
66	51.19
67	51.56
68	51.92
69	52.29
70	52.65
71	53.01
72	53.37
73	53.72
74	54.07
75	54.43
76	54.78
77	55.13
78	55.48
79	55.83
80	56.18
81	56.53
82	56.89
83	57.24
84	57.59
85	57.95
86	58.31
87	58.67
88	59.04
89	59.4
90	59.78
91	60.15
92	60.53
93	60.92
94	61.31
95	61.7
96	62.11
97	62.52
98	62.94
99	63.37
100	63.8
101	64.25
102	64.71
103	65.18
104	65.66
105	66.16
106	66.67
107	67.2
108	67.75
109	68.32
110	68.9
111	69.52
112	70.16
113	70.82
114	71.52
115	72.26
116	73.03
117	73.85
118	74.73
119	75.67
120	76.69
121	77.81
122	79.05
123	80.47
124	82.15
125	84.23
109	68.32
110	68.9
111	69.52
112	70.16
113	70.82
114	71.52
115	72.26
116	73.03
117	73.85
118	74.73
119	75.67
120	76.69
121	77.81
122	79.05
123	80.47
124	82.15
125	84.23
126	87.05
127	91.75
128	100