Refinement and Validation of the Social Participation Restrictions Questionnaire: An Application of Rasch Analysis and Traditional Psychometric Analysis Techniques.

Eithne Heffernan[1,2], David W. Maidment[1,2], Johanna G. Barry[3,4], and Melanie A. Ferguson[1,3]

[1] National Institute for Health Research Nottingham Biomedical Research Centre, Ropewalk House, 113 The Ropewalk, Nottingham, UK; [2]Hearing Sciences Section, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, UK; [3] Nottingham University Hospitals National Health Service Trust, Nottingham, UK; [4] Medical Research Council Institute of Hearing Research, Nottingham, UK.

*Corresponding author: Eithne Heffernan, National Institute for Health Research Nottingham Biomedical Research Centre, Ropewalk House, 113 The Ropewalk, Nottingham, NG1 5DU, UK, eithne.heffernan1@nottingham.ac.uk

Keywords:

Hearing loss, social participation, psychometric analysis, Rasch analysis, classical test theory, outcome measurement, questionnaire validation

ABSTRACT

Objectives

The primary aim of this research was to refine and validate the Social Participation Restrictions Questionnaire (SPaRQ). The SPaRQ is a hearing-specific, patient-reported outcome measure that was originally developed through consultation with adults with hearing loss, clinicians, and researchers. This research comprised two studies. Study 1 aimed to assess the psychometric properties of the SPaRQ and to improve these properties by amending the questionnaire (e.g. removing items) as required. Study 2 aimed to validate the refined SPaRQ.

Design

In Study 1, 279 adults with hearing loss completed a long-form, 53-item SPaRQ. Rasch analysis, a modern psychometric analysis technique, was used to assess a range of psychometric properties for the questionnaire (e.g. unidimensionality, fit to the Rasch model). The properties of the individual items were also assessed (e.g. response dependency, differential item functioning).

In Study 2, 102 adults with hearing loss completed the refined SPaRQ. In addition, they completed three questionnaires that had been designed to measure related constructs. These were a hearing-specific questionnaire (Hearing Handicap Inventory for the Elderly), a generic health and disability questionnaire (shortened World Health Organization Disability Assessment Schedule 2.0), and a brief depression and anxiety screening questionnaire (Patient Health Questionnaire-4). Traditional psychometric analysis techniques (e.g. Cronbach's alpha) were used to assess the construct validity and internal consistency of the refined SPaRQ.

Results

Rasch analysis was used to refine the SPaRQ. The result was a 19-item measure divided into two subscales. The 9-item *Social Behaviors* subscale measured difficulties with performing actions in a social context due to hearing loss. The 10-item *Social Perceptions* subscale measured negative thoughts and feelings experienced in a social context due to hearing loss. Both Rasch analysis and the traditional psychometric analysis techniques demonstrated that each subscale had strong psychometric properties. In particular, each subscale passed the test of unidimensionality, displayed good fit to the Rasch model, and had high internal consistency. In addition, it was found that, as predicted, each subscale had strong, positive correlations with the hearing-specific questionnaire and moderate, positive correlations with the generic health and disability questionnaire and the depression and anxiety screening questionnaire. Taken together, these findings support the construct validity of the 19-item SPaRQ.

Conclusions

This was one of the first studies to devise a new hearing-specific outcome measure using Rasch analysis. Rasch analysis proved to be a powerful technique for supporting decisions regarding which items to retain in order to achieve a psychometrically robust questionnaire. Additional support for the robustness of this questionnaire came from the utilization of traditional psychometric analysis techniques. Therefore, this questionnaire has the potential to be used in research and clinical practice to evaluate whether auditory rehabilitation interventions improve social participation in adults with hearing loss. The next stage of this research will be to further validate this questionnaire by assessing its responsiveness in a clinical population. The combined use of modern and traditional psychometric analysis techniques should be considered in future questionnaire development and validation research.

INTRODUCTION

Outcome measurement is an integral element of both research and practice in the field of auditory rehabilitation (Saunders et al. 2005; Wong and Hickson 2012). It can be defined as the process of gathering information about healthcare services and interventions in order to demonstrate patient benefit, demonstrate cost effectiveness, inform the development of clinical practice guidelines, inform the allocation of healthcare resources, and ensure professional credibility (Beck 2000; Nemes 2003; Saunders et al. 2005). In the current era of evidence-based practice, it is essential to supplement patients' anecdotal reports and clinicians' expert opinions with quantitative evidence obtained using validated, high-quality outcome measures (Beck 2000; Bentler and Kramer 2000; Wong and Hickson 2012). A high-quality outcome measure is one that possesses robust measurement properties (e.g. content validity, construct validity) in accordance with published standards (e.g. Terwee et al. 2007). Furthermore, a high-quality outcome measure is one that has been developed and validated through a series of research studies of good methodological quality (e.g. appropriate statistical methods, adequate sample size) in accordance with published standards (e.g. Mokkink et al. 2010b; 2012).

Development of the Social Participation Restrictions Questionnaire

Numerous studies have demonstrated that many individuals with hearing loss experience participation restrictions (Kramer et al. 2002; Gopinath et al. 2012; Heffernan et al. 2016). Participation restrictions have been defined in the International Classification of Functioning, Disability, and Health (ICF) as the difficulties an individual experiences with involvement in life situations (World Health Organization 2013). These situations include family relationships, informal social relationships, recreation and leisure, community life, and remunerative employment (Danermark et al. 2013). The ultimate aim of adult auditory

rehabilitation is to reduce the participation restrictions that stem from hearing loss (Boothroyd 2007; Ferguson et al. 2017). As a result, high-quality patient-reported outcome measures (PROMs) are required that specifically assess the effectiveness of auditory rehabilitation interventions in terms of improved participation in adults with hearing loss.

To this end, we developed the Social Participation Restrictions Questionnaire (SPaRQ), a hearing-specific PROM, in accordance with best practice recommendations (Terwee et al. 2007; Mokkink et al. 2012). The first stage of development entailed extensive qualitative research with key stakeholders. Specfically, the items were generated by **(1)** conducting semi-structured interviews with adults with hearing loss, researchers, and clinicians (Heffernan et al. 2016) and **(2)** reviewing the literature, including the ICF Core Sets for Hearing Loss (Danermark et al. 2013) and existing questionnaires (Seekins et al. 2012; Granberg et al. 2014). The findings demonstrated that hearing-related participation restrictions primarily concerned difficulties experienced in social situations (e.g. family life, friendships, community life). The title and content of the SPaRQ were designed to reflect this finding. Where possible, the phrases and words of the adults with hearing loss were used to construct the items. Once a prototype of the SPaRQ had been developed, its relevance, clarity, comprehensiveness, and acceptability were evaluated and enhanced through **(1)** cognitive interviews with adults with hearing loss and **(2)** a subject matter expert survey with researchers and clinicians (Heffernan et al. Reference Note 2).

The product of this qualitative research was a 53-item version of the measure (SPaRQ-53) divided into three potential dimensions of participation restrictions: **(1)** 26 ‘behavior items’, which represented problems with performing actions in a social context due to hearing loss (e.g. difficulty with taking part in group discussions), **(2)** 16 ‘emotion items’, which represented negative feelings experienced in a social context due to hearing loss (e.g. feeling isolated at get-togethers), and **(3)** 11 ‘identity items’, which represented negative

social attributes perceived to stem from hearing loss (e.g. being perceived as unfriendly). At this stage of development, it is good practice to include all potentially relevant items, even if some of these items later prove to be tangential to the target construct. This is because it is easier to remove existing irrelevant items during psychometric analysis than it is to add new relevant items after psychometric analysis (Clark and Watson 1995; Reise et al. 2000). The items were accompanied by an 11-point response scale ranging from zero ("*Completely Disagree*") to ten ("*Completely Agree*"). Response scales with a broad range of response options are thought to allow participants to provide nuanced responses and to maximize responsiveness and reliability (Stewart and Archbold 1993; Weng 2004; Krosnick and Presser 2010). Furthermore, the cognitive interviews supported the use of an 11-point response scale. The next stage of this research utilized psychometric analysis techniques to further refine and validate the SPaRQ-53.

Traditional and modern psychometric analysis techniques

Currently, there are two prominent psychometric analysis techniques in the field of PROM development. The majority of hearing-specific PROMs that have been refined and validated have used traditional psychometric analysis techniques, such as using factor analysis to assess dimensionality and Cronbach's alpha to assess internal consistency (Cano and Hobart 2011; Hospers et al. 2016). However, despite their dominance, traditional techniques have several important limitations in comparison to modern psychometric analysis techniques, which include Rasch analysis and Item Response Theory (IRT) (Turk et al. 2006; Cano and Hobart 2011).

Firstly, various traditional psychometric analysis techniques are based on an underlying measurement theory known as Classical Test Theory (CTT). In essence, CTT posits that observed scores are the sum of true scores and random error (Hobart and Cano

2008). A true score can be defined as the expected value of observed scores across an infinite number of independent test-taking occasions. Random error can be defined as the remainder after the subtraction of the true score from the observed score. It has been argued that CTT cannot be adequately tested and falsified because it is based on definitions, rather than on assumptions that can be demonstrated to be true or false. Whilst it is possible to add assumptions to the main definitions of CTT in order to create falsifiable CTT models, these assumptions are determined by individual investigators (Hobart and Cano 2008; Brennan 2010; Raykov and Marcoulides 2011; Trafimow 2013). In contrast, modern psychometric analysis techniques are based on measurement theories that can be adequately tested and falsified (Hobart et al. 2007; da Rocha et al. 2013). These theories, known as Latent Trait Theories (LTTs), assume that the response to an item is a result of the interaction between the level of severity (e.g. participation restrictions) present in the respondent and the characteristics of the item, such as the level of severity captured by that item. Modern psychometric analysis techniques based on LTTs essentially assess the difference between predicted scores and observed scores for each item (Cano and Hobart 2011; da Rocha et al. 2013).

It has been argued that the use of traditional psychometric analysis techniques alone provides weak or circumstantial evidence for the validity of a test (Hobart and Cano 2008). In particular, the use of factor analysis and Cronbach's alpha to identify items for removal from a test and to assess the dimensionality of a test have been called into question. In terms of factor analysis, currently there is no consensus in the test development literature on how best to determine the number of factors to be extracted, with over-extraction and under-extraction being common problems (Wright 1996; Prieto et al. 2003). In terms of Cronbach's alpha, it has been demonstrated that a test can obtain a high alpha value, despite multidimensionality or low average item inter-correlation, simply by retaining a sufficient (i.e. ≥20) number of

items (Cortina 1993; Streiner 2003). By comparison, modern psychometric analysis techniques provide a stringent means of testing the unidimensionality of a test (Tennant and Pallant 2006; Yorke et al. 2012). This is important because unidimensionality, which means that a test measures a single latent trait, is an essential component of construct validity. It is also a prerequisite for calculating a total score for a test by summing its items, as it ensures that the test is a meaningful measure of a construct, rather than a list of somewhat miscellaneous items (Pesudovs et al. 2007; Yorke et al. 2012). In addition, modern psychometric analysis techniques provide a more accurate and comprehensive assessment of item characteristics than traditional techniques. For instance, modern techniques can be used to assess whether an item is biased in terms of characteristics such as age and gender. They can also be used to establish whether the response scale is functioning as expected for each individual item (Tennant and Conaghan 2007; Petrillo et al. 2015).

Another limitation of traditional psychometric analysis techniques is that the performance of a test is dependent on the sample in which that test is assessed, whereas it is preferable that tests are stable across samples (Turk et al. 2006; Cano and Hobart 2011). Therefore, the properties of the test (e.g. reliability, validity) and the item statistics (e.g. item difficulty, item discrimination) for that test are sample dependent (Hobart and Cano 2009). An advantage of modern psychometric analysis techniques is that several, though not all, of the test properties and item statistics produced are sample independent (Hagquist et al. 2009; Bonsaksen et al. 2013).

A further limitation of traditional psychometric analysis techniques is that they lead to the development of tests that produce ordinal, rather than interval, data. This has implications for the analysis of clinical trials data, as difference scores and change scores are most meaningful and accurate when they are calculated using interval data. Furthermore, ordinal data are more appropriate for group-level measurement than individual-level measurement

due to the wide confidence intervals around an individual ordinal score (Prieto et al. 2003; Hobart et al. 2007; Cano and Hobart 2011). The use of modern psychometric analysis techniques mean that the transformation of test data to interval scaling is possible (van der Velde et al. 2009).

Despite these limitations, traditional psychometric analysis techniques are still considered to be an important part of PROM development. For example, they are a crucial component of the COSMIN (i.e. Consensus-based Standards for the selection of health Measurement Instruments) checklist, which is used to assess the methodological quality of studies that investigate the measurement properties of PROMs (Mokkink et al. 2010b; 2012). Furthermore, certain important psychometric properties (e.g. test-retest reliability) that have been included in quality criteria guidelines for PROMs (e.g. Terwee et al. 2007) can only be assessed using traditional techniques. In addition, traditional techniques, which focus on the validity of the summed scale score, are a useful complement to modern techniques, which focus on the performance of items within a scale (Belvedere and de Morton 2010; da Rocha et al. 2013). For these reasons, many PROMs outside of hearing research have been refined and validated using a combination of traditional and modern psychometric techniques (e.g. Chachamovich et al. 2008; Whiteneck et al. 2011).

On this basis, the primary aim of this research was to refine and validate the SPaRQ using both modern and traditional psychometric analysis techniques. To achieve this aim, two consecutive quantitative studies were conducted. In Study 1, Rasch analysis, a modern psychometric analysis technique, was used to evaluate the psychometric properties of the SPaRQ-53 and to refine the questionnaire, such as by removing items with poor psychometric properties. In Study 2, traditional psychometric analysis techniques were used to validate the refined version of the questionnaire (SPaRQ).

METHODS AND MATERIALS

Study 1: Refinement

Procedure. This research was approved by the University of Nottingham Sponsor, the North East - Tyne and Wear South Research Ethics Committee, and the Research and Innovation Department of the Nottingham University Hospitals NHS Trust. Participants completed a paper-and-pen version of the SPaRQ-53 from home at their own pace. Participants who normally wear a hearing aid were instructed to answer the SPaRQ-53 as if they were currently wearing the hearing aid. In addition, the participants self-reported hearing loss via a validated screening instrument (Davis et al. 2007) and answered a series of demographic questions (e.g. age, gender, hearing aid use). All participants were entered into a prize draw for a chance to win one of six gift vouchers (1x£100, 5x£50).

Participants. In order to obtain stable Rasch analysis outputs, the minimum required sample size for Study 1 was 250 participants (Linacre 1994). The inclusion criteria were: **(1)** ability to complete and return questionnaires by post, **(2)** aged 18 years or older, **(3)** self-reported mild, moderate, or severe hearing loss, **(4)** good standard of written English, and **(5)** normal, or corrected-to-normal, vision.

The questionnaires were returned by 295 individuals. The majority were recruited via the National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre (BRC) participant database (n=227, 76.95%). Forty individuals (13.56%) were recruited via two UK audiology clinics: the Nottingham Audiology Service and the Queen Elizabeth Hospital Audiology Centre in Birmingham. Audiologists at these clinics identified potential participants during appointments from any stage of the care pathway and provided them with the study materials and questionnaires for completion at home. In addition, 28 individuals (9.49%) were recruited through an advertisement posted on the ‘Deafness and Hearing Loss’

online forum on the web site of a UK-based charity: Action on Hearing Loss. A total of 279 participants were included in the data analysis (Table 1), following the removal of 16 individuals who did not have hearing loss (n=6), had a profound hearing loss (n=3), or had >50% missing responses to the SPaRQ-53 (n=7).

Statistical analysis. Rasch analysis is the formal testing of a test, such as a PROM, against the Rasch model (Hobart and Cano 2009). In Rasch analysis, estimates of person severity (e.g. an individual's level of participation restrictions) and item severity (e.g. the level of participation restrictions expressed by an item) are used to obtain the predicted responses that should have occurred for the items and persons to satisfy the Rasch model (Hobart et al. 2007; da Rocha et al. 2013). Item and person estimates are both measured in logits, or the log odds of a respondent of a given level of severity, as indicated by their response to all the of items combined, having a 50% chance of affirming that item (Pallant and Tennant 2007; Yorke et al. 2012). Rasch analysis then assesses the degree to which observed and predicted responses differ. Where the differences between observed and predicted responses are deemed to be within an acceptable range, it is concluded that the data fit the Rasch model and thus fulfil the core requirements of measurement. These requirements are unidimensionality, invariance, and interval-level measurement (Hagquist 2001; Hobart et al. 2007; Hagquist et al. 2009).

The data were analyzed using the Rasch Unidimensional Measurement Models software program, known as RUMM2030 (Andrich et al. 2010). Rasch analysis was conducted and reported in accordance with recommendations of Tennant and Conaghan (2007). The specific psychometric properties assessed were as follows:

Fit to the Rasch model - A fundamental component of Rasch analysis is testing the fit of the data to the Rasch model for each individual item and for the test as a whole. One

indicator of fit to the model is the fit residual: a standardized summation of the differences between observed and predicted scores for an item (Tennant and Conaghan 2007; Hobart and Cano 2009). For the overall test, the mean fit residual of all of the items together should be close to zero and the standard deviation should be approximately ≤1.5. For individual items, the fit residual should fall within the range of ±2.5 (Pallant and Tennant 2007; Yorke et al. 2012; Darzins et al. 2014; Randall et al. 2014). In addition, the chi-square statistic of each item, which compares observed and predicted scores for groups of respondents who have different levels of severity, should be non-significant following a Bonferroni adjustment. Finally, the item characteristic curve, which plots the predicted response to an item against the person location on the severity continuum, should display a close alignment between observed and predicted scores for each item (Hagquist et al. 2009; Yorke et al. 2012).

Local independence - One of the primary assumptions of the Rasch model is that the test is unidimensional, which means that no meaningful clusters remain in the residuals once the Rasch factor (i.e. the main scale) has been extracted. In this study, a robust test of unidimensionality was conducted as part of Rasch analysis to ensure that a single dominant factor underpinned the questionnaire (Smith 2002; Tennant and Pallant 2006; Tennant and Conaghan 2007). Specifically, a Principal Component Analysis (PCA) of the residuals was performed. Subsequently, two sets of items were created that represented high-positive and high-negative PCA loadings on the first residual factor. These two item sets were used to form separate person estimates, which were compared through a series of independent *t* tests. The percentage of significant tests should be no greater than 5%. A 95% Binomial Proportions Confidence Interval (CI) was then applied to this percentage. If the lower bound of the CI overlaps 5%, the unidimensionality of the questionnaire is confirmed (Smith 2002; Tennant and Pallant 2006; Tennant and Conaghan 2007; Hagquist et al. 2009; Yorke et al. 2012).

In addition to multidimensionality, the assumption of local independence is violated by response dependency. Response dependency occurs when the response to an item is dependent on or determines the response to another item in the same test (Tennant and Conaghan 2007; Hagquist et al. 2009). For example, the answer to an item about socializing several times a week would determine the answer to an item about socializing once a week. Dependent items are considered to be redundant because they either replicate other items or they do not provide additional important information (Kurtaiş et al. 2011; Randall et al. 2014). Such items artificially inflate internal consistency and give undue weight to the concept measured by the redundant items in the total score (Hagquist et al. 2009; Kurtaiş et al. 2011; Lundgren Nilsson and Tennant 2011). In the absence of response dependency, the residuals should represent random error and therefore the correlations among the residuals should be low. Response dependency is detected when the residuals of two or more items have a correlation of ≥0.3 (Hobart and Cano 2009; Randall et al. 2014).

Targeting - This refers to the match between the range of severity measured by the items in the test and the range of severity present in the sample of individuals to be tested (Hagquist et al. 2009). A well-targeted scale is one with a high level of measurement precision. Targeting was assessed by examining whether the mean location of the persons on the severity continuum was close to the mean location of the items, which is always zero. Additionally, the person-item threshold distribution graph, which compares the distribution of item thresholds to the distribution of persons on the severity continuum, was inspected (Pallant and Tennant 2007; Tennant and Conaghan 2007).

Person separation reliability - The Person Separation Index (PSI) is a reliability index that quantifies the extent to which the scale detects differences between individual respondents. It is the Rasch analysis equivalent of Cronbach's alpha. The PSI value should be

≥0.7 for group use and ≥0.85 for individual use (Tennant and Conaghan 2007; van der Velde et al. 2009).

Differential item functioning (DIF) – This is a form of item bias that is detected when different participant groups (e.g. men and women) display different response patterns to a particular item, despite having the same level of the underlying trait (e.g. participation restrictions) that is being measured. DIF is assessed by conducting an analysis of variance for each item (Pallant and Tennant 2007; Tennant and Conaghan 2007; Hagquist et al. 2009). As recommended, DIF related to gender and age group was investigated (Tennant and Conaghan 2007).

Threshold ordering - A threshold is the transition point between two response options where either response option is equally probable (Pallant and Tennant 2007; Hagquist et al. 2009). In an 11-point response scale, there are 10 thresholds. For the thresholds to be ordered, an increase in the response options from zero to ten should represent an increase in the underlying dimension: participation restrictions. Thresholds are disordered when participants do not select response options in a manner that is consistent with their degree of the underlying dimension. For example, the transition between response options one and two should not represent a higher level of participation restrictions than the transition between response options three and four. Disordered thresholds affect fit to the model and indicate that the response scale is confusing or difficult to use. Thresholds can be re-ordered by collapsing response categories. Threshold ordering is assessed by inspecting category probability curves for each item (Pallant and Tennant 2007; Hagquist et al. 2009).

Person fit - A small number of participants who display an abnormal response pattern, as indicated by high, positive person fit residuals, can have a detrimental impact on Rasch analysis results, particularly item fit (Tennant and Conaghan 2007). In Study 1, 11

participants were removed from the dataset due to having a high (≥3), positive person fit residual. The Rasch analysis was repeated on the data from all 279 participants and it was found that there were no notable discrepancies between the analysis including these individuals and the analysis excluding these individuals. The removed participants varied in terms of their demographics (e.g. age, hearing aid use).

Study 2: Validation

Procedure. The procedure of Study 2 was identical to that of Study 1, with the exception that all study materials were delivered online.

Participants. In order to obtain robust construct validity and internal consistency estimates, the minimum required sample size for Study 2 was 100 participants (Terwee et al. 2007; Hobart et al. 2012). The inclusion criteria were identical to Study 1, with the exception that participants had to be able to complete and return questionnaires via the internet. All participants were recruited through the Nottingham BRC participant database. Those who had participated in Study 1 were not invited to participate in Study 2.

The questionnaires were returned by 109 individuals, of whom 102 were included in the data analysis (Table 1). Seven individuals were excluded because they had profound hearing loss (n=5) or they had a different otological condition (i.e. hyperacusis, auditory processing disorder) to hearing loss (n=2).

Materials. In addition to the refined version of the SPaRQ, each participant completed the three questionnaires outlined below. These questionnaires were selected in order to provide a means of assessing the construct validity of the refined SPaRQ (see 'Statistical Analysis' section).

Hearing Handicap Inventory for the Elderly (HHIE) - This 25-item, hearing-specific questionnaire consists of a 12-item 'social/situational' subscale and a 13-item 'emotional'

subscale accompanied by a 3-point response scale ("*Yes*"=4, "*Sometimes*"=2, and "*No*"=0) (Ventry and Weinstein 1982). Respondents are instructed to answer as if they are not wearing a hearing aid. There is evidence to support its psychometric properties from studies using traditional psychometric analysis techniques (Ventry and Weinstein 1982; Weinstein and Ventry 1983; Weinstein et al. 1986).

Shortened World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) - This 12-item, generic questionnaire measures functioning in six activity and participation domains: understanding and communication, mobility, self-care, getting along with others, life activities, and societal participation (Üstün et al. 2010a; 2010b). Respondents rate how much difficulty they have experienced in these domains across the past 30 days using a 5-point scale ("*None*"=1, "*Mild*"=2, "*Moderate*"=3, "*Severe*"=4, "*Extreme or cannot do*"=5). The WHODAS 2.0 has both a simple scoring method, suitable for analyzing a specific sample, and a complex, IRT-based scoring method, suitable for comparative analyses of different populations. There is evidence to support the psychometric properties of this questionnaire from studies using both traditional and modern psychometric analysis techniques (Luciano et al. 2010; Üstün et al. 2010b; Kim et al. 2015).

Patient Health Questionnaire-4 (PHQ-4) - This mental health screening tool consists of a two-item depression subscale and a two-item anxiety subscale (Kroenke et al. 2009). Respondents rate how often they have been affected by anxiety or depression symptoms in the last two weeks using a 4-point scale ("*Not at All*"=0, "*Several Days*"=1, "*More than Half the Days*"=2, "*Nearly Every Day*"=3). A score of at least three on the depression subscale identifies potential depressive disorder cases. A score of at least three on the anxiety subscale identifies potential anxiety disorder cases. The total score represents a general measure of mental health disability and symptom burden. The PHQ-4 has been validated through several studies using traditional psychometric analysis techniques and can be used to screen for

anxiety and depression in general population (Kroenke et al. 2009; Löwe et al. 2010; Kerper et al. 2014).

Statistical Analysis. The data were analyzed using IBM SPSS Statistics for Windows Version 22.0. The analysis was conducted and reported in accordance with recommendations from the literature (Terwee et al. 2007). The following psychometric properties were assessed:

Construct validity - This refers to the extent to which the scores of a test are consistent with hypotheses based on the assumption that the test is a valid measure of the target construct. This includes hypotheses about the relationship between the test and other measures, as well as hypotheses about the expected differences in the scores of relevant subgroups. At least 75% of these hypotheses should be confirmed (Terwee et al. 2007; Scholtes et al. 2011).

Firstly, Spearman's rank order correlation coefficient (r_s) was used to assess predictions about the convergent validity the SPaRQ. Convergent validity refers to the extent to which an instrument is correlated with another instrument that measures a related or similar construct (Oppenheim 2000). It was predicted that the SPaRQ would have a strong, positive correlation (±0.6 or above) with the hearing-specific measure (HHIE) and a moderate, positive correlation (±0.3 to ±0.59) with the generic measure (WHODAS 2.0). In addition, it was predicted that the SPaRQ would have a moderate, positive correlation with the mental health screening tool (PHQ-4), as both hearing loss and participation restrictions have previously been associated with depression and anxiety (Kramer et al. 2002; Glass et al. 2006; Nachtegaal et al. 2009).

Secondly, the Mann-Whitney *U* test was used to test predictions about the discriminative validity of the SPaRQ. Discriminative validity refers to the ability of a PROM

to distinguish between patient subgroups (De Vet et al. 2011). It was predicted that participants with high (≥3) PHQ-4 scores on the anxiety subscale and the depression subscale would obtain significantly higher SPaRQ scores than participants with low PHQ-4 scores.

Internal consistency - This refers to the level of inter-relatedness amongst the items (Scholtes et al. 2011). Cronbach's alpha was used to assess this property. This statistic should fall within the range of 0.7-0.95 for a PROM or its subscales (De Vet et al. 2011; Raykov and Marcoulides 2011). Additional indicators of internal consistency were the mean corrected inter-item correlation, which should fall within the range of 0.3-0.7, and the mean corrected item-total correlation, which should be ≥0.3 (Ferketich 1991; Kline 2013).

RESULTS

Study 1: Refinement

The derivation of the Rasch model used in the analysis was the Partial Credit Model, which assumes that each item has an individual rating scale structure (Masters 1982).

Creation of subscales. Rasch analysis demonstrated that the SPaRQ-53 was multidimensional, as it failed the test of unidimensionality (percentage of significant *t* tests=16.97%). It was decided, based on an inspection of the statistical results, as well as the results of the earlier qualitative studies used to generate the content of the questionnaire, that the items should be divided into two subscales and that Rasch analysis should be conducted separately on each subscale. The first subscale, *Social Behaviors*, contained all 26 behavior items. The second subscale, *Social Perceptions*, contained all 27 emotion and identity items.

Iterative item reduction. Items that displayed poor psychometric properties were removed iteratively from their respective subscales (see Supplemental Digital Content 1 and Supplemental Digital Content 2). The majority of items were removed due to poor fit, response dependency, or both. For example, Item 4, which concerned watching live events or

shows, displayed poor fit (fit residual=2.91, chi square statistic=1.94, DF=3, *p*=0.585). A small number of items were removed due to having >15% missing responses, as this is indicative of serious flaws (i.e. irrelevance, ambiguity, or intrusiveness). For example, Item 15, which concerned participation in workplace conversations, had 43.01% blank responses. The item reduction process ceased when each subscale displayed the requisite psychometric properties (e.g. fit to the Rasch model and local independence). The removed items were then reintroduced one-by-one to determine whether they could be reinstated without negatively affecting these properties. Ultimately, the item reduction process yielded a 9-item *Social Behaviors* subscale and a 10-item *Social Perceptions* subscale (Table 2).

Fit to the Rasch model. The overall item fit residual demonstrated that the *Social Behaviors* subscale displayed relatively good fit to the Rasch model (M=0.01, SD=1.54). In addition, the *Social Perceptions* subscale displayed good fit to the Rasch model (M=0.45, SD=1.17). Each individual item displayed good fit within their subscale (Table 2).

Local independence. The *Social Behaviors* subscale passed the test of unidimensionality (percentage of significant *t* tests=4.33%). The *Social Perceptions* subscale passed the test of unidimensionality when the 95% Binomial Proportions Confidence Interval was inspected (percentage of significant *t* tests=6.53%, 95%CI=3.8%-9.3%). Response dependency was not a concern in either subscale, as the residual correlations between the items fell below 0.3 (see Supplemental Digital Content 3 and Supplemental Digital Content 4).

Targeting. The *Social Behaviors* subscale was well-targeted, as demonstrated by the person location (M=-0.23, SD=1.28) and the person-item threshold distribution (Fig. 1). The *Social Perceptions* subscale was also well-targeted according to the person location (M=-0.12, SD=1.64) and the person-item threshold distribution (Fig. 1). However, as these

distributions show, in each subscale, there were a small number of individuals at the extreme ends of the severity continuum who were not captured by the items.

Person separation reliability. The *Social Behaviors* person separation reliability was high, both including extreme scores (PSI=0.914) and excluding extreme scores (PSI=0.927). In addition, the *Social Perceptions* person separation reliability was high, both including extreme scores (PSI=0.925) and excluding extreme scores (PSI=0.94).

Differential item functioning. The analysis of variance results demonstrated that none of the items displayed DIF related to gender or age (see Supplemental Digital Content 5 and Supplemental Digital Content 6).

Threshold ordering. The *Social Behaviors* items had ordered thresholds. In the *Social Perceptions* subscale, Item 34 and Item 41 had disordered thresholds, as demonstrated by their category probability curves (see Supplemental Digital Content 7). Collapsing individual thresholds so that these items had seven, rather than 11, response options achieved sufficiently ordered thresholds (see Supplemental Digital Content 8).

Person fit. The overall person fit for both *Social Behaviors* (M=-0.4, SD=1.25) and *Social Perceptions* (M=-0.28, SD=1.23) was acceptable.

Study 2: Validation

The two subscales of the refined, 19-item SPaRQ were analyzed separately, as they are independent, unidimensional subscales (Pesudovs et al. 2007).

Construct validity. All convergent validity predictions were confirmed for the *Social Behaviors* subscale and the *Social Perceptions* subscale (Table 3). In addition, all discriminative validity predictions were confirmed. Specifically, as predicted, participants who scored at least three on the PHQ-4 anxiety subscale obtained a significantly higher mean rank on the *Social Behaviors* subscale (M=66.16, n=25) than participants who scored less

than three (M=46.74, n=77): U(104)=1329, Z=2.85, p=0.004. Similarly, participants who scored at least three on the PHQ-4 depression subscale obtained a significantly higher mean rank on the *Social Behaviors* subscale (M=72.66, n=25) than participants who scored less than three (M=44.63, n=77): U(104)=1491.5, Z=4.12, p<0.001.

Furthermore, participants who scored at least three on the PHQ-4 anxiety subscale obtained a significantly higher mean rank on the *Social Perceptions* subscale (M=69.12, n=25) than participants who scored less than three (M=45.78, n=77): U(104)=1403, Z=3.43, p<0.001. Finally, participants who scored at least three on the PHQ-4 depression subscale obtained a significantly higher mean rank on the *Social Perceptions* subscale (M=71.04, n=25) than participants who scored less than three (M=45.16, n=77): U(104)=1451, Z=3.8, p<0.001.

Internal consistency. All of the internal consistency statistics fell within the required range (Table 4).

DISCUSSION

The primary aim of this research was to refine and validate a hearing-specific, patient-reported outcome measure (PROM) called the Social Participation Restrictions Questionnaire (SPaRQ). This study was novel in that it used a combination of modern and traditional psychometric analysis techniques to achieve this aim; an approach rarely used in hearing research. The results demonstrated that the SPaRQ had a range of strong psychometric properties. Therefore, this measure has the potential to serve as a PROM in research and clinical practice to evaluate the impact of auditory rehabilitation interventions on participation restrictions in adults with hearing loss.

Refinement of the SPaRQ

In Study 1, Rasch analysis was used to refine a 53-item version of the questionnaire (SPaRQ-53) so that it ultimately became the finalized SPaRQ, which comprised a 9-item *Social Behaviors* subscale and a 10-item *Social Perceptions* subscale. Each subscale is unidimensional and thus independent of one another. This means that, whilst there may be value in combining the subscale scores to generate an overall total score for the SPaRQ in some research and clinical contexts, this may not be advisable in high stakes scenarios, such as high-quality clinical trials. Each subscale displayed strong psychometric properties, including fit to the Rasch model, person separation reliability, targeting, and local independence. These findings provide support for the construct validity of the measure (Pesudoves et al. 2007; Scholtes et al. 2011).

In terms of the individual items, Rasch analysis led to the selection of items that displayed adequate fit, low response dependency, and an absence of DIF. This indicates that, within their subscales, the items measured the same underlying construct and that redundancy, gender bias, and age bias were not a concern. The majority of the items had ordered thresholds, which provides statistical justification for the 11-point response scale (Pesudovs et al. 2007). However, two *Social Perceptions* items should ideally be re-scored from an 11-point scale to a 7-point scale in order to achieve sufficiently ordered thresholds, particularly in high stakes contexts. This is similar to the complex scoring system of the WHODAS 2.0. In addition, in high stakes contexts, it would be advantageous to covert the raw (i.e. ordinal) SPaRQ scores into logit (i.e. interval) scores that have been produced by the Study 1 Rasch analysis (Hobart and Cano 2009).

The iterative item reduction process led to the removal of 34 items that displayed poor psychometric properties. Firstly, three items were removed from the *Social Behaviors* subscale primarily because they were associated with a large number of missing responses, which can be indicative of irrelevance, ambiguity, and intrusiveness (De Vet et al. 2011).

Two of these items asked about participation in workplace conversations and in educational or training activities and thus were likely to be irrelevant to the large proportion (66%) of retired participants. Though these items may be relevant to some participants, it was important to remove them to ensure that the SPaRQ did not attract large numbers of blank responses in the future. The other item concerned participation in community, voluntary, or religious activities. Participants may have overlooked this item because they do not take part in such activities or because the wording of the item was too complex.

Several *Social Behaviors* items were removed due to displaying poor fit, which suggests that they measure a construct that is different from, though possibly related to, the target construct. A review of the content of these items showed that some may have exhibited poor fit because they were not necessarily social in nature, such as one item that asked about listening to the radio or music. However, other items removed due to poor fit appeared to be social in nature, such as an item that asked about attending social gatherings. Though Rasch analysis identifies poorly fitting items, it does not reveal the causes of this poor fit. Nevertheless, it was important to adhere to the pre-specified item removal criteria and thus delete these items, as the aim of the research was to produce a psychometrically robust questionnaire.

In terms of the *Social Perceptions* subscale, the majority of the identity items were removed due to poor fit, including items about being perceived as incapable and demanding. This finding suggests that items concerning the negative social attributes associated with hearing loss measure a distinct construct from items concerning the negative emotions associated with hearing loss. Also, it may be that some identity items, which pertained to pretending to understand others and preferring to sit quietly whilst others talk, displayed poor fit because they measured coping, rather than participation restrictions. An alternative

explanation for this poor fit is that these items were longer and more complex than the other items and consequently were more difficult to complete.

A number of items were removed from each subscale because they displayed response dependency, which means that they overlapped with or were dependent on the responses to other items. For example, in the *Social Behaviors* subscale, items about interactions with family were dependent on items about interactions with friends. Following the iterative item reduction process, the items about family interactions were removed. Though family interactions were identified as being relevant content in the qualitative phase of the SPaRQ development, it was necessary to discard these items because their dependency means that they do not provide important information that cannot be gained from the other items (Kurtaiş et al. 2011; Randall et al. 2014). Retaining dependent items can increase the length of the measure and thus respondent burden, artificially inflate internal consistency, over-weight their content in the total score, and violate the assumption of local independence (Hagquist et al. 2009; Kurtaiş et al. 2011; Lundgren Nilsson and Tennant 2011).

It is likely that the results of the item reduction process would have differed had a more traditional psychometric analysis approach been utilized. Petrillo et al. (2015) evaluated a visual functioning questionnaire and found that modern techniques (i.e. Rasch analysis and IRT) identified many item flaws (e.g. poor fit, poor targeting, and disordered thresholds) that were not identified by traditional techniques. The Study 1 results might also have differed had IRT been utilized, rather than Rasch analysis. IRT involves identifying the statistical model that best explains the data, whereas Rasch analysis involves identifying the data that satisfy the Rasch model requirements (Hobart and Cano 2009, da Rocha et al. 2013). Also, the Rasch model has one parameter (i.e. item difficulty), whereas IRT models can have several parameters (e.g. item discrimination, guessing). Proponents of IRT often consider Rasch analysis to be restrictive and a possible threat to content validity, whereas proponents

of Rasch analysis tend to view the modelling of data in IRT as too accepting of poor quality data that are a potential threat to construct validity. Rasch analysis was selected for Study 1 because it is unique in that it facilitates the stringent assessment of unidimensionality and it tests whether the data fulfil the core requirements of measurement; thus providing a means for transforming the data into interval scaling (Hobart and Cano 2009; Cano and Hobart 2011; da Rocha et al. 2013).

Validation of the SPaRQ

In Study 2, traditional psychometric analysis techniques were used to assess the construct validity of each subscale in the 19-item finalized SPaRQ. Support was obtained for the all of the convergent validity and discriminative validity predictions. This finding, when combined with the Study 1 results pertaining to fit to the Rasch model and unidimensionality, provides strong evidence in favor of the construct validity of the SPaRQ (Terwee et al. 2007). This indicates that this PROM measures the construct that it purports to measure: hearing-related social participation restrictions.

The internal consistency of the subscales was also established. Specifically, each subscale had Cronbach's alpha values within the required range of 0.7-0.95 (Terwee et al. 2007; De Vet et al. 2011). It has been suggested that Cronbach's alpha values greater than 0.9 are indicative of the presence of redundant items (Pesudovs et al. 2007). However, the Rasch analysis demonstrated that the items had low response dependency, which indicates that the high Cronbach's alpha values are not attributable to redundancy in this case. It may be that the large number of response options in the response scale contributed to these high values (Weng 2004). Moreover, Cronbach's alpha has several limitations that mean that it must be interpreted with caution (Pesudovs et al. 2007). As stated previously, it is sample dependent

and can be artificially inflated by the inclusion of a large number (≥20) of items (Cortina 1993; Streiner 2003).

The next stage of this research will assess the responsiveness and interpretability of the SPaRQ. This will be achieved by administering both the SPaRQ and a clinical global impression scale to adults with hearing loss before and after receiving an auditory rehabilitation intervention. A range of statistical analyses will be performed, including investigating floor and ceiling effects and establishing the smallest detectable change (SDC) and minimal important change (MIC) of the SPaRQ (Terwee et al. 2007).

Limitations

A potential limitation of this research is that the Study 1 sample size is lower than commonly recommended sample sizes for Rasch analysis studies (e.g. ≥500) (Cappelleri et al. 2014). Whilst certain analyses (e.g. person separation reliability) do not appear to be affected by small sample sizes, other analyses (e.g. threshold ordering, individual item fit) can be adversely affected (Chen et al. 2014). However, Linacre (1994) has advised that a minimum sample size of 250 cases is are generally sufficient to produce stable Rasch analysis outputs that are unlikely to differ noticeably from outputs produced by larger (e.g. 1000 cases) datasets. This suggests that the Study 1 sample size is adequate. Furthermore, Study 1 was supported by qualitative research that provided in-depth information about each item, as well as by traditional psychometric analyses, which is the advised approach for Rasch analysis studies with relatively small samples (Chen et al. 2014).

Another limitation of this research is that the majority of participants were recruited from the NIHR Nottingham BRC participant database, which means that the sample may not fully represent all individuals with hearing loss. For example, the database contains a large number of experienced hearing aid users and experienced research participants. Furthermore,

in Study 2, it was necessary to administer the SPaRQ and the three questionnaires used for assessing construct validity to participants on one single occasion. Ideally, the questionnaire undergoing development should be administered on a prior occasion to the other questionnaires so that responses to the former do not influence responses to the latter (De Vet et al. 2011). However, it is common practice in investigations of construct validity to administer all of the relevant questionnaires to participants on one occasion (e.g. Chisolm et al. 2005; Post et al. 2008).

A further limitation pertained to the testing of the discriminative validity hypotheses. Specifically, the sample size of the subgroups who had high depression and anxiety scores on the PHQ-4 was small (n=25). Ideally, the subgroups in a discriminative validity analysis should have 50 cases (Terwee et al. 2007). Furthermore, research on additional subgroups is needed in order to thoroughly assess the discriminative validity of the SPaRQ. A final potential limitation of this research is that a paper-and-pen SPaRQ-53 was administered in Study 1, whereas an online, electronic SPaRQ was administered in Study 2. Many previous studies have established that paper-and-pen questionnaires are equivalent to their electronic counterparts, particularly if their appearance remains very similar across the different formats (Cook et al. 2004; Gwaltney et al. 2008; Bishop et al. 2010). However, some studies have found that the responses and psychometric properties of online questionnaires differ from their paper-and-pen counterparts (Vallejo et al. 2007; Thorén et al. 2012). Consequently, the optimal approach would have been to have used a single format for both Study 1 and Study 2. Unfortunately, it was not possible to post paper-and-pen questionnaires to participants in Study 2 due to the limited resources and strict timeframe of this study.

Implications

This research has provided insights on the conceptualization and measurement of participation restrictions in adults with hearing loss. The ICF conceptualizes participation restrictions as consisting of numerous potentially challenging categories of functioning, such as conversation, employment, and family relationships (Danermark et al., 2013; WHO 2013). Others have conceptualized hearing handicap, an equivalent construct, as consisting of a situational domain, such as avoiding groups of people, and an emotional domain, such as feeling irritable (Ventry and Weinstein 1982). Social isolation, a related construct, has been conceptualized as consisting of an objective domain, such as a reduced social network size, and a subjective domain, or the perceived, emotional experience of isolation (Hawthorne 2008). In the SPaRQ, hearing-related participation restrictions are conceptualized as comprising *Social Behaviors* (i.e. problems with performing actions in a social context due to hearing loss) and *Social Perceptions* (i.e. negative thoughts and feelings experienced in a social context due to hearing loss). The *Social Behaviors* dimension is comparable to the ICF conceptualization of participation restrictions, whilst the *Social Perceptions* dimension is similar to the subjective and emotional domains of handicap and isolation. Both of these subscales can provide distinct yet equally important insights regarding hearing-related participation restrictions. For instance, a respondent could have a low *Social Behaviors* score and a high *Social Perceptions* score, indicating that they feel isolated despite partaking in various social activities. The SPaRQ does not measure objective participation restrictions, such as counts of social contacts or counts of social activities, because many individuals are satisfied with having a relatively quiet social life and a small social network.

This research has highlighted the value of utilizing modern psychometric analysis techniques to refine hearing-specific PROMs. One of the main advantages of these techniques is that they allow unique measurement properties to be examined. In particular, unidimensionality can be rigorously tested and the performance of each response option on

the response scale can be assessed. In addition, a greater deal of information is available on the individual items, including their degree of DIF and response dependency (Yorke et al. 2012). Furthermore, the use of modern techniques in outcome measurement can ultimately facilitate the transformation of PROM data to interval scaling, the development of item banks for Computer Adaptive Testing, and potentially the linking of scores from different PROMs that measure the same construct (Belvedere and de Morton 2010; da Rocha et al., 2013; Turk et al., 2006).

To date, modern techniques have been used to re-develop or re-validate just a small number of existing, hearing-specific PROMs. For example, IRT has been applied to two measures of listening and communication difficulties that were originally developed using traditional techniques: the Amsterdam Inventory for Auditory Disability and Handicap and the Communication Profile for the Hearing Impaired (Mokkink et al. 2010a; Demorest et al. 2011; Chenault et al. 2013; Hospers et al. 2016). Also, Rasch analysis has been applied to the Tinnitus Functional Index (Fackrell et al. Reference Note 1; Meikle et al. 2012). It appears that even fewer hearing-specific questionnaires were originally developed using modern psychometric analysis techniques. Exceptions include the following pediatric questionnaires: the Classroom Communication Ease Scale (Garrison et al. 1994) and the Evaluation of Children's Listening and Processing Skills questionnaire (Barry and Moore, 2014; Barry et al. 2015).The SPaRQ is one of the first PROMs targeted at adults with hearing loss to have been created through the use of a modern psychometric analysis technique.

Conclusions

This research has produced the SPaRQ: a new 19-item PROM with strong psychometric properties that could be used in research or clinical practice to evaluate the impact of auditory rehabilitation interventions on participation restrictions in adults with hearing loss. This

research also highlighted the benefits of using modern psychometric analysis techniques in conjunction with more traditional approaches in order to develop high-quality PROMs. Research is currently underway to further validate the SPaRQ, including investigations of its responsiveness and interpretability.

ACKNOWLEDGEMENTS

The authors would like to thank Sandra Smith (NIHR Nottingham BRC, University of Nottingham) for her assistance with data management. We would also like to thank Mike Horton (Psychometric Laboratory for Health Sciences, Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds) for providing Rasch analysis training and guidance. We are grateful to Richard Nicholson, Julie Brady, Asma Hussain, Julie Carr, Karenbir Basi, Huw Cooper, the staff of the Nottingham Audiology Service, and the staff of the Queen Elizabeth Hospital Audiology Centre in Birmingham for their assistance with participant identification. Finally, we would like to thank the participants for taking the time to be part of this research.

This paper presents independent research supported by the NIHR Biomedical Research Unit Programme. The views expressed in this paper are those of the authors and not necessarily those of the NHS, the NIHR, or the UK Department of Health and Social Care.

The authors have no conflicts of interest to disclose.

Please contact hearingnihr@nottingham.ac.uk to obtain a copy of the Social Participation Restrictions Questionnaire (SPaRQ).

Received August 31, 2017; accepted April 10, 2018.

REFERENCES

Andrich, D., Lyne, A., Sheridan, B., et al. (2010). RUMM 2030. Perth: RUMM Laboratory.

Barry, J. G., Tomlin, D., Moore, D. R., et al. (2015). Use of questionnaire-based measures in the assessment of listening difficulties in school-aged children. *Ear Hear*, 36(6), e300-e313.

Barry, J. G. & Moore, D. R. (2014). *Evaluation of children's listening and processing skills (ECLiPS)*. London: MRC-T.

Beck, L. B. (2000). The role of outcomes data in health-care resource allocation *Ear Hear*, 21(4), 89S-96S.

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *J Clin Epidemiol*, 63(12), 1287-1297.

Bentler, R., & Kramer, S. (2000). Guidelines for choosing a self-report outcome measure. *Ear Hear*, 21(4), 37S-49S.

Bishop, F. L., Lewis, G., Harris, S., et al.(2010). A within-subjects trial to test the equivalence of online and paper outcome measures: the Roland Morris Disability Questionnaire. *BMC Musculoskelet Disord*, 11(1), 1-7.

Bonsaksen, T., Kottorp, A., Gay, C., et al. (2013). Rasch analysis of the General Self-Efficacy Scale in a sample of persons with morbid obesity. *Health Qual Life Outcomes*, 11(1), 202.

Boothroyd, A. (2007). Adult aural rehabilitation: what is it and does it work? *Trends Amplif*, 11(2), 63-71.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, 1-21.

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

Cappelleri, J. C., Lundy, J..J, & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther*, 36, 648-662.

Chachamovich, E., Fleck, M. P., Trentini, C. M., et al. (2008). Development and validation of the Brazilian version of the Attitudes to Aging Questionnaire (AAQ): An example of merging classical psychometric theory and the Rasch measurement model. *Health Qual Life Outcomes*, 6(5), 1-10.

Chen, W.-H., Lenderking, W., Jin, Y., et al. (2013). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Qual Life Res*, 23(2), 485-493.

Chenault, M., Berger, M., Kremer, B., et al. (2013). Quantification of experienced hearing problems with item response theory. *Am J Audiol*, 22(2), 252-262.

Chisolm, T. H., Abrams, H. B., McArdle, R., et al. (2005). The WHO-DAS II: psychometric properties in the measurement of functional health status in adults with acquired hearing loss. *Trends Amplif*, 9(3), 111-126.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychol Assess*, 7, 309-319.

Cook, A. J., Roberts, D. A., Henderson, M. D., et al. (2004). Electronic pain questionnaires: A randomized, crossover comparison with paper questionnaires for chronic pain assessment. *Pain*, 110(1), 310-317.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *J Appl Psychol*, 78(1), 98-104.

da Rocha, N. S., Chachamovich, E., de Almeida Fleck, M. P., et al. (2013). An introduction to Rasch analysis for psychiatric practice and research. *J Psychiatr Res*, 47(2), 141-148.

Danermark, B., Granberg, S., Kramer, S. E., et al. (2013). The creation of a comprehensive and a brief core set for hearing loss using the international classification of functioning, disability and health. *Am J Audiol*, 22(2), 323-328.

Darzins, S., Imms, C., Di Stefano, M., et al. (2014). Evaluation of the internal construct validity of the Personal Care Participation Assessment and Resource Tool (PC-PART) using Rasch analysis. *BMC Health Serv Res*, 14, 543.

Davis, A., Smith, P., Ferguson, M., et al. (2007). Acceptability, benefit and costs of early screening for hearing disability: A study of potential screening tests and models. *Health Technology Assessment*, 11(42).

De Vet, H. C., Terwee, C. B., Mokkink, L. B., et al. (2011). *Measurement in Medicine: A Practical Guide*. Cambridge: Cambridge University Press.

Demorest, M. E., Wark, D. J., & Erdman, S. A. (2011). Development of the Screening Test for Hearing Problems. *Am J Audiol*, 20(2), 100-110.

Ferketich, S. (1991). Focus on psychometrics. Aspects of item analysis. *Res Nurs Health*, 14(2), 165-168.

Ferguson, M. A., Kitterick, P. T., Chong, L. Y. et al. (2017). Hearing aids for mild to moderate hearing loss in adults. *Cochrane Database Syst Rev*, 9.

Garrison, W., Long, G., & Stinson, M. (1994). The classroom communication ease scale: Development of a self-report questionnaire for mainstreamed deaf students. *Am Ann Deaf*, 139(2), 132-140.

Glass, T. A., De Leon, C. F. M., Bassuk, S. S., et al. (2006). Social engagement and depressive symptoms in late life longitudinal findings. *J Aging Health*, 18(4), 604-628.

Gopinath, B., Hickson, L., Schneider, J., et al. (2012). Hearing-impaired adults are at increased risk of experiencing emotional distress and social engagement restrictions five years later. *Age Ageing*, 41(5), 618-623.

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.

Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of Electronic and Paper-and-Pencil Administration of Patient-Reported Outcome Measures: A Meta-Analytic Review. *Value Health*, 11(2), 322-333.

Hagquist, C. (2001). Evaluating composite health measures using Rasch modelling: An illustrative example. *Soz Praventivmed*, 46(6), 369-378.

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(3), 380-393.

Hawthorne, G. (2008). Perceived social isolation in a community sample: its prevalence and correlates with aspects of peoples' lives. *Soc Psychiat Epidemiol*, 43(2), 140-150.

Heffernan, E., Coulson, N. S., Henshaw, H., et al. (2016). Understanding the psychosocial experiences of adults with mild-moderate hearing loss: An application of Leventhal's self-regulatory model. *Int J Audiol*, 55(S3), S3-S12.

Hobart, J., & Cano, S. (2008). Rating scales for clinical studies in neurology - Challenges and opportunities. *US Neurology*, 4(1), 12-18.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment* 2009, 13(12), 1-200.

Hobart, J. C., Cano, S. J., Warner, T. T., et al. (2012). What sample sizes for reliability and validity studies in neurology? *J Neurol*, 259(12), 2681-2694.

Hobart, J. C., Cano, S. J., Zajicek, J. P., et al. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol*, 6(12), 1094-1105.

Hospers, J. M. B., Smits, N., Smits, C., et al. (2016). Reevaluation of the Amsterdam Inventory for Auditory Disability and Handicap Using Item Response Theory. *J Speech Lang Hear Res*, 59(2), 373-383.

Kerper, L., Spies, C., Tillinger, J., et al. (2014). Screening for depression, anxiety and general psychological distress in preoperative surgical patients: A psychometric analysis of the Patient Health Questionnaire 4 (PHQ-4). *Clin Health Promot*, 4(1), 5-14.

Kim, J.-I., Long, J. D., Mills, J. A., et al. (2015). Performance of the 12-item WHODAS 2.0 in prodromal Huntington disease. *Eur J Hum Genet*, 23(11), 1584.

Kline, P. (2013). *Handbook of Psychological Testing*. London: Routledge.

Kramer, S. E., Kapteyn, T. S., Kuik, D. J., et al. (2002). The association of hearing impairment and chronic diseases with psychosocial health status in older age. *J Aging Health*, 14(1), 122-137.

Kroenke, K., Spitzer, R. L., Williams, J. B. W., et al. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ–4. *Psychosomatics*, 50(6), 613-621.

Krosnick, J. A. & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of Survey Research* (pp. 263-314). Emerald Group Publishing.

Kurtaiş, Y., Őztuna, D., Küçükdeveci, A. A., et al. (2011). Reliability, construct validity and measurement potential of the ICF comprehensive core set for osteoarthritis. *BMC Musculoskelet Disord*, 12, 255-255.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7, 328.

Löwe, B., Wahl, I., Rose, M., et al. (2010). A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *J Affect Disord*, 122(1), 86-95.

Luciano, J. V., Ayuso-Mateos, J. L., Aguado, J., et al. (2010). The 12-item World Health Organization disability assessment schedule II (WHO-DAS II): a nonparametric item response analysis. *BMC Med Res Methodol*, 10(1), 45.

Lundgren Nilsson, Å., & Tennant, A. (2011). Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. *J Rehabil Med*, 43(10), 884-891.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Meikle, M.B., Henry, J. a, Griest, S.E., et al. (2012). The Tinnitus Functional Index: development of a new clinical measure for chronic, intrusive tinnitus. *Ear Hear,* 33(2), 153-76.

Mokkink, L. B., Knol, D. L., van Nispen, R. M., et al. (2010a). Improving the quality and applicability of the Dutch scales of the Communication Profile for the Hearing Impaired using item response theory. *J Speech Lang Hear Res,* 53(3), 556-571.

Mokkink, L. B., Terwee, C. B., Knol, D. L., et al. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol,* 10(1), 22.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., et al. (2012). *The COSMIN Checklist Manual.* Amsterdam: VU University Medical Center.

Nachtegaal, J., Smit, J. H., Smits, C., et al. (2009). The association between hearing status and psychosocial health before the age of 70 years: results from an internet-based national survey on hearing. *Ear Hear,* 30(3), 302-312.

Nemes, J. (2003). Despite benefits of outcomes measures, advocates say they're underused. *The Hearing Journal*, 56(8), 19-20.

Oppenheim, A. N. (2000). *Questionnaire Design, Interviewing and Attitude Measurement*. London: Bloomsbury Publishing.

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), 1-18.

Pesudovs, K., Burr, J. M., Harley, C., et al. (2007). The development, assessment, and selection of questionnaires. *Optom Vis Sci*, 84(8), 663-674.

Petrillo, J., Cano, S.J., McLeod, L.D., et al. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health*, 18(1), 25-34.

Post, M. W., de Witte, L. P., Reichrath, E., et al. (2008). Development and validation of IMPACT-S, an ICF-based questionnaire to measure activities and participation. *J Rehabil Med*, 40(8), 620-627.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes*, 1(1), 1-13.

Randall, M., Imms, C., Carey, L. M., & Pallant, J. F. (2014). Rasch analysis of the Melbourne assessment of unilateral upper limb function. *Dev Med Child Neurol*, 56(7), 665-672.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. London: Routledge.

Reise, S. P., Waller, N. G. & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychol Assess*, 12, 287.

Saunders, G. H., Chisolm, T. H., & Abrams, H. B. (2005). Measuring hearing aid outcomes - Not as easy as it seems. *J Rehabil Res Dev*, 42(4), 157-168.

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury*, 42(3), 236-240.

Seekins, T., Shunkamolah, W., Bertsche, M., et al. (2012). A systematic scoping review of measures of participation in disability and rehabilitation research: A preliminary report of findings. *Disabil Health J*, 5, 224-232.

Smith, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *J Appl Meas*, 3(2), 205-231.

Stewart, B. J. & Archbold, P. G. (1993). Nursing intervention studies require outcome measures that are sensitive to change: Part two. *Res Nurs Health*, 16, 77-81.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess*, 80(1), 99-103.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res*, 57(8), 1358-1362.

Tennant, A., & Pallant, J. (2006). Unidimensionality matters! (A tale of two smiths?). *Rasch Measurement Transactions*, 20(1), 1048–1051.

Terwee, C. B., Bot, S. D., de Boer, M. R., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*, 60(1), 34-42.

Thorén, E. S., Andersson, G., & Lunner, T. (2012). The use of research questionnaires with hearing impaired adults: Online vs. paper-and-pencil administration. *BMC Ear Nose Throat Disord*, 12, 12.

Trafimow, D. (2013). Are measurement theories falsifiable, and should we care? *Theory & Psychology*, 23, 397-400.

Turk, D. C., Dworkin, R. H., Burke, L. B., et al. (2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain*, 125(3), 208-215.

Üstün, T., Kostanjsek, N., Chatterji, S., et al. (2010a). *Measuring Health and Disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0.* Geneva: World Health Organization.

Üstün, T. B., Chatterji, S., Kostanjsek, N., et al. (2010b). Developing the World Health Organization disability assessment schedule 2.0. *Bull World Health Organ*, 88(11), 815-823.

Vallejo, M. A., Jordán, C. M., Díaz, M. I., et al. (2007). Psychological assessment via the internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). *J Med Internet Res*, 9(1), e2.

van der Velde, G., Beaton, D., Hogg-Johnston, S., et al. (2009). Rasch analysis provides new insights into the measurement properties of the neck disability index. *Arthritis Care Res*, 61(4), 544-551.

Ventry, I. M., & Weinstein, B. E. (1982). The hearing handicap inventory for the elderly: A new tool. *Ear Hear*, 3(3), 128-134.

Weinstein, B. E., Spitzer, J. B., & Ventry, I. M. (1986). Test-retest reliability of the Hearing Handicap Inventory for the Elderly. *Ear Hear*, 7(5), 295-299.

Weinstein, B. E., & Ventry, I. M. (1983). Audiometric correlates of the hearing handicap inventory for the elderly. *J Speech Hear Disord*, 48(4), 379-384.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educ Psychol Meas*, 64(6), 956-972.

Whiteneck, G. G., Dijkers, M. P., Heinemann, A. W., et al. (2011). Development of the Participation Assessment With Recombined Tools–Objective for Use After Traumatic Brain Injury. *Arch Phys Med Rehabil*, 92(4), 542-551.

Wong, L., & Hickson, L. (2012). *Evidence-Based Practice in Rehabilitative Audiology Evaluating: Interventions for Children and Adults With Hearing Impairment*. San Diego: Plural Publishing.

World Health Organization. (2013). *How to use the ICF: A Practical Manual for using the International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis, *Struct. Equ. Modeling*, 3(1), 3-24.

Yorke, J., Horton, M., & Jones, P. W. (2012). A critique of Rasch analysis using the Dyspnoea-12 as an illustrative example. *J Adv Nurs*, 68(1), 191-198.

REFERENCE NOTE

Fackrell, K., Hall, D.A., Barry, J.G., et al. (July 25, 2015). Measurement properties of the eight subscales in the Tinnitus Functional Index (TFI) questionnaire evaluated using Rasch analysis. *13th European Conference on Psychological Assessment (ECPA)*, Zurich, Switzerland.

Heffernan, E., Barry, J., Coulson, N., et al. (September 21, 2016). 'The evaluation of a new outcome measure to assess social participation in adults with mild-moderate hearing loss', *33rd World Congress of Audiology*, Vancouver, Canada.

Table 1. Demographic information of the study 1 and study 2 participants

	Study 1		*Study 2*	
Gender	*n*	*%*	*n*	*%*
Male	156	55.9	54	52.9
Female	123	44.1	48	47.1
Age	*Years*		*Years*	
Mean	65.67		60.92	
SD	12.73		10.53	
Range	24-89		25-90	
Hearing aid use	*n*	*%*	*n*	*%*
Everyday	176	63.1	63	61.8
Sometimes	40	14.3	25	24.5
Never	63	22.6	14	13.7

Table 2. Item fit statistics for the SPaRQ subscales.

Social Behaviors Subscale					
Item	*Item Summary*	*Logit Position*[1]	*Fit Residual*[2]	*Chi-Square*	*p*[3]
6	Carry out pastimes	0.42	2.38	5.68	0.129
8	Follow talk or lecture	-0.5	1.74	1.23	0.745
11	Manage stressful situations	0.25	0.42	1.94	0.586
12	Take part in a group discussion	-0.48	-2.47	6.68	0.083
14	Persevere with lengthy conversations	-0.11	-1.7	6.5	0.090
16	Talk with staff	-0.17	0.08	3.46	0.327
17	Manage responsibilities	0.52	-0.8	4.93	0.177
19	Attend get-togethers	0.24	-0.17	1.68	0.642
23	Converse with friends	-0.18	0.62	4.69	0.196
Social Perceptions Subscale					
Item	*Item Summary*	*Logit Position*[1]	*Fit Residual*[2]	*Chi-Square*	*p*[3]
28	Find social gatherings stressful	-0.12	0.63	3.03	0.387
29	Worry about missing sounds	-0.57	1.19	7.61	0.055
30	Isolated during group conversations	-0.33	-1.82	8.27	0.041
32	Frustrated by being left out	-0.3	1.26	0.17	0.983
34	Lonely, even around others	0.76	0.15	8.35	0.039
36	Isolated at get-togethers	0.44	0.29	6.58	0.087
38	Unenthusiastic about conversations	-0.43	1.38	0.51	0.916
39	Upset during conversations	-0.06	0.77	2.35	0.503
41	Worry about talking to strangers	0.53	-1.23	6.12	0.106
51	Look foolish by saying the wrong thing	0.08	1.89	3.27	0.352

[1] Negative values indicate low severity and positive values indicate high severity.
[2] The acceptable range for fit residuals = ±2.5.
[3] The Bonferroni adjusted α=0.005 for chi-square statistic should be non-significant.

Table 3. Spearman's Rank Order correlation coefficients to test convergent validity for the SPaRQ subscales.

Scale	*Social Behaviors Subscale*		*Social Perceptions Subscale*	
	r_s	*p*	r_s	*p*
HHIE overall scale	0.692	<.001	0.793	<.001
HHIE situational subscale	0.656	<.001	0.712	<.001
HHIE emotional subscale	0.664	<.001	0.788	<.001
WHODAS 2.0 simple scoring	0.511	<.001	0.454	<.001
WHODAS 2.0 complex scoring	0.492	<.001	0.442	<.001
PHQ4 overall scale	0.430	<.001	0.474	<.001
PHQ4 anxiety subscale	0.388	<.001	0.464	<.001
PHQ4 depression subscale	0.412	<.001	0.410	<.001

Table 4. Internal consistency statistics for the SPaRQ subscales.

Subscale	*Social Behaviors*	*Social Perceptions*
Cronbach's alpha[1]	0.941	0.950
Mean inter-item correlation[2]	0.639	0.655
Mean corrected item-total correlation[3]	0.772	0.786

[1] Value should fall between 0.7-0.95
[2] Value should fall between 0.3-0.7,
[3] Value should be >0.3

Figure 1. Person-Item Threshold Distribution graphs for the Social Behaviors and Social Perceptions subscales showing the distribution of persons (upper half) and items thresholds (lower half). The x-axes display the level of severity on a logit scale, y-axes display the frequencies.