OPEN

Refinement and Validation of the Empowerment Audiology Questionnaire: Rasch Analysis and Traditional Psychometric Evaluation

Rebecca J. Bennett,[1,2,3,4] Josefina Larsson,[5] Sarah Gotowiec,[5] and Melanie Ferguson[2,6]

Objectives: Empowerment is the process through which individuals with hearing-related challenges acquire and use knowledge, skills and strategies, and increase self-efficacy, participation, and control of their hearing health care, hearing solutions, and everyday lives. The aim was to refine and validate the Empowerment Audiology Questionnaire (EmpAQ), a hearing-specific measure of empowerment. This was achieved through (1) refinement via Rasch analysis (study 1), and (2) traditional psychometric analysis of the final survey structure (study 2).

Design: In study 1, 307 adult hearing aid owners completed the initial empowerment measure (33 items) online. To inform an intended item reduction, Rasch analysis was used to assess a range of psychometric properties for individual items. The psychometric properties included analysis of individual items (e.g., response dependency, fit to the polytomous Rasch model, threshold ordering) and the whole EmpAQ (e.g., dimensionality). Item reduction resulted in a 15-item version (EmpAQ-15) and a short-form 5-item version (EmpAQ-5), validated using modern (Rasch), and traditional (Classical Test Theory) psychometric analysis (study 2). In study 2, 178 adult hearing aid owners completed the EmpAQ-15 and EmpAQ-5, alongside 5 questionnaires to measure related constructs. These included two hearing-specific questionnaires (Social Participation Restrictions Questionnaire and Self-Assessment of Communication), two general health-related questionnaires (Patient Activation Measure and World Health Organization Disability Assessment Schedule 2.0), and a general empowerment questionnaire (Health Care Empowerment Questionnaire). Modern (Rasch) and traditional psychometric analysis techniques (internal consistency, construct validity, and criterion validity) were used to assess the psychometric properties of the EmpAQ-15 and EmpAQ-5.

Results: Rasch analysis of the initial 33-item measure of empowerment identified 18 items with high response dependency, poor fit to the Rasch model, and threshold disordering, which were removed, resulting in a long-form (EmpAQ-15) hearing-specific measure of empowerment. A short-form (EmpAQ-5) version was developed for use in the clinic setting. Validation of the two EmpAQ measures using Rasch analysis showed good item fit to the Rasch model, appropriate threshold targeting, and the existence of unidimensionality. Traditional psychometric evaluation showed that both questionnaires had high internal consistency and positive correlations with the hearing-specific questionnaires. However, in contrast with our hypotheses, correlations with general health questionnaires were stronger than with hearing-specific questionnaires; all questionnaires were correlated with the EmpAQ and in the direction hypothesized. Taken together, these findings support the construct validity of the EmpAQ-15 and EmpAQ-5.

Conclusions: The EmpAQ-15 and EmpAQ-5 are the first self-report measures to be developed specifically for the measurement of empowerment. The EmpAQ-15 and EmpAQ-5 were found to meet the Rasch model criteria for interval-level measurements. Traditional psychometric evaluation supports the construct validity of both measures. The EmpAQ measures have the potential to be used in both research and clinical practice to evaluate empowerment along the hearing journey. The next stage of this research will be to further validate these measures by assessing their responsiveness, minimal clinically important difference, and clinical interpretability in a clinical population.

Key words: Empowerment, Hearing aid, Hearing loss, Psychometric, Questionnaire development, Rasch, Self-report measure, Validation.

(Ear & Hearing 2024;45;583–599)

INTRODUCTION

Empowerment is the granting of the power, right, or authority to perform various acts or duties. Within the healthcare context, empowerment is defined as "an individuals' capacity to make decisions about their health (behavior) and to have, or take control over aspects of their lives that relate to health" (McAllister et al. 2012). Empowerment has gained prominence in healthcare as service delivery has moved from a biomedical model of care to a biopsychosocial model of care, delivering holistic, equitable, and collaborative healthcare (McAllister et al. 2012). The process of empowerment can be driven by others or by oneself. For example, patients can be empowered by their healthcare providers through education, counseling, and patient-centered care, or patients can empower themselves through self-education, help-seeking, or by participating in patient/community organizations or activism (Holmström & Röing 2010). Empowerment as it relates to healthcare is not a uniform experience and not all people want to be, or can be, empowered at all times (McAllister et al. 2012). In the case of acute care, some patients may prefer their doctor to make treatment decisions for them, at least in the short-term. Conversely, in the case of chronic health conditions, it is untenable for patients to rely on clinicians for all care decisions and processes. Consequently, modern healthcare seeks to empower people with chronic conditions to feel confident to self-manage their own health over the long-term (McAllister et al. 2012).

Research shows that empowered patients have a greater understanding of how to navigate the healthcare system (Khuntia et al. 2017), experience improved health outcomes

[1]Brain and Hearing, Ear Science Institute Australia, Perth, Australia; [2]Curtin enAble Institute, Curtin enAble Institute, Curtin University, Perth, Australia; [3]School of Medicine, The University of Western Australia, Perth, Australia; [4]National Acoustic Laboratories, Sydney, Australia; [5]ORCA Europe, WS Audiology, Stockholm, Sweden; and [6]School of Allied Health, Curtin University, Perth, Australia.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

<zdoi; 10.1097/AUD.0000000000001449>

(Yeh et al. 2016), and are more satisfied with the healthcare they receive (Yeh et al. 2018). While a range of self-report measures have been developed for the measurement of empowerment (including nonspecific scales, and those developed for a particular condition), literature reviews highlight important gaps in scale development processes (Barr et al. 2015; Cyril et al. 2016). Guidelines exist to support the development and validation of patient-reported outcome measures (PROMs) (Mokkink et al. 2019; Gagnier et al. 2021), yet assessment of patient empowerment PROMs against these guidelines reveal common flaws relating to (1) lack of a comprehensive process to ensure that the PROM encapsulates all empowerment domains pertinent to a specific ailment or setting; (2) failure to use a combination of modern and Classical Test Theory (CTT) to guide PROM refinement and validation; and (3) failure to include analysis of many psychometric properties, or where these properties were tested, there was limited evidence to support reliability and validity (Barr et al. 2015). This aligns well with a growing awareness of the need to develop measures that conform to best practice guidelines (Mokkink et al. 2019), with the "patient voice" firmly embedded in the development of such measures (Heffernan et al. 2018a; Hughes et al. 2021; Allen et al. 2022). Given the growing importance of empowerment in the audiology literature and the lack of a measure of patient empowerment specific to the audiology context, we set out to develop a hearing-specific measure of empowerment on the hearing health journey. The foundational research and development of the Empowerment Audiology Questionnaires (EmpAQ-5 and EmpAQ-15) used participatory methods to first conceptualize empowerment on the hearing journey with representatives of both people living with hearing loss and those who work with them, and the process of development and validation of the EmpAQs has closely followed evidence-based guidelines (Mokkink et al. 2019; Gagnier et al. 2021).

Empowerment as it relates to the management of chronic hearing loss has only recently emerged as a byproduct of research findings (Laplante-Lévesque et al. 2013; Bennett et al. 2019; Maidment et al. 2019, 2020; Bennett et al. 2021; Gomez et al. 2021). Qualitative studies of adults with hearing loss have emphasized the importance of empowerment with respect to acquiring hearing devices (Poost-Foroosh et al. 2011); trouble shooting problems that arise with hearing devices (Bennett et al. 2019); use of smartphone-connected hearing devices (Maidment et al. 2019; Gomez et al. 2021); and multimedia educational resources for hearing aid owners (Maidment et al. 2020). Our recent qualitative investigation identified five dimensions conceptualizing empowerment on the hearing health journey from first discovery of hearing challenges through to becoming an active hearing aid user: knowledge, skills and strategies, participation, self-efficacy, and control (Gotowiec et al. 2022). Knowledge refers to the acquisition and assimilation of information, leading to an understanding of an individual's hearing, hearing-related challenges and hearing solutions. Skills and strategies refer to the acquired ability to do something well. Participation is defined as the active involvement in both hearing rehabilitation and all aspects of social life, including family and informal social relationships, and encompasses decisions, processes, and actions relating to the hearing healthcare. Self-efficacy refers to the belief in ones' ability to successfully manage hearing-related challenges and hearing solutions. Control refers to a sense of power to influence and manage hearing-related challenges and hearing solutions in everyday life. Overall, we define hearing-specific empowerment as the process through which individuals with hearing-related challenges acquire and use knowledge, skills, and strategies, and increase self-efficacy, participation, and the feeling of control of their hearing health care, hearing solutions, and everyday lives. The value of developing a questionnaire to measure the domains of hearing-specific empowerment is multifaceted. Such a measure could complement clinical insights and inform patient outcomes, be used to evaluate the magnitude of intervention effects, or to explore the mediators and moderators of empowerment.

Building on our previous research, this study aimed to develop a measure of empowerment following international consensus-based standards for the development and validation of self-report measures (Mokkink et al. 2019; Gagnier et al. 2021). We followed a five-phase process of scale development (Brod et al. 2009; Patrick et al. 2011; Heffernan et al. 2019): (1) concept elicitation and conceptualization; (2) item generation; (3) content validation; (4) modern psychometric evaluation based on Item Response Theory (IRT) (item refinement); and (5) traditional psychometric evaluation of the final structure based on CTT (validation). Phase 1 used qualitative interviews to explore (a) how empowerment manifests itself from individuals' first awareness of hearing loss through to hearing aid fitting and then to becoming an active hearing aid user, (b) identify points when the different dimensions of empowerment are most relevant, and (c) conceptualize empowerment (Gotowiec et al. 2022). Phase 2 drew on the interview data to generate a pool of potential items for a measure of empowerment (Gotowiec et al. 2023). Phase 3 used cognitive interviews with adult hearing aid users and expert panel review to evaluate the content validity of the pool of items and develop an initial draft self-report measure of empowerment (33 items) (Gotowiec et al. 2023). Here, we focus on phases 4 and 5 to further refine the item pool, finalize the PROMs, and evaluate their psychometric properties.

Approaches to Psychometric Evaluation

Psychometric evaluation aims to establish whether the questionnaire's conceptualization of the target construct/variable of interest has been successfully captured by the set of items. Psychometric evaluation of self-report measures is driven by two schools of thought: CTT and IRT (Hambleton & Jones 1993). CTT is the traditional, more common, approach used to evaluate reliability and validity of a scale as it examines how measurement error affects rating scale scores (Cappelleri et al. 2014). Example applications include factor analysis for item reduction, and Cronbach alpha as a measure of internal consistency reliability. IRT is a modern approach for examining the pattern of responses that respondents make to the items (Cano & Hobart 2011). Example applications include Rasch analysis to guide item reduction and explore dimensionality. The key differences between CTT and IRT are (1) CTT focuses on the overall score obtained by an individual on a test or scale, whereas IRT focuses on the response patterns to individual test items; (2) CTT assumes that all respondents have the same level of ability for completing the measure, while IRT assumes that respondents have varying levels of ability; and (3) CTT measures the difficulty and discrimination of items based on the overall performance of respondents, whereas IRT measures

these parameters based on the responses of individual respondents (Cano & Hobart 2011). Modern IRT techniques are being increasingly reported alongside traditional CTT analyses in studies of PROM development and validation (Gagnier et al. 2021), including in hearing research (Heffernan et al. 2018b; Hughes et al. 2021). The benefits of doing so includes both improved item analysis, as IRT can provide more detailed information about the psychometric properties of individual items, such as their difficulty and discrimination parameters. This can be used to improve the overall quality of the measure, and improved test construction, as IRT can help developers to create measures with items that are well-matched to the ability levels of respondents, which can improve the accuracy and fairness of the test.

The Rasch model is considered to be a variant of IRT from the IRT perspective, but is considered distinct from IRT from the Rasch model perspective (Andrich 2011). The orientation of IRT is to find a model that best accounts for the data, and if the simplest of the possible models does not work, then one with more parameters is tested. In general, the relevant Rasch model is the simplest model in terms of the number of parameters that can be considered. In contrast, in Rasch Measurement Theory, the relevant Rasch model specifies a criterion for measurement to have been achieved (Andrich 2011). Rasch Measurement Theory as a psychometric method is increasingly used alongside CTT to develop and validate self-report measures (Aryadoust et al. 2019). Rasch analysis is the term used to describe the formal evaluation of a self-report measure against Rasch mathematical measurement model. Rasch analysis uses a probabilistic model to evaluate the measurement properties of rating scales (Andrich 2011). The Rasch model for ordered categories, such as Likert rating scales, is based on the assumption that the probability of a response is governed by a person's ability on the variable and the difficulty of the item. In Rasch analysis, estimates of ability (e.g., an individual's degree of empowerment) and item severity (e.g., the degree of empowerment evaluated by an item) are obtained. These estimates can be interpreted to the degree in which the responses fit the Rasch model. The process of checking fit to the model provides an opportunity to evaluate the psychometric properties of PROMs (Andrich 2011). For example, Rasch analysis can be used to identify items that could be removed or rewritten to improve the performance of the questionnaire. It can also be used to establish whether an item's response scale is functioning as expected, helping to guide decisions around optimal response scales for a PROM. Rasch analysis also provides an effective tool for exploring potential response bias, a systematic tendency for research participants to respond to a self-report survey in a way that is not reflective of their true attitudes, beliefs, or behaviors. This can help identify which items contribute most to response bias (Bradley et al. 2015). Any discrepancies between the data and the Rasch model requirements are indicative of anomalies in the responses to the PROM as a measurement instrument. These discrepancies provide diagnostic information to assist understanding and empirical improvement of the questionnaire at both the item and the scale level (Hobart & Cano 2009). If data are found to conform to the Rasch model, then PROM developers can theoretically be confident that an individual's responses accurately reflect their location on a continuum that measures the construct under investigation (e.g., low to high degrees of empowerment) (Hobart & Cano 2009).

This article describes refinement of a measure of empowerment via Rasch analysis (study 1), and validation of the final PROM structure using both modern (Rasch) and traditional (CTT) psychometric analysis (study 2). We aimed to first reduce the number of items that emerged (33 items) following the content evaluation of the initial measure of empowerment as participants explicitly requested the measure to be shorter (Gotowiec et al. 2023), as well as the likelihood that a shorter measure would more likely be used in a research or clinical context in the future. Furthermore, we wanted to ensure that we had the best-quality items in the PROM, which was a consequence of the item reduction. We also aimed to undertake a detailed psychometric assessment of the final PROM(s) using both Rasch and CTT analyses (Mokkink et al. 2019; Gagnier et al. 2021). Rasch analysis was used to explore dimensionality of the scale, assess the response format, suitability of the items and item bias, and reduce the number of items accordingly. CTT was used to assess internal consistency reliability and to explore the scale's association with existing measures and selected demographic characteristics (convergent and criterion validity).

METHODS AND RESULTS

Ethics approval was provided by the Swedish Ethical Review Authority (DNR: 2020-04562) and the Human Research Ethics Office of The University of Western Australia (2021/ET000766).

Study 1: Rasch Analysis (Item Refinement)

Methods • Rasch analysis was used to guide item reduction and refinement of the initial 33-item measure of empowerment. Specifically, the partial credit parameterization approach, wherein a separate set of threshold parameters were estimated for each item.

Materials • The initial version of the empowerment measure generated by Gotowiec et al. (2023), a relatively long, 33-item measure with a 6-point Likert scale format from strongly disagree to strongly agree. In addition, demographic questions (age, gender, years of hearing loss, impact of hearing loss, hearing device ownership, and daily hours of hearing aid use) were administered via an online survey (Qualtrics). Although it is common to "force" responses with electronic surveys, that is, not allow participants to move on to subsequent items until they have responded to all prior items, we opted not to "force" responses as we wished to explore how participants would interact with the items naturally, enabling us to measure the proportion of missing responses for each item.

Participants • Hearing aid owners were recruited from a hearing service provider in Western Australia. Inclusion criteria were (1) aged 18 years or older, and (2) used hearing aid(s) for a minimum of 6 months. Exclusion criteria were (1) self-reported nonfluency in written and spoken English, and (2) self-report cognitive decline or dementia that would require assistance in completing questionnaire items. The sample size was based on international guidelines for Rasch analysis that require a minimum sample size of 250 participants (Mokkink et al. 2019).

Procedure • Potential participants were identified from our partner clinic's client database and recorded on a spreadsheet in random order using a random number generator in Microsoft Excel. The first group of 50 potential participants were sent an invitation to participate in the study via e-mail, 24 hr later

TABLE 1. Description of the psychometric properties assessed within the Rasch analysis

Fit to the Rasch model	Fit to the model can be evaluated by examining the mean and SD of the fit residuals across all items. The fit residual is a standardized summation of the differences between observed and predicted scores for an item. A mean fit residual close to zero and an SD of approximately ≤1.5 for all items together is an indication of good fit to the Rasch model.
Item fit to the Rasch model	The item fit to the Rasch model can be described as a goodness-of-fit statistic evaluating the degree of discrepancy between the observed item performance and expected item performance, for an individual item. Item fit was assessed using three statistics available in RUMM2030: Fit residual: The fit residual measures the difference between the expected response to an item based on the Rasch model and the actual response from the participants (Tennant & Pallant 2006). The fit residual gives an indication of the magnitude and direction of misfit, with large positive values indicating low discrimination and large negative values indicating high discrimination relative to the average discrimination of all the items. Chi-square: This statistic measures the difference between the observed and expected response frequencies for an item. A significant Chi-square value indicates poor fit between the observed and expected responses, while a nonsignificant value indicates good fit. F statistic: This statistic measures the ratio of the Chi-square value to its DF. An F statistic value close to 1 indicates good fit, while values greater than 1 indicate poor fit.
Response dependency	Response dependency is where items are linked in some way, such that the response on one item is dependent on or determines the response to another item. Dependent items are considered redundant as they either replicate an item or do not provide additional important information.
Missing responses	When a sizeable number of participants fail to respond to an item, it can be indicative of serious flaws (i.e., irrelevance, ambiguity, or intrusiveness) and the item may need to be reworded or removed. Others have used a cut off of >15% missing responses to indicate the need for removal of individual items (Heffernan et al. 2018b).
Threshold ordering	Threshold ordering is used to evaluate the functioning of a PROM's rating scale. Threshold ordering was assessed using the location of the thresholds within each item on the common scale, with the expectation that higher score categories require more empowerment to endorse, and therefore the threshold between two high score categories should be located higher along the scale than the threshold between two lower score categories.
DIF	If an item is functioning differently for different groups of people, it is said to exhibit DIF. Rasch analysis is used to detect DIF in items by comparing item responses across different groups of people who have similar abilities or trait levels (e.g., gender). This allows researchers to identify items that may be functioning differently for certain groups, and to adjust or remove them as necessary to ensure fair and accurate measurement (Hagquist & Andrich 2017).
Targeting	Targeting refers to the precision with which a PROM can differentiate individuals (e.g., a person with high empowerment from a person with low empowerment) (Hagquist & Andrich 2017).
PSI	PSI is a measure of reliability. It indicates how well the Rasch model can distinguish between people with different levels of ability or proficiency. It has been recommended that the PSI value should be ≥0.7 for group use and ≥0.85 for individual use (Pallant & Tennant 2007).
Dimensionality	A PROM's dimensionality refers to the number and nature of the variables reflected in its items. It is important to identify whether the final measure developed is multidimensional (tapping into different domains) or unidimensional (tapping into a single domain) as this affects how the PROM is scored. Items for a unidimensional PROM (such as the EmpAQ-15) can be legitimately be summed.

DF, degrees of freedom; DIF, differential item functioning; EmpAQ-15, Empowerment Audiology Questionnaire-15; PROM, patient-reported outcome measure; PSI, person separation index.

another batch of 50 were dispatched, followed by another 24 hr after that. This process was repeated until the required minimum of 250 participants had been reached. A total of 1200 potential participants were invited to participate, with 356 starting the survey (response rate 29.67%). Although survey responses from 356 participants were collected, 49 participants consented to participate but did not commence the survey set (provide a response for at least one survey question), thus only 307 entries were included in the analysis.

E-mail invitations included a digital Participant Information Form (PIF) and a link to the electronic survey. The landing page of the survey provided an overview of the project and the PIF. All participants provided consent to participate by ticking a consent box within the survey before gaining access to the subsequent survey items. There were no fees or incentives for participation. The study 1 items took approximately 15 min to complete.

Data analysis • The Rasch measurement model was applied using RUMM2030 Software (Andrich et al. 2022). Reporting of Rasch analysis is in line with the Reporting Guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research (Van de Winckel et al. 2022). Table 1 describes the psychometric properties assessed.

Results • Data from the 307 usable surveys were included in the analysis. Participants ranged in age from 29 to 93 years (mean 72.55; SD 9.34). There were more male participants (61.56%; n = 189) than female (38.43%; n = 118). When asked *How much does your hearing loss impact your daily life?* 4.56% (n = 14) reported *Not at all*; 36.81% (n = 113) reported *Mildly*; 41.69% (n = 128) reported *Moderately*; and 16.93% (n = 52) reported *Significantly*. See Table 2 for demographic data.

Of the 307 participants, 3 selected the highest category for every survey question, giving them the maximum total score possible on the survey. As these entries looked to be valid based on the detailed demographic information and open text responses provided, the data from these 3 participants were analyzed but their estimate of empowerment needed to be extrapolated.

Missing data • The presence of missing data, where participants did not respond to one or more survey questions, was addressed using the estimation algorithm in the RUMM2030 software (Andrich & Luo 2003).

Iterative process of item reduction • Rasch analysis informed the intended item reduction and refinement of the PROM. Specifically, we looked at individual items' fit to the

TABLE 2. Demographic information for studies 1 and 2 sample

	Study 1	Study 2
Gender (n)		
Male	189	95
Female	118	82
Other	0	1
Age (yrs)		
Mean	72.55	74.0
SD	9.34	9.38
Range	29–93	40–92
Duration of hearing loss (yrs)		
Mean	n/a	13
SD	n/a	11.57
Range	n/a	1–70
Impact of hearing loss on daily life (n [%])		
Not at all	14 (4.56)	20 (11.24)
Mild	113 (36.81)	78 (43.82)
Moderate	128 (41.69)	55 (30.90)
Significant	52 (16.93)	25 (14.04)
Device use (n [%])		
Monaural (hearing aid)	21 (6.95)	14 (7.87)
Binaural (hearing aid)	278 (90.55)	157 (88.20)
Bimodal (hearing aid and hearing implant)	8 (2.65)	7 (3.93)
Time spent using device per day (hr)		
Mean	10.44	10.51
SD	4.87	4.82
Range	0–20	1–24
Duration of hearing device ownership (yrs)		
Mean	7.43	7
SD	8.58	6.83
Range	0.25–50	1–40

The question on duration of hearing loss was asked of study 2 participants and not study 1 participants. Demographic data were missing for two participants in study 2 (n = 178). "Other" responses regarding device usage were as follows: normally wear hearing devices but do not currently due to cost; normally wear hearing devices but do not currently due to difficulty when required to also use a face mask; transmitter; contralateral routing of signals. n/a, not applicable.

Rasch model, response dependency, proportion of missing responses, threshold ordering, and differential item functioning (DIF) (Table 1). The magnitude of the Rasch statistics and the wording of the items were considered when determining whether to reject or retain individual items.

Rasch analysis for the purpose of item reduction was conducted over three iterative rounds (Fig. 1). In the first round of the analysis of the responses, all 33 items were analyzed with interpretation of the data resulting in rejection of 14 items (19 items retained). In the second round, 19 items were reanalyzed and interpretation of the data resulted in rejection of 4 additional items (15 items retained). In the third and final round, the remaining 15 items were analyzed as well as a subset of 5 items and no items were rejected from either the 15 or 5 item sets. The majority of items rejected in rounds one and two were due to poor fit to the Rasch model, response dependency, or both (see Supplemental Digital Content 1, http://links.lww.com/EANDH/B249). Later, we provide a description of the reasons for item rejection with examples demonstrating how these decisions were made.

Person separation index • Where CTT uses coefficient alpha to evaluate the internal consistency of a test or assessment, the person separation index (PSI) too is a reliability index but based on Rasch model parameter estimates and their standard errors to estimate the variance of the true scores and standard errors. The PSI is a reliability index defined as the ratio of an estimated true variance to the total estimated variance, which is the sum of the true and error variances. PSI tells us how well the Rasch model can distinguish between people with different degrees of the construct under investigation. The PSI ranges from 0 to 1, where a high score is considered >0.7, indicating that the Rasch model is doing a good job of separating people with different degrees of empowerment. The PSI score was 0.938 for the first analysis of 33 items.

Fit to the Rasch model • Overall fit to the Rasch model refers to the extent to which the data from a set of items fit the assumptions of the Rasch model. In other words, it evaluates whether the observed data match what is expected by the model. The mean (0.49) and SD (3.49) of the fit residuals for all 33 items from the initial analysis before rescoring suggest reasonable fit to the Rasch model. The SD is larger than acceptable, largely due to the high fit residual scores for the two reversed items (see later for further discussion on these items). Recalculation following removal of the two reversed items demonstrates much improved scores (mean = −0.19; SD = 2.27).

Item fit to the Rasch model • The Rasch model is built on the underlying logic that participants are more likely to agree with easier items and less likely to agree with more difficult items (i.e., in the case of the EmpAQ, respondents who are more empowered are more likely to endorse EmpAQ items that represent higher degrees of empowerment). The item fit statistics (fit residual, Chi-square, and F statistic) provide information about whether participants perform on an individual item as expected according to the Rasch model. Interpretation of the fit statistics is typically done by comparing the observed values to predetermined thresholds. Generally, fit residuals should be within ±2.5, Chi-square values should be nonsignificant ($p > 0.05$), and F statistic values should be close to 1 (i.e., between 0.5 and 1.5). It is important to note that there are no absolute criteria for interpreting fit statistics and the earlier criteria are only provided as a guide to help identify potential problems (Andrich & Marais 2019). If an item has fit statistics that fall outside these ranges, it may indicate that the item is not measuring what it is intended to measure, or that there is some other problem with the item. In such cases, the item may need to be revised or removed from the test to improve its validity and reliability. However, fit statistics should not be interpreted in isolation, and data should be interpreted in conjunction with other sources of evidence such as item content, IRT model assumptions, and expert judgment. For example, item 20 *My hearing loss stops me from taking part in social activities* had a fit residual of 10.864, Chi-square of 192.061, and F statistic of 16.4692. Figure 2 shows the poor fit of the observed class interval means to the expected value curve. The curve represents the expected value, according to the model given the item's severity estimate, for a person's location (severity/empowerment estimate). The dots represent the observed class interval means. There are six class intervals here (based on the initial EmpAQ six-point Likert response options) of approximate equal sample size, with their calculated mean location displayed (the short vertical lines on the *x* axis). Each class interval's mean score for this item is shown on the *y* axis (observed value). In this sample, the highest-class interval (most empowered) has a lower mean score (response category chosen) than expected by the model (expected value

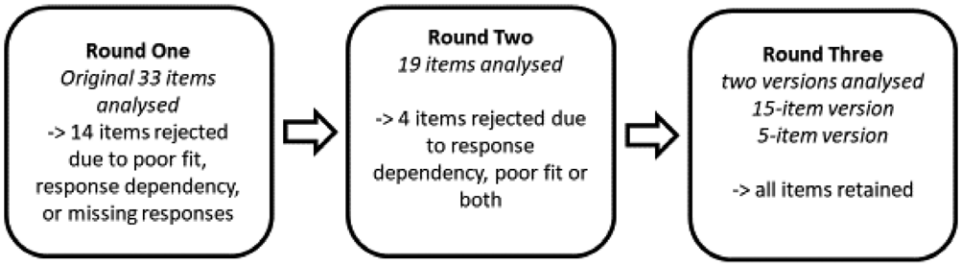

Fig. 1. Rasch analysis was conducted over three rounds.

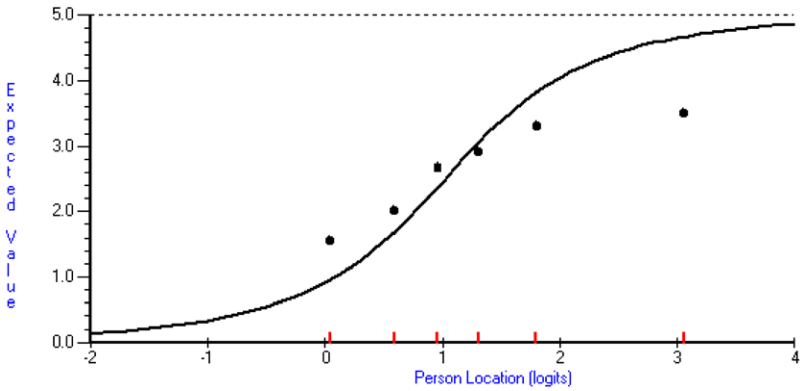

Fig. 2. The fit of item 20 to the Rasch model is shown. The curve represents the expected value for a person's location (severity/empowerment estimate) according to the model. The dots represent the observed class interval means. There are six class intervals here of approximate equal sample size (the short vertical lines on the *x* axis). Their mean score for this item is shown on the *y* axis (observed value). In this sample, the highest-class interval (most empowered) has a lower mean score (response category chosen) than expected by the model. The opposite is true of the lowest class interval.

curve). The opposite is true of the lowest class interval. This is referred to as low discrimination because the item does not discriminate between persons of different empowerment as well as the average discrimination of all the items. Put simply, how far the dots are from the curve, shows how different the actual responses were (within a class interval) from what is expected by the model. There are not cutoffs for the item fit statistics. However, when two items had high residual correlations with one another, indicating possible response dependence, the relative item fit statistics were used to guide the decision around whether to reject, retain or reword the item. In the case of item 20 *My hearing loss stops me from taking part in social activities*, other sources of evidence suggested that the wording of the item may have contributed to its misfit and in this instance the item was not rejected, but rather reworded (described in further detail later).

Response dependency • Response dependency was assessed by examining patterns within the item residual correlation matrix which indicates whether pairs of items are correlated after accounting for their expected correlation from assessing a common latent variable (Andrich & Marais 2019). There is no set cut off for an item residual correlation, however, researchers often investigate item residual correlations greater than 0.3 (Hobart & Cano 2009) to determine if the item pairs assess content from the same content domain and if there may be any other substantive reason why they were correlated. Items with high residual correlations (i.e., >0.3) were examined. We note an alternative approach that identifies local dependence (Christensen et al. 2017), which may be preferable to use to this rule of thumb approach used here. Along with the magnitude of the residual correlation, the other Rasch analysis statistics for the two items (e.g., fit to the Rasch model, threshold ordering) and the wording of the item, were taken into consideration when determining whether to reject or retain an individual item. For example, item 17 *I contact my hearing care professional whenever I need anything* and item 29 *I am confident talking with my hearing care professional about any problems I have with my hearing or hearing device(s)* had an item residual correlation of 0.352. There is no set cut off for an item residual correlation and researchers may use other variables and their own judgment to determine whether to reject or retain items with a residual correlation around 0.3. While we could have retained both items based on the item residual correlation, in this instance, item 29 demonstrated a poor fit to the Rasch model (round one: fit residual = −2.064; Chi-square = 16.601; *F* statistic = 4.264), yet item 17 demonstrated a good fit to the Rasch model (round one: fit residual = 0.066; Chi-square = 3.489; *F* statistic = 0.5674). As item 17 was a better fit to the Rasch model (i.e., 2 of the 3 fit statistics were closer to 0) and encompassed a broader concept (e.g., "need anything" as opposed to "problems I have with my hearing or hearing device(s)") we opted to reject item 29 and retain item 17.

Missing responses • One item was rejected due to having >15% missing responses (Heffernan et al. 2018b); 22% of participants did not respond to the item 19 *I ask my hearing care professional to give me information in another way when needed (e.g., written, diagrams, videos)* and it was thus rejected.

Threshold ordering • Threshold ordering is used to evaluate the functioning of a PROM's rating scale. A greater degree of empowerment should equate to higher scores on the EmpAQ.

Category probability curves are generated for each item to show the probability of observing each category according to the Rasch model. Disordering occurs when respondents select a response option that is inconsistent with their degree (in this case, of empowerment) and implies that a rating scale's categories may be confusing or difficult to use (Pallant & Tennant 2007). Disordered thresholds thus identify items that should be studied to understand why the categories are not working as intended as well as identifying potential issues with a rating scale if the majority of items demonstrate disordering (Andrich & Marais 2019). Rasch analysis of the initial EmpAQ 33-item PROM identified both individual item threshold disordering and broad disordering indicating issues with the rating scale functioning.

Threshold ordering—individual item functioning • Four items had reversed thresholds, indicating that the categories may not have been functioning as intended. All four of these items were rejected as they also demonstrated other issues (e.g., response dependency, poor fit to the Rasch model).

In addition, 2 of the original 33-item were worded such that they needed to be reverse scored. Threshold ordering suggested that participants may have responded inappropriately to these items, in that the reversal of the response options may have caused confusion. For the two reverse scored items, the frequency of responses in each score category did not reflect the pattern seen among the other items which suggests that some participants may not have identified that the wording of these items required negative scoring. While we chose to reject one of the items, due to poor fit to the Rasch model, high residual correlation (0.899) and similar wording of the two items, we opted to reject item 21 *My hearing loss stops me from feeling included in my social activities* and retain item 20 *My hearing loss stops me from taking part in social activities*. However, we decided to reword item 20 so that the responses were no longer reverse scored in relation to the other items (final wording *My hearing loss doesn't stop me from taking part in social activities*). The new item wording was evaluated in the stage 2 psychometric testing.

Threshold ordering—rating scale functioning • The majority of items showed threshold disordering, in that participants did not appear to be using the categories of slightly disagree and slightly agree as frequently as would be expected based on their overall degree of empowerment. Specifically, among many of the items, there was no range of the common scale where the probability of a participant choosing the slightly disagree or slightly agree category was at a maximum.

Category malfunction may be resolved by collapsing one or more rating scale categories with adjacent categories (Boone & Noltemeyer 2017). In this instance, collapsing the six-point Likert scale into a four-point scale resolved the disordered thresholds. An example of this can be seen in Figure 3 where the category characteristic curves for item 8 *I know where to find useful information about my hearing device(s)* are shown. The top image depicts the original six-point Likert scale (0 = strongly disagree; 1 = disagree; 2 = slightly disagree; 3 = slightly agree; 4 = agree; 5 = strongly agree) and bottom image depicts the collapsing of responses into a four-point Likert scale with original categories disagree and slightly disagree merged and categories slightly agree and agree merged. The curves represent the probability of a category being chosen by a respondent located at each point along the scale. In the six-point Likert scale example, the curves show a disordered relationship between response categories. For example, there is no region of the continuum in which a score of 2 is the most likely. That is, even in the region of proficiencies where the expected (mean) score is 2, people are more likely to obtain one of the other scores. In contrast, the four-point Likert scale depicts an even distribution of response score categories across the scale. Given this finding and the notion that a simpler response scale provides less participant burden, we changed the response scale to a four-point Likert scale, which was subsequently evaluated using a new sample in study 2.

It is important to note that, collapsing response categories in a Rasch analysis can affect other parameters of the data. The specific effects will depend on the nature of the data and the analysis, but in general, collapsing categories can result in changes to item fit and targeting. Collapsing categories can result in changes to item fit statistics, as items may no longer fit the Rasch model as well after collapsing categories. Collapsing categories can also impact targeting, which is the degree to which the distribution of item difficulties matches the distribution of person abilities. If the collapsed categories result in a less precise measure, this can cause the targeting to be less optimal, with the result that some items may be too easy or too difficult for the sample being tested. For these reasons, we first conducted the Rasch analysis with the full 33-item data using the original six-point Likert scale, and then, after identifying the issues potentially caused by the response categories, we collapsed the response categories and re-ran the Rasch analysis on the full 33-item with the collapsed categories (four-point Likert scale). When seeking to refine the items, we looked at data from both the original analysis and the secondary analysis with collapsed response categories.

Differential item functioning • There was no evidence of DIF for gender observed in the analysis of variance (ANOVA) results and the IntraClass Correlations (ICCs) when each gender group was plotted separately.

Dimensionality • Dimensionality is assessed by forming subtests for each of the five domains (knowledge, skills and strategies, self-efficacy, participation, and control; Gotowiec et al. 2022). For example, all of the items relating to self-efficacy were grouped together to create a subtest. A separate analysis was conducted on each subtest facilitating assessment of whether the items in each subtest measure a single underlying dimension of empowerment. The possibility of multidimensionality was investigated through comparing reliability (measured by the PSI) after forming subtests, as well as the percentage of nonerror variance that is common among the content domains (Andrich 2016). Forming subtests compensates for items in a subtest having something in common in addition to the variable being measured by the PROM (multidimensionality), which inflates the reliability (PSI). Therefore, a reduction in PSI after forming subtests is an indication of multidimensionality.

For the dimensionality of the initial 33-item measure before rescoring, the percentage of common nonerror variance among the content domains, which was calculated by forming 5 subtests, was estimated at 84% indicating that there may be some nonunidimensionality among the original items (and original scoring) of the subscales.

Final configurations of the scales • Initially, we had intended to aim for a PROM with approximately 10 items, to minimize respondent burden and maximize the chance that it

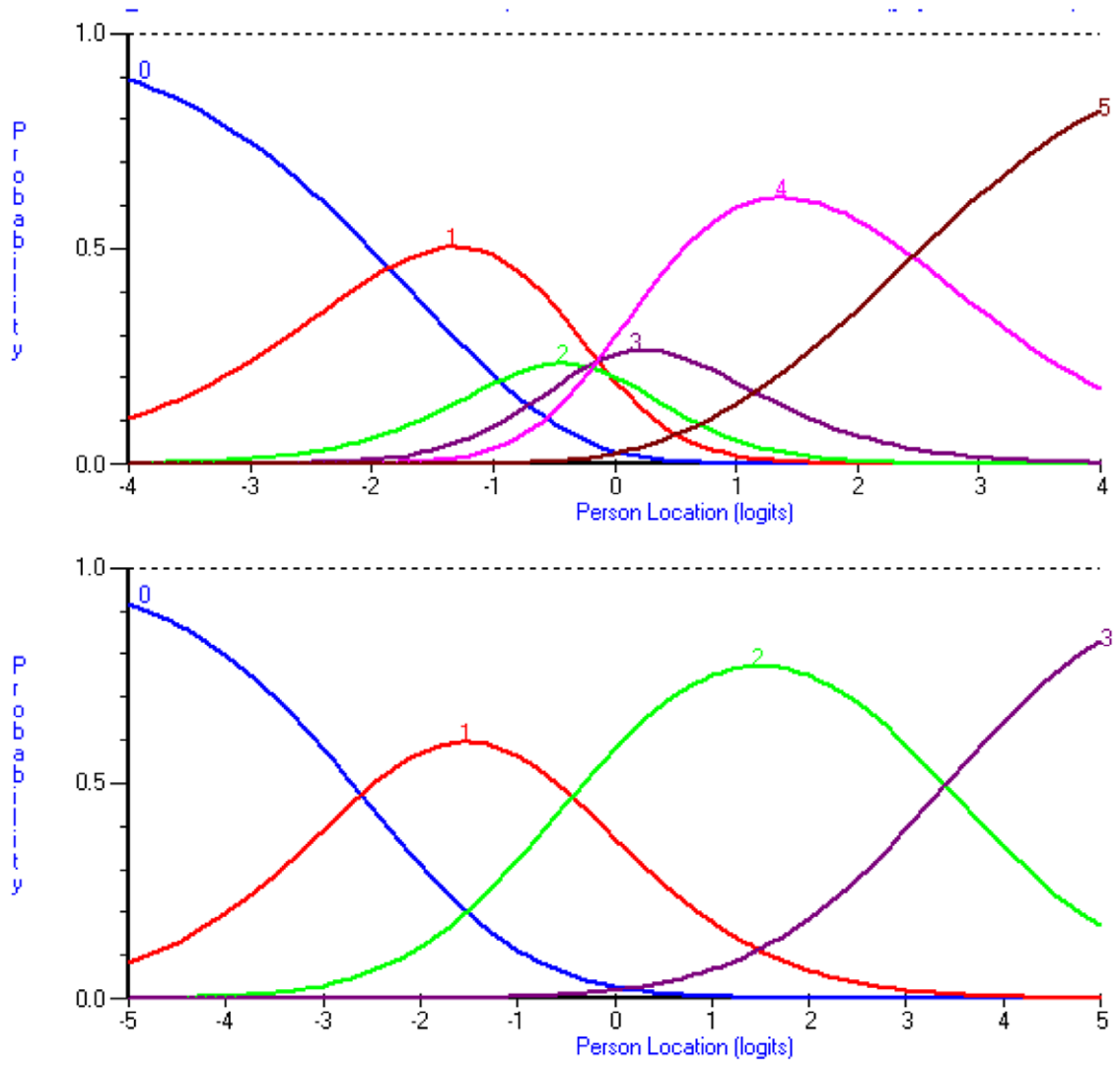

Fig. 3. These two graphs show the category characteristic curves for item 8 "I know where to find useful information about my hearing device(s)." The colored curves represent the probability of participants selecting each response option. The top image depicts the category characteristic curves of the original six-point Likert scale (0 = strongly disagree; 1 = disagree; 2 = slightly disagree; 3 = slightly agree; 4 = agree; 5 = strongly agree) and the bottom image depicts a four-point Likert scale (0 = strongly disagree; 1 = disagree or slightly disagree; 2 = slightly agree or agree; 3 = strongly agree). Note that the categories are disordered in the original six-point Likert scale with the two middle categories not functioning as intended, but the categories are ordered in the rescored four-point Likert scale.

would be adopted. However, the process of iterative item reduction that followed ceased when each scale displayed the requisite psychometric properties (e.g., low response dependency, fit to the Rasch model), and ultimately yielded a 15-item scale. Thinking of how the questionnaire might be used in practice, we opted to develop two versions. First, a long-form of 15 items (EmpAQ-15), retaining all well-performing items, which could be used in a research setting because time constraints were less likely and a more detailed analysis could be done. Second, a short-form of five items (EmpAQ-5), with one item from each dimension, which could be used clinically. The EmpAQ-5 was generated by selecting one item from each domain based on the item fit and response dependency statistics. Each version of the questionnaire underwent a final iteration of analysis.

The psychometric properties of the long-form 15-item and short-form 5-item scales are provided later.

Rasch analysis was used instead of the traditional factor analysis approach as recommended by Terwee et al. (2007).

Fit to the Rasch model • The EmpAQ-5 (mean = −0.92; SD = 2.97) and EmpAQ-15 (mean = −0.51; SD = 2.63) demonstrated good fit to the Rasch model.

Item fit to the Rasch model • The item fit statistics from the analysis of the EmpAQ-15, and in particular the deviation of observed means from their expected values, suggested one item (item 20: *My hearing loss stops me from taking part in social activities*) stood out as being a poor fit. This indicated that the functioning of this item was not consistent with that of the other items. Specifically, it showed much less discrimination than expected with a Chi-square statistic on 5 degrees of freedom greater than 80, relative to an expected value of 5. The particular item was that with negative wording and reversed scoring relative to the other items. Given the explanation, indicated earlier, for this item's misfit, the analysis was repeated with this item removed. The PSI of the EmpAQ-15 remained high with a value of 0.878 for 14 items, indicating strong power in the test of fit. The fit of these items was considered

sufficiently consistent with the model that the PROM was considered adequate to administer to a further sample. Therefore, these 14 items, together with modified wording to eliminate the reverse scoring of item 20, were administered in study 2. Item 20 did not stand out as misfitting in the analysis of the 5-item PROM, however the PSI was not as high as that of the 15-item PROM and therefore there was relatively less power in the test of fit.

Response dependency • The two largest item residual correlations in the analysis of the 15-item PROM were between items 7 *I know where to find useful information about hearing loss* and 8 *I know where to find useful information about my hearing device(s)*, and between items 11 *I use tactics to help me communicate in challenging situations (e.g., move to a quieter location)* and 14 *I search for other ways to help me cope with my hearing loss when I need to (e.g., look online or ask a friend)*. The item pairs demonstrated a residual correlation of the order of 0.36 and 0.31, respectively, with the items in each pair assessing the same content domain likely explaining this result. Despite assessing the same content domain, these items assessed different concepts and were thus retained.

Missing responses • No items had >15% missing responses.

Threshold ordering • No items displayed threshold disordering of the magnitude that suggest poor performance after the rescoring was performed, which indicates that the four response categories within each item were functioning as intended.

Differential item functioning • There was no evidence of DIF for gender in the 15- and 5-item PROMs, which was demonstrated by the ANOVA results and ICCs. Furthermore, the mean estimate of empowerment for each of the gender groups was similar (females: N = 116, M = 1.79, SD = 1.50; males: N = 186, M = 1.85, SD = 1.51).

Targeting • Adequate targeting of the severity of the survey items to the severity of the participants' empowerment was assessed using the plot of the person and item distributions on the same scale. The 15- and 5-item empowerment scales were both adequately targeted for the majority of respondents, with respondents at the extremes not well targeted, as demonstrated by the person-item threshold distribution (Fig. 4). The bars presented in the upper half of the graph (pointing upward) depict the person frequency distribution based on their estimates (which is the degree of empowerment measured by the PROM). The bars presented in the lower half of the graph (pointing downward) represent the distribution of the item thresholds' severity and appropriate targeting is when the spread of items aligns with the spread of persons, as seen in Figure 4. Ideally, we would want to see even distribution of the items across the ability axis, as if items are clustered at one end of the ability continuum, the test may not provide adequate information about participants at the other end of the continuum, and if items are clustered at the center of the continuum, the test may not be sensitive enough to detect differences between participants at the extremes of the continuum. However, in this instance, while items were not evenly distributed across the ability axis, items were spread across the continuum with only slight clustering, suggesting that the test can accurately measure participants' abilities across a wide range of the construct being assessed.

These findings suggest that the PROM is appropriate (not too hard or too easy) for individuals.

The preferred spread of persons depends on the purpose of the PROM. In this instance, we sought to develop a PROM that could be used to detect individual level differences as well as group differences, and these applications of the PROM are supported by the large spread of persons observed. That is, the spread of persons (estimate of empowerment) was well distributed, suggesting that the PROMs could differentiate empowered from disempowered individuals. Of note, the spread of persons was not centered at the mid-point (zero) of the item estimates, but rather, there was a heavier distribution of persons to the right of the mid-point, suggesting that the cohort was considerably empowered relative to the degree of empowerment measured by the items. Although on average the cohort self-rates as high in empowerment, no ceiling effects were noted and a large spread of persons was observed.

Person separation index • The PSI measures of reliability for the 15- and 5-item PROMs were 0.875 and 0.633, respectively (where a value greater than 0.7 shows good person separation). The PSI for the 15-item version was very good, although a little low for the 5-item version. The lower PSI for the 5-item version fell just short of 0.7, however this is consistent with the expectation that fewer items will not be as successful at separating persons based on their degree of empowerment compared with having more items. It is noted that the lower the value of the PSI, the weaker the power of the tests of fit (Andrich & Marais 2019).

Dimensionality • The 15-item empowerment PROM demonstrated no evidence of multidimensionality. In particular, the percentage of common nonerror variance among the content domains, which was calculated by forming 5 subtests, was estimated at 95% for the 15-item PROM. Also, the reduction in reliability (measured by the PSI) was not large after forming the five subtests (from 0.87 to 0.83) which is consistent with unidimensionality. It was not possible to test for dimensionality of the five-item PROM as it only included one item from each dimension.

STUDY 2: RASCH ANALYSIS AND TRADITIONAL PSYCHOMETRIC EVALUATION (PROM VALIDATION)

Methods

The second study aimed to explore validity of the EmpAQ-15 and EmpAQ-5 PROMs through (1) Rasch analysis and (2) traditional psychometric analysis, in adult hearing aid owners. Where Rasch analysis was used to guide item reduction in study 1, here Rasch analysis was used to explore validity of the final structure of the two EmpAQ PROMs. This facilitated an examination of the performance of (1) the reverse scored item and (2) the collapsed response category options (reduced from six to four-point Likert) in a new cohort sample.

It is important to note that, with the reduction of the initial 33-item measure of empowerment by 18 items to one with 15 items, there is a possibility that the selection of items was affected by chance effects and that the items selected would not work as well together in a new sample of data. Therefore, study 2 was aimed at establishing that the set of items worked together as a scale in a new sample of responses.

Traditional psychometric evaluation of the EmpAQ measures included internal consistency, construct validity, and criterion validity. Internal consistency was used to examine the

extent to which PROM items correlate and thus measure the same concept (Terwee et al. 2007). Construct validity refers to how closely the new questionnaire is related to another questionnaire that measures a related or similar construct (De Vet et al. 2011). Not only should the construct correlate with related variables but it should not correlate with dissimilar, unrelated ones. Criterion validity refers to the extent to which scores on the questionnaire relate to a gold standard or other closely related self-report PROM in the absence of a gold standard (Terwee et al. 2007).

Materials • The survey set used in study 2 included both final PROMs of empowerment from study 1 (the EmpAQ-15 and EmpAQ-5), four outcome measures expected to have varying degrees of correlation with the EmpAQ PROMs (to measure construct validity), and a generic measure of empowerment (to measure criterion validity).

Social Participation Restrictions Questionnaire • This is a validated hearing-specific, PROM comprising two subscales, social behaviors (9 items) and social perceptions (10 items). It is scored on an 11-point scale (0 = completely disagree to 10 = completely agree). The SPaRQ as a whole is not a unidimensional scale, however each subscale is unidimensional. We included the first subscale (social behaviors; nine items) in the survey set as items explore behaviors relating to social participation and the EmpAQ PROM also focuses on behaviors. SPaRQ social behaviors scores were calculated by summing scores on the nine items comprising this subscale, with higher scores indicating higher degrees of difficulty in social situations (Heffernan et al. 2018b).

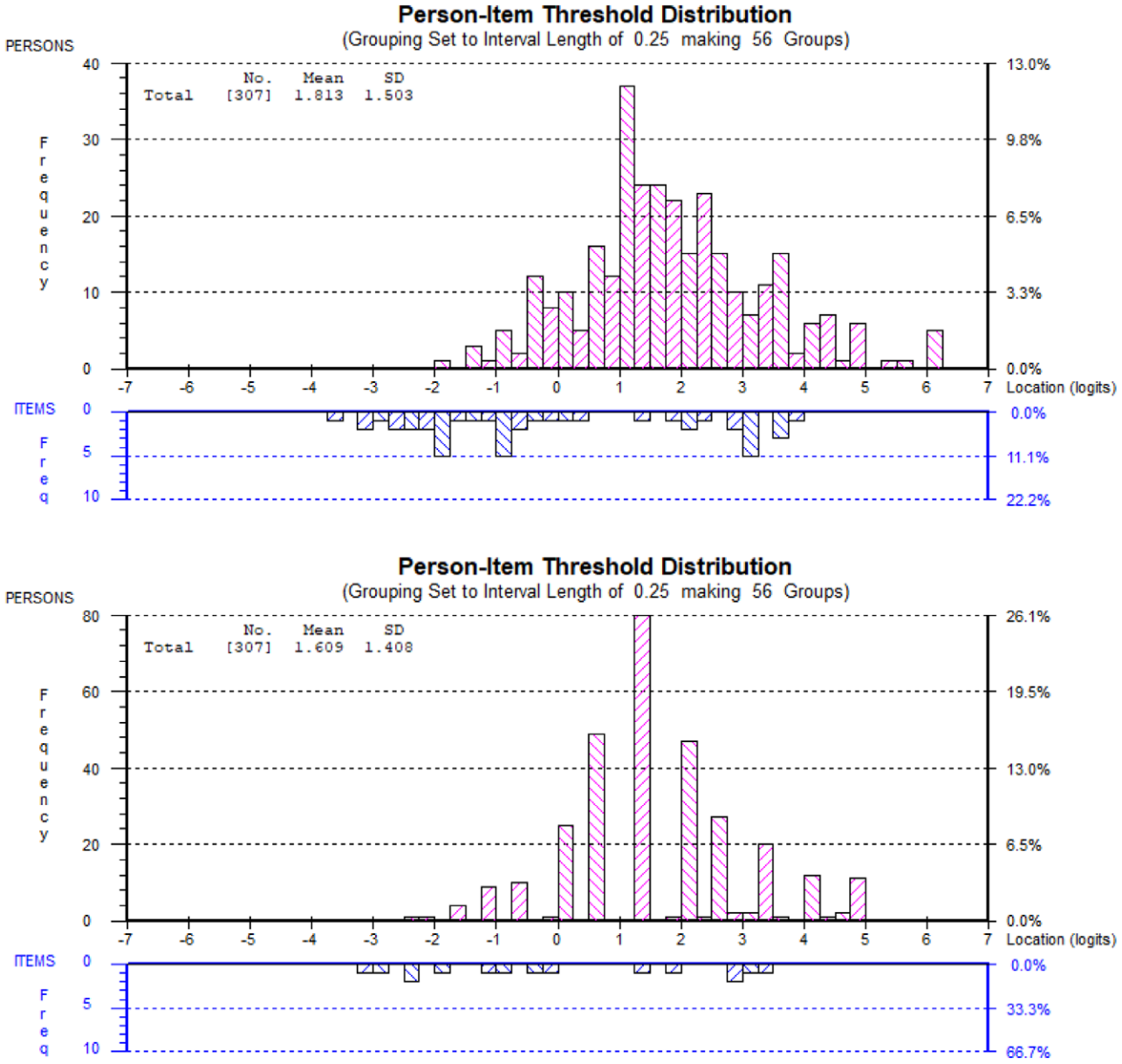

Fig. 4. Person-item threshold distribution for the 15-item empowerment PROM (upper image) and the 5-item empowerment PROM (lower image). The red bars (top half of the image) represent the spread of participants' location estimates across the common scale demonstrating the distribution of the persons' severity. The blue bars (lower half of each image) represent the distribution of the item thresholds' severity. PROM indicates patient-reported outcome measures.

Self-Assessment of Communication • This is a validated 10-item measure of communication with two parts focusing on (1) disability or activity limitation (five-item), and (2) handicap or participation restriction (four-item; Schow & Nerbonne 1982). It is scored on a five-point Likert scale (*Almost never or never* to *Practically always or always*). All 10 items of the Self-Assessment of Communication (SAC) were included in the survey set. SAC scores were calculated by summing scores across these 10 items, with higher scores indicating higher degrees of communication difficulties.

Patient activation measure (PAM) • This is a validated 13-item scale to determine patient engagement in healthcare (Hibbard et al. 2005). It is scored on a four-point Likert scale (*Disagree Strongly* to *Agree Strongly*, and includes a *N/A* option). PAM scores were calculated using a proprietary scoring algorithm provided by Insignia Health. This algorithm produces two scores. The first places the respondent in one of four broad categories. The second falls on a 0 to 100 scale. We analyzed scores on the 0 to 100 scale, with higher scores indicating higher levels of activation.

World Health Organization Disability Assessment Schedule 2.0 • This is a validated 12-item scale generic questionnaire assessing activity and participation domains: understanding and communication, mobility, self-care, getting along with others, life activities, and societal participation (Luciano et al. 2010; Üstün et al. 2010). Respondents rate how much difficulty they have experienced for each item within the past 30 days using a five-point scale (*None* to *Extreme or cannot do*). The 12-item World Health Organization Disability Assessment Schedule 2.0 (WHO-DAS 2.0) can be scored using one of two methods: complex, which involves weighting items based on IRT, and simple, which involves summing responses to each item (Rehm et al. 1999). We elected to use the simple scoring method. Psychometric evaluation indicates that these two scoring methods yield highly similar scores within samples (rs > 0.98; Andrews et al. 2009).

Health Care Empowerment Questionnaire (HCEQ) • This is a 10-item measure of patient empowerment (Gagnon et al. 2006). The HCEQ was developed with a sample of aging adults in Canada to measure three aspects of patient empowerment (1) degree of control (consideration of who is involved in making decisions), (2) involvement in interactions (the ability and opportunity to communicate needs and initiate requests with a healthcare provider), and (3) involvement in decisions (actively obtaining the information necessary to make healthcare decisions). The original HCEQ requires respondents to answer each of the 10 questions in two ways: with regard to their experience (*Did you feel that…*) and their perceptions of the importance of the item (*How important is it that…*). In the present study, we included only the scale of experience as it evaluates behaviors relating to empowerment, as do the EmpAQ PROMs. For example, HCEQ *During the last 6 months, did you feel…* (item 5) *that your choices were respected?* Participants rated their agreement with each item on a five-point Likert scale (5 = strongly agree to 1 = strongly disagree). Subscale scores were calculated by summing scores on items comprising each subscale (degree of control, involvement in interactions, and involvement in decisions). Total scores were calculated by summing scores on all 10 items. Higher scores indicate higher degrees of empowerment.

The survey set included the earlier surveys presented in the following order: (1) EmpAQ-5; (2) SPaRQ; (3) SAC; (4) PAM; (5) WHO-DAS 2.0; (6) HCEQ; (7) EmpAQ-15; and (8) demographic questions.

Participants • Hearing aid owners were recruited from a hearing service provider in Western Australia. Inclusion criteria were (1) clients aged 18 years or older, and (2) who had been using a hearing aid for a minimum of 6 months. Exclusion criteria were (1) self-reported nonfluency in written and spoken English, or (2) self-report cognitive decline or dementia that would require assistance in completing questionnaire items. No exclusions were placed on daily hours of hearing aid use. Based on international guidelines, we aimed for a minimum sample size of 150 adult hearing aid owners (Mokkink et al. 2019).

Procedure • The approach to recruiting rounds of 50 potential participants in study 1 was repeated for the recruitment of participants in study 2 until the required 150 sample size had been reached. A total of 650 potential participants were invited to participate, with 187 starting the survey (response rate 28.77%). Of the 187 respondents, 7 were excluded from analyses due to not having fully completed at least the first survey in the survey set (the EmpAQ-5). Two further participants completed the majority of the survey but did not complete the two final questionnaires (i.e., the EmpAQ-15 and demographics items). These 2 participants were retained in analyses relating to the EmpAQ-5.

E-mail invitations included a digital PIF and a link to the survey (Qualtrics). The landing page of the survey provided an overview of the project and another opportunity to view the PIF. All participants provided consent to participate by ticking a consent box within the survey before gaining access to the subsequent survey items. There were no fees or incentives for participation. Study 2 items took 35 to 45 min to complete.

Data Analysis • Survey responses were exported and analyzed using RUMM2030 and IBM SPSS Statistics 28.0 software.

Rasch analysis of the study 2 dataset investigated item fit to the Rasch model, response dependency, missing responses, threshold ordering, DIF, targeting, person separation, and dimensionality as per study 1.

Internal consistency was assessed using Cronbach alpha (Terwee et al. 2007). Given that each of the empowerment measures were designed to allow "not applicable" responses, we calculated Cronbach alpha using the method outlined in Weaver and Maxwell (2014), which allows for calculation of reliability in the presence of missing values (i.e., "not applicable" responses). This method uses the expectation maximization algorithm within SPSS to calculate a matrix of interitem correlations. This matrix was then entered into the reliability analysis. Cronbach alpha was expected to fall within the range of 0.7 to 0.95, as required for a PROM or its subscales (De Vet et al. 2011).

Construct validity was explored using Pearson correlation coefficient to assess predictions about the construct validity of the two empowerment measures. It was hypothesized that the EmpAQ-15 and EmpAQ-5 would have a moderate, negative correlation with the two hearing-specific measures (SPaRQ: social participation restriction; and SAC: communication difficulties), and a low correlation with the two generic health measures

*These groupings of the response categories were chosen because the ANOVA requires groups to be of near equal size.

TABLE 3. Correlations between each of the EmpAQ-5 and EmpAQ-15 PROMs and other questionnaires for the purpose of evaluating construct validity

	EmpAQ-5			EmpAQ-15		
	r	*p*	n	*r*	*p*	n
SPaRQ—social behaviors	−0.36	<0.001	179	−0.28	<0.001	177
Patient SAC	−0.49	<0.001	179	−0.47	<0.001	177
PAM	0.55	<0.001	179	0.62	<0.001	177
WHO-DAS 2.0—12-item	−0.38	<0.001	178	−0.46	<0.001	175
HCEQ—total score	0.37	<0.001	177	0.29	<0.001	175
HCEQ—involvement in decisions	0.18	0.014	179	0.05	0.549	177
HCEQ—involvement in interactions	0.44	<0.001	177	0.50	<0.001	174
HCEQ—degree of control	0.17	0.020	179	0.07	0.360	177

EmpAQ-5, Empowerment Audiology Questionnaire-5; EmpAQ-15, Empowerment Audiology Questionnaire-15; HCEQ, Health Care Empowerment Questionnaire; PAM, patient activation measure; PROMs, patient-reported outcome measures; SAC, Self-Assessment of Communication; SPaRQ, Social Participation Restrictions Questionnaire; WHO-DAS 2.0, World Health Organization Disability Assessment Schedule 2.0.

(positive correlation with the PAM as it measures patient activation; and negative correlation with the WHO-DAS 2.0 as it measures global disability). Criterion validity was assessed using Pearson correlation coefficient to compare EmpAQ-15 and EmpAQ-5 scores with HCEQ scores. Interpretability of the size of correlations followed Cohen's guidelines of: small $r = 0.10$; medium $r = 0.30$; and large $r = 0.50$ (Cohen 1988).

Results

Participant Characteristics • Data from 180 participants were examined in study 2; all of who completed the EmpAQ-5 and 178 completed both the EmpAQ-5 and the EmpAQ-15.

Rasch Analysis • Rasch analysis was applied to the study 2 dataset, with special interest in the performance of (1) the reverse scored item and (2) the response category threshold ordering in a new cohort sample.

Fit to the Rasch model • Both the EmpAQ-15 (mean = −0.14; SD = 1.42) and EmpAQ-5 (mean = −0.39; SD = 1.87) performed well with good item fit to the Rasch model; an improvement from the results observed in study 1.

Item fit to the Rasch model • Both the EmpAQ-15 and EmpAQ-5 performed well with good item fit to the Rasch model. The wording of item 20 (item number in study 1) was changed from negatively to positively worded (item 9 in the EmpAQ-15 and item 3 in the EmpAQ-5 in study 2) and performed well in this sample, demonstrating good fit to the model (fit residual 1.591 in EmpAQ-15 analysis and −0.137 in EmpAQ-5 analysis) and the four response categories functioned as intended (see Supplemental Digital Content 2, http://links.lww.com/EANDH/B250).

Response dependency • Residual correlations were all within acceptable levels.

Missing responses • No items had >15% missing responses.

Threshold ordering • The four-point Likert response option worked better than the six-point Likert used in study 1 with no items displaying reversed threshold ordering.

Differential item functioning • There was no evidence of DIF for gender in the 15- and 5-item PROMs, which was demonstrated by the ANOVA results and ICCs. There was also no DIF for age when the response categories were combined to be 69/below, 70 to 74, 75 to 79, 80/above*; no DIF for years with hearing loss when the response categories were combined to be 5/below, 6 to 10, 11 to 19, 20/above*; no DIF for hours using device each day when the response categories were combined to be 1 to 4, 5 to 8, 9 to 11, 12, 13 to 15, and 16 to 24.*

Targeting • The 15- and 5-item empowerment scales were both adequately targeted for the majority of respondents, with respondents at the extremes not well targeted, as demonstrated by the person-item threshold distribution (Supplemental Digital Content 3, http://links.lww.com/EANDH/B251).

Person separation index • The PSI measures of reliability for the 15- and 5-item PROMs were 0.874 and 0.655 respectively. The value for the 15-item PROM demonstrates good reliability, in terms of what is typically considered good reliability for a PROM, as it is well above the 0.7 value (Tennant & Conaghan 2007). Whereas the value for the five-item PROM falls a little short of 0.7, as would be expected due to the PROM having fewer items.

Dimensionality • Dimensionality analysis supported the existence of unidimensionality of the EmpAQ-15.

Traditional Psychometric Analysis •

Outliers • The data were screened for multivariate outliers via examination of Mahalanobis distances (Tabachnick et al. 2007). Data points were considered multivariate outliers in any given analysis if the *p* value for their Mahalanobis distance was <0.001. Where multivariate outliers were identified, these were removed before calculation of relevant correlation coefficients. This process resulted in up to four outliers being removed from any one analysis. Resulting *n*s are reported in Table 3.

Internal consistency • Internal consistency was within the recommended range for PROMs (0.70 to 0.9; De Vet et al. 2011; Raykov & Marcoulides 2011) for both the EmpAQ-5 ($\alpha = 0.72$; n = 180) and EmpAQ-15 ($\alpha = 0.90$; n = 178).

Agreement between EmpAQ-5 and EmpAQ-15 scores • There was a positive correlation between EmpAQ-5 and EmpAQ-15 scores $r = 0.62$, $p < 0.001$ ($n = 176$).

Construct validity • Construct validity was evaluated by calculating Pearson correlations between EmpAQ-5 and EmpAQ-15 and selected questionnaires measuring related constructs (Table 3). Higher scores on the EmpAQ-5 and EmpAQ-15 (higher degrees of empowerment) were significantly associated with lower levels of social participation restriction (SPaRQ social behaviors scale), and lower degrees of communication difficulties related to hearing loss (patient SAC), as hypothesized. However, the magnitude of these associations was lower than predicted. Higher scores on the EmpAQ-5 and EmpAQ-15

(higher degrees of empowerment) were significantly associated with higher levels of patient activation (PAM) and lower levels of global disability (WHO-DAS 2.0), as hypothesized. However, the strength of the associations between the EmpAQ and PAM was higher than predicted.

Criterion validity • A moderate, positive correlation between the HCEQ and EmpAQ measures was observed, supporting criterion validity of the measures (Table 3).

DISCUSSION

The overall aim of this study was to refine and validate a hearing-specific measure of empowerment. The process resulted in two measures, a 15-item (EmpAQ-15) and 5-item (EmpAQ-5) measure. The EmpAQ-15 and EmpAQ-5 are the first measures to be developed specifically for the measurement of empowerment in adults with hearing loss, and are among the few hearing-related self-report measures to be developed and validated using modern psychometric techniques (Heffernan et al. 2018b; Hughes et al. 2021). In following best practice guidelines (Gagnier et al. 2021), this study used both modern and traditional methods for item refinement and psychometric analysis. The results demonstrate that the EmpAQ-15 and EmpAQ-5 have strong psychometric properties. Specifically, they satisfy the Rasch model requirements for interval-level measurement, in that the data meet the key assumptions for unidimensionality (i.e., measuring a single underlying construct or trait), local independence (i.e., responses to each item are independent of the responses to other items), item homogeneity (i.e., items are similar in terms of their difficulty levels and the extent to which they are able to measure the underlying construct being assessed), and absence of DIF (i.e., the probability of answering an item as it corresponds with their actual experience of the condition should be the same across different subgroups of the population being assessed). In satisfying the Rasch model, both the EmpAQ-15 and EmpAQ-5 have the potential to be used as both research and clinical tools for measuring individual changes in empowerment along the hearing journey as well as for making group comparisons (Browne & Cano 2019).

Rasch Analysis

Using Rasch analysis to develop the EmpAQ measures offered several advantages. First, Rasch offers a systematic and data-driven process for item refinement. We used evidence-driven, participatory processes to generate the initial item pool of 47 potential survey items (Gagnier et al. 2021; Gotowiec et al. 2022, 2023), which then underwent content analysis to refine and reduce the pool to 33-items (Gagnier et al. 2021; Gotowiec et al. 2022, 2023). Rasch analysis facilitated selection of the items that were able to separate persons with different degrees of empowerment. Specifically, they satisfy the Rasch model requirements for interval-level. The PSI is a reliability index but based on Rasch model parameter estimates and their standard errors, to estimate the variance of the true scores and standard errors. The PSI was useful for evaluating the quality of the EmpAQ terms and their ability to accurately measure individual differences in empowerment (i.e., the degree to which the EmpAQ is able to distinguish between individuals who have different degrees of empowerment).

Second, using Rasch analysis allowed us to explore the potential for multidimensionality. When we first embarked upon this research program, we based our initial understanding of empowerment primarily on the well-established framework, Zimmerman's theory of empowerment (Zimmerman 1995). This framework is built on the assertion that empowerment is a multidimensional construct with three different necessary components: intrapersonal, interactional, and behavioral elements (Zimmerman 1995). In different settings, these three components may comprise different specific dimensions. Following in-depth interviews with adults with hearing loss, we came to understand empowerment as it relates to hearing loss to comprise five domains: knowledge, skills and strategies, participation, control, and self-efficacy (Gotowiec et al. 2022). Generation of the item pool for the EmpAQ measures drew on these domains and thus it was important to explore whether the final measure developed was multidimensional (tapping into individual domains) or unidimensional. The Rasch results demonstrated that the EmpAQ-15 is a unidimensional measure of empowerment, and is thus scored as a single score.

Third, Rasch analysis is an iterative process, and while it can be a lengthy procedure, the multiple stages allow iterative refinement to optimize PROM development. For example, traditional item refinement, such as factor analysis, can identify items to be removed based on the extent to which each variable in the dataset is associated with a common theme or factor, whereas Rasch analysis provides data and information on how each individual item performs. This enables researchers to understand possible reasons why the item may not be performing as expected, and thus whether removal or revision of the item is required. For example, we had previously used cognitive interviews with adults with hearing loss and expert panel review to inform clarity and relevance of items (Gotowiec et al. 2023). However, there was disagreement between participants about the wording of the two items describing participation in social activities. Although several adults with hearing loss preferred the wording for one of the items *My hearing loss doesn't stop me from taking part in social activities*, other adults with hearing loss and several expert panel members suggested that this double negative might be confusing ("doesn't stop"), and suggested flipping the statement to simplify the wording: *My hearing loss stops me from taking part in social activities*. However, flipping the item in this way meant that for these two items, participants needed to reverse score the items. That is, for all other items the right-hand side of the response scale indicated a positive response (high empowerment) and the left-hand side represented low empowerment, yet the opposite was true for these two items, requiring participants to mentally reverse the response options. Given that this was the recommendation of both adults with hearing loss and expert panel members during the item development phase, this reversed wording was used in study 1. Traditional item reduction processes such as factor analysis would have likely identified these two items as outliers and likely led to their exclusion from the PROM. However, through exploring the category functioning and frequency of responses of these items in relation to the other items, Rasch analysis identified that participants responded to these items differently. We removed one of the items (due to redundancy), and we reversed the wording of the other item, described earlier, and included it in study 2. Once reversed, the item performed more strongly. In this way, Rasch analysis was able to help us

iteratively sculpt the measure rather than simply identify which items to remove, as do traditional methods of item reduction. It is important to note that, the final wording of this item was considered valid to the end users (cognitive interview participants) and a good fit to the Rasch model based on study 2 results. Overall, the iterative item reduction process ultimately led to the removal of 18 items that displayed poorer psychometric properties relative to those retained. While researcher judgment was used to guide the process, in most instances, decisions to remove items were data-driven and helped to reduce researcher bias when selecting the final PROM items.

Fourth, Rasch analysis enabled us to detect and address category malfunctioning. The first version of the empowerment measure generated from the qualitative study to conceptualize empowerment (Gotowiec et al. 2023), included 33-items measured with a six-point Likert scale format. Category probability curves generated based on the six-point Likert scale identified that many items showed a disordered relationship between response categories, a sign that the Likert scale is not functioning as intended or is not suitable for the particular population being measured. Reasons for the scale not functioning as intended can include poorly defined response categories, ambiguous wording, or cultural differences in the interpretation of the response categories (Boone & Noltemeyer 2017). Collapsing the response scale from a six-point Likert scale to a four-point Likert scale results in less disordering because the reduced number of response categories makes it easier for individuals to distinguish between them and choose the one that best reflects their level of the latent trait. That is, if more categories are working as intended then they provide more information; but if there are too many categories for respondents to deal with readily, then they can provide more "noise" than information, as was the case for the EmpAQ. Although the version of the measure that was analyzed via content validity used a six-point Likert scale, participants did not comment on whether they required the six categories or not. Collapsing the six-point Likert scale into a four-point scale resolved the disordered thresholds as demonstrated in study 1 and confirmed in study 2.

Fifth, the Rasch model establishes a "fixed ruler" that represents a continuous spectrum of the phenomenon of interest (Riff et al. 2017). Consistent with other Rasch-developed PROMs (Heffernan et al. 2018b; Hughes et al. 2021), the raw score of the EmpAQ is converted into a Rasch-transformed 0 to 100 scale, intended to enhance its usability for both clinicians and patients. However, we acknowledge the ongoing discussion within the literature around Rasch and interval versus pseudo-interval -level measurement (Salzberger 2010). While some have suggested that the Rasch model transforms or converts ordinal scales into interval scales (Tennant & Conaghan 2007), others suggest that the Rasch model is merely capable of constructing linear measures from counts of qualitatively-ordered observations (Linacre & Wright 1993). In this context, it is important to note that while linear transformations have been applied to convert EmpAQ scores to a 0 to 100 scale, this transformation affects the origin and unit of measure rather than creating a linear scale with equal step sizes. The logit scale maintains the relative ordering of individuals' empowerment levels while accounting for the nonuniform distribution of responses inherent to the Rasch model. While the transformed EmpAQ scores on the 0 to 100 scale provide a convenient representation of perceived empowerment levels, interpretation of differences between EmpAQ scores must still consider the nonlinearity of the underlying measurement and statistical analysis should use approaches appropriate for logit-based measurements (e.g., nonparametric tests).

With regard to clinical interpretation, high scores on the EmpAQ survey indicate a strong sense of empowerment, whereas low scores suggest a lower degree of empowerment. Examining individual items with both high and low scores can provide valuable insights into areas where the client is empowered as well as areas where they may require specific assistance. Trends in the magnitude across time of EmpAQ scores within an individual can provide valuable insights into changes in their perceived empowerment levels. In addition, comparing EmpAQ scores between clients may offer a broader understanding of relative empowerment levels within the sample. As with all self-report measures utilizing ordinal response options, it is important to be aware of the nonlinearity of logit-based scales.

Last, application of the Rasch model supports the development of short forms and multiple versions of a questionnaire due to psychometric analysis occurring at the item level rather than at the scale level. For example, the social isolation measure (Heffernan et al. 2019) was a 5-item short form of the 19-item SPARQ (Heffernan et al. 2018b). Following development of the EmpAQ-15, we used information on item fit to the model, item difficulty, and item discrimination to select a small subset of items; the EmpAQ-5. Development of a short form will likely optimize PROM uptake, while ensuring adequate measurement of persons located across the common scale. The EmpAQ-5 might have use as a screener and preferred over the EmpAQ-15 by clinicians, given time constraints of the clinical setting.

Traditional Psychometric Analysis

In study 2, traditional psychometric analyses were used to validate the final structure of the two measures. Both demonstrated acceptable internal consistency, construct validity and criterion validity.

Although evaluation of construct validity was predominantly as expected (with correlations detected and, in the direction, hypothesized), the strength of the correlations was not as we had predicted. We hypothesized that the two hearing-specific measures (SPaRQ and SAC) would be more highly correlated with the EmpAQ measures than the two general health measures, given their shared specificity to the lived experience of hearing loss. However, this was not the case. Instead, the EmpAQ measures demonstrated strongest correlation with the PAM, a quantifiable scale determining patient activation in healthcare. The items comprising the PAM explore a person's skills, confidence, and knowledge to manage their own health (Hibbard et al. 2005). Skills and strategies, knowledge and self-efficacy (confidence) were three of the five dimensions identified as conceptualizing empowerment on the hearing health journey (Gotoweic et al. 2022) and informed the EmpAQ item generation (Gotoweic et al. 2022). The strong correlation observed between the PAM and EmpAQ measures likely represents the overlap in concepts underpinning patient activation and empowerment.

Criterion validity was explored through correlation analysis of the EmpAQ measures and the general HCEQ. The positive correlation between the measures further supports the notion that the EmpAQ measures are measuring related underlying constructs. Correlation analysis between the three dimensions of the HCEQ and the EmpAQ measures revealed a strong

relationship with the dimension *Involvement in Interactions* and a weak relationship with the dimensions *Involvement in Decisions* and *Degree of Control*. The low correlations possibly reflect the broad concepts of the HCEQ being less sensitive to hearing-specific empowerment than the context-specific items on the EmpAQ. General measures designed to assess broad aspects of health that are not specific to a particular disease (e.g., HCEQ) facilitate comparison of patients across different health conditions (Al Sayah et al. 2021). Whereas condition-specific measures assess aspects of health specific to a disease (e.g., EmpAQ) and can be used to detect changes in aspects of a specific health condition.

A Brief Measure of Empowerment

We initially set out to develop a single measure of empowerment. However, cognitive interview data from adults with hearing loss and survey data from an expert panel (audiology clinicians and researchers) highlighted the need for the measure to be brief so as to reduce participant burden and increase the likelihood that clinicians would make time to administer the measure within routine clinical practice. The advantages of a brief questionnaire compared with a more detailed, longer version include: time-saving for both the patients/respondents and the clinicians/researchers; high response rates due to lower burdensomeness; and ease of administration as due to length it may be adaptable for administration in person, online, over the phone or via mobile devices (e.g., for ecological momentary assessment). However, the disadvantages of a brief measure include: it may be less comprehensive and thus have a limited ability to differentiate between individuals with similar scores; have limited validity and reliability; and may be more susceptible to response bias. Using Rasch analysis to identify poorly performing items and iteratively reduce the number of items, the first iteration whereby all items performed well, was a 15-item measure. As a research team, we thus chose to retain the 15-item version so as to provide a comprehensive and theoretically driven measure of empowerment, and also develop a brief version for clinical use.

STRENGTHS AND LIMITATIONS

A key strength of this study was the application of both modern and traditional methods to inform PROM refinement and validation. In addition, given that the EmpAQ-5 does not contain any items that specifically describe hearing aid ownership, a strength is that it has potential to be used in evaluation of a range of hearing-related interventions and in pre-post comparisons. Responsiveness, minimally important change, and clinical interpretability, with multi-national cohorts, as well as clinical implementation, will be the subject of future investigation.

This study is not without limitations. Participants self-selected for the study, potentially contributing to response biases. Response rates of 29.67% (study 1) and 28.77% (study 2) were observed, which could potentially have been improved through use of follow-up e-mail reminders or participant incentives. Participants were all recruited from a single organization in Western Australia, and while they were diverse in age, gender, years of hearing loss, and severity of hearing-related impact on daily life, they mostly reported high levels of hearing aid use, which may have biased the results. Overall, the cohorts recruited in both rounds appeared to demonstrate high degrees of empowerment, which may be related to the quality of service and support they received from the clinic from which they were recruited. As such, further research exploring hearing-related empowerment in a more diverse sample is required to understand the normal range of degree of empowerment for people with hearing loss. Furthermore, participants were all hearing aid owners, and further research is needed to explore how the measures might perform in other populations of people with hearing difficulties, such as those in early help-seeking stages of their hearing health journey.

In study 1, we inadvertently only looked at DIF for gender despite recommendations of assessing DIF for both age and gender (Tennant & Conaghan 2007). Subsequently, we looked at DIF for gender, age, and years of hearing loss in study 2. None the less, DIF was found to be acceptable in study 2. Regarding construct validity, guidelines require that "The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured, and that at least 75% of the results are in accordance with pre-defined specific hypotheses" (Terwee 2007). However, we did not achieve at least 75% of the results in accordance with our predefined hypotheses, and thus we will reevaluate construct validity in our future research that also aims to evaluate test-retest reliability and responsiveness.

In study 2, participants completed both the 15-item measure and 5-item measure (i.e., they completed the same items twice) within the same survey set, usually within 35 to 45 min. This may have impacted the results, and further research is needed to explore how the measures might perform independent of each other.

Last, a recent article by Ekstrand et al. (2022) explored the application of the least measurable difference, the standard error of measurement, and the least significant difference as means to transform the logit scale into ranges that preserve their linear properties. Our future research will explore the possibility of using these approaches to logit transformation to develop a scale that will enable researchers to analyze data using parametric tests and clinicians to make interval-level comparison. As part of this exploration, we must consider the needs of all potential user groups and consider whether requiring clinicians to transform the data using a logit scale, requiring them to perform calculations or conversions, might be seen as an extra task that adds complexity to their workflow and inevitably work as a barrier to clinical uptake of the EmpAQ, especially as clinicians are used to working with traditional measurement scales with raw scores. In this way, preserving the original scale may align better with the needs and preferences of clinical stakeholders. These considerations will be explored in our future work exploring implementation needs for clinicians wanting to use the EmpAQ.

CONCLUSION

The refinement and validation of the 15- and 5-item EmpAQ, a new self-report measure of empowerment is described. The questionnaire has strong psychometric properties fit for use as both a research or clinical tool, and can also be used as a guide for dialogue between clinician and patient. This study also highlights the benefits of using modern psychometric analysis techniques in conjunction with traditional approaches to develop high-quality self-report measures and provides detailed

examples of how to apply these methods. In future work that builds on this study, it will be important to further investigate responsiveness and clinical interpretability, as well as investigate the possibilities and value of clinical implementation of the EmpAQ-5 and EmpAQ-15.

The EmpAQ-15 and EmpAQ-5 can be freely downloaded from https://osf.io/caj84/.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of David Andrich and Sonia Sappl for guiding data analysis, interpretation of results and reporting. The authors also acknowledge the contributions of Ellen Bothe to the data analysis, Ear Science Institute Australia clinics for distributing recruitment invitations, and all those who participated in the study.

This project was funded by WSAudiology. R.J.B. and M.F. received financial support from WSAudiology.

S.G. and J.L. are employees of WSAudiology.

All authors contributed equally to the design of this research project. R.J.B. and M.F. conducted data collection. All authors contributed to data analysis, interpretation, and wrote the article. All authors discussed the results and implications and commented on the article at all stages.

Address for correspondence: Rebecca J. Bennett, Audiological Sciences, National Acoustic Laboratories, Macquarie University, Sydney 2109, Australia. E-mail: bec.bennett@nal.gov.au

Received October 24, 2022; accepted October 22, 2023; Published online ahead of print December 12, 2023.

REFERENCES

Al Sayah, F., Jin, X., Johnson, J. A. (2021). Selection of patient-reported outcome measures (PROMs) for use in health systems. *J Patient Rep Outcomes*, *5*, 1–5.

Allen, D., Hickson, L., Ferguson, M. (2022). Defining a patient-centred core outcome domain set for the assessment of hearing rehabilitation with clients and professionals. *Front Neurosci*, *16*, 787607.

Andrews, G., Kemp, A., Sunderland, M., Von Korff, M., Ustun, T. B. (2009). Normative data for the 12 item WHO Disability Assessment Schedule 20. *PLoS One*, *4*, e8343.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res*, *11*, 571–585.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *J Appl Meas*, *4*, 205–221.

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. *D. Andrich y I. Marais (Coords.), Measuring in the Educational, Social and Health Sciences*: 41–53.

Andrich, D., Sheridan, B., Luo, G. (2022). *RUMM2030: Rasch unidimensional models for measurement*. Perth, Western Australia: RUMM Laboratory, *3*, 1–10.

Andrich, D. (2016). Rasch rating-scale model. In *Handbook of item response theory* (pp. 75-94). Chapman and Hall/CRC.

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.

Barr, P. J., Scholl, I., Bravo, P., Faber, M. J., Elwyn, G., McAllister, M. (2015). Assessment of patient empowerment—A systematic review of measures. *PLoS One*, *10*, e0126553.

Bennett, R. J., Barr, C., Montano, J., Eikelboom, R. H., Saunders, G. H., Pronk, M., Heffernan, E. (2021). Identifying the approaches used by audiologists to address the psychosocial needs of their adult clients. *Int J Audiol*, *60*, 104–114.

Bennett, R. J., Laplante-Lévesque, A., Eikelboom, R. H. (2019). How do hearing aid owners respond to hearing aid problems? *Ear Hear*, *40*, 77–87.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Educ*, *4*, 1416898.

Bradley, K. D., Peabody, M. R., Akers, K. S., Knutson, N. (2015). Rating scales in survey research: Using the Rasch model to illustrate the middle category measurement flaw. *Surv Pract*, *8*, 1–11.

Brod, M., Tesler, L. E., Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Qual Life Res*, *18*, 1263–1278.

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.

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.

Christensen, K. B., Makransky, G., Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas*, *41*, 178–194.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cyril, S., Smith, B. J., Renzaho, A. M. (2016). Systematic review of empowerment measures in health promotion. *Health Promot Int*, *31*, 809–826.

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

Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Med Res Methodol*, *22*, 1–10.

Gagnier, J. J., Lai, J., Mokkink, L. B., Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res*, *30*, 2197–2218.

Gagnon, M., Hébert, R., Dubé, M., Dubois, M.-F. (2006). Development and validation of an instrument measuring individual empowerment in relation to personal health care: The Health Care Empowerment Questionnaire (HCEQ). *Am J Health Promot*, *20*, 429–435.

Gomez, R., Habib, A., Maidment, D. W., Ferguson, M. A. (2021). Smartphone-connected hearing aids enable and empower self-management of hearing loss: A qualitative interview study underpinned by the behavior change wheel. *Ear Hear*, *43*, 921–932.

Gotowiec, S., Bennett, R., Larsson, J., Ferguson, M. (2023). Development of a self-report measure of empowerment along the hearing health journey: A content evaluation study. *Int J Audiol*, 1–11.

Gotowiec, S., Larsson, J., Incerti, P., Young, T., Smeds, K., Wolters, F., Ferguson, M. (2022). Understanding patient empowerment along the hearing health journey. *Int J Audiol*, *61*, 148–158.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes*, *15*, 1–8.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educ Meas: Issues Pract*, *12*, 38–47.

Heffernan, E., Coulson, N. S., Ferguson, M. A. (2018a). Development of the Social Participation Restrictions Questionnaire (SPaRQ) through consultation with adults with hearing loss, researchers, and clinicians: A content evaluation study. *Int J Audiol*, *57*, 791–799.

Heffernan, E., Habib, A., Ferguson, M. (2019). Evaluation of the psychometric properties of the social isolation measure (SIM) in adults with hearing loss. *Int J Audiol*, *58*, 45–52.

Heffernan, E., Maidment, D. W., Barry, J. G., Ferguson, M. A. (2018b). Refinement and validation of the Social Participation Restrictions Questionnaire: An application of Rasch analysis and traditional psychometric analysis techniques. *Ear Hear*, *40*, 328–339.

Hibbard, J. H., Mahoney, E. R., Stockard, J., Tusler, M. (2005). Development and testing of a short form of the patient activation measure. *Health Services Res*, *40*, 1918–1930.

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

Holmström, I., & Röing, M. (2010). The relation between patient-centeredness and patient empowerment: A discussion on concepts. *Patient Educ Couns*, *79*, 167–172.

Hughes, S. E., Watkins, A., Rapport, F., Boisvert, I., McMahon, C. M., Hutchings, H. A. (2021). Rasch analysis of the listening effort questionnaire—Cochlear implant. *Ear Hear*, *42*, 1699–1711.

Khuntia, J., Yim, D., Tanniru, M., Lim, S. (2017). Patient empowerment and engagement with a health infomediary. *Health Policy Technol*, *6*, 40–50.

Laplante-Lévesque, A., Jensen, L. D., Dawes, P., Nielsen, C. (2013). Optimal hearing aid use: Focus groups with hearing aid clients and audiologists. *Ear Hear*, *34*, 193–202.

Linacre, M., & Wright, B. (1993). Constructing linear measures from counts of qualitative observations. *Paper presented at the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics, Berlin, Germany*.

Luciano, J. V., Ayuso-Mateos, J. L., Aguado, J., Fernandez, A., Serrano-Blanco, A., Roca, M., Haro, J. M. (2010). The 12-item World Health Organization Disability Assessment Schedule II (WHO-DAS II): A nonparametric item response analysis. *BMC Med Res Methodol*, *10*, 1–9.

Maidment, D. W., Ali, Y. H., Ferguson, M. A. (2019). Applying the COM-B model to assess the usability of smartphone-connected listening devices in adults with hearing loss. *J Am Acad Audiol*, *30*, 417–430.

Maidment, D. W., Heyes, R., Gomez, R., Coulson, N. S., Wharrad, H., Ferguson, M. A. (2020). Evaluating a theoretically informed and cocreated mobile health educational intervention for first-time hearing aid users: Qualitative interview study. *JMIR Mhealth Uhealth*, *8*, e17193.

McAllister, M., Dunn, G., Payne, K., Davies, L., Todd, C. (2012). Patient empowerment: The need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Serv Res*, *12*, 157.

Mokkink, L. B., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., Terwee, C. B. (2019). *COSMIN Study Design Checklist for Patient-Reported Outcome Measurement Instruments*.

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., Leidy, N. K., Martin, M. L., Molsen, E., Ring, L. (2011). 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.

Poost-Foroosh, L., Jennings, M. B., Shaw, L., Meston, C. N., Cheesman, M. F. (2011). Factors in client-clinician interaction that influence hearing aid adoption. *Trends Amplif*, *15*, 127–139.

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

Rehm, J., Üstün, T. B., Saxena, S., Nelson, C. B., Chatterji, S., Ivis, F., Adlaf, E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *Int J Methods Psychiatr Res*, *8*, 110–122.

Riff, K. W. W., Tsangaris, E., Goodacre, T., Forrest, C. R., Pusic, A. L., Cano, S. J., Klassen, A. F. (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.

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale. *Rasch Meas Trans*, *24*, 1273–1275.

Schow, R. L., & Nerbonne, M. A. (1982). Communication screening profile: Use with elderly clients. *Ear Hear*, *3*, 135–147.

Tabachnick, B. G., Fidell, L. S., Ullman, J. B. (2007). *Using Multivariate Statistics* (Vol. *5*). Pearson.

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., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Meas Trans*, *20*, 1048–1051.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*, *60*, 34–42.

Üstün, T. B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., Epping-Jordan, J., Pull, C. (2010). Developing the World Health Organization Disability Assessment Schedule 20. *Bull World Health Organ*, *88*, 815–823.

Van de Winckel, A., Kozlowski, A. J., Johnston, M. V., Weaver, J., Grampurohit, N., Terhorst, L., Melvin, J. (2022). Reporting guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research—Explanation & elaboration manuscript. *Arch Phys Med Rehabil*, *103*, 1487–1498.

Weaver, B., & Maxwell, H. (2014). Exploratory factor analysis and reliability analysis with missing data: A simple method for SPSS users. *Quant Meth Psych*, *10*, 143–152.

Yeh, H.-Y., Ma, W.-F., Huang, J.-L., Hsueh, K.-C., Chiang, L.-C. (2016). Evaluating the effectiveness of a family empowerment program on family function and pulmonary function of children with asthma: A randomized control trial. *Int J Nurs Stud*, *60*, 133–144.

Yeh, M.-Y., Wu, S.-C., Tung, T.-H. (2018). The relation between patient education, patient empowerment and patient satisfaction: A cross-sectional-comparison study. *Appl Nurs Res*, *39*, 11–17.

Zimmerman, M. A. (1995). Psychological empowerment: Issues and illustrations. *Am J Community Psychol*, *23*, 581–599.