Validation of a new questionnaire measure of tinnitus functioning and disability for use in the UK: the Tinnitus Functional Index (TFI).

KATHRYN LOUISE FACKRELL, BSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

NOVEMBER 2015

ABSTRACT

The Tinnitus Functional Index (TFI) was developed in the USA as a standard for assessing the functional impact of tinnitus based on eight tinnitus-related domains. The finalised 25-item version was never formally validated. This PhD seeks to assess the psychometric properties of the questionnaire and evaluate its suitability as the tool of choice for use in the diagnostic and outcome assessment of tinnitus for clinical and research purposes in the UK.

The primary objectives were to (i) determine whether the TFI is reliable, (ii) verify its factor structure, and (iii) evaluate its responsiveness to treatment-related change. These objectives were evaluated in two UK studies. The first was a prospective multi-centre longitudinal validation study in which 255 NHS patients were recruited from audiology clinics to complete the TFI over four different time points in a nine-month period. The second was a retrospective analysis of data collected on the TFI and a battery of other health questionnaires from 294 members of the general public who had previously participated in two-centre randomised controlled trial of a novel tinnitus device. Approaches to psychometric analysis included classical and modern test theories, including Rasch measurement theory. Both approaches led to similar conclusions. Seven of the eight subscales were reliable and valid in both studies, although not as sensitive as the original developers proposed. Classical testing showed the auditory subscale to be reasonably reliable, but Rasch modelling indicated that it did not measure the functional impact of tinnitus. The overall factor structure was not confirmed. The sleep and auditory subscales did not relate to the other subscales and did not fit the model. My recommendation is to calculate the composite TFI score using only six subscales.

The sleep subscale should be scored separately and the auditory subscale should not be used.

ACKNOWLEDGEMENTS

I would firstly like to express my gratitude to my excellent supervisory team; Dr. Derek Hoare for always having time to answer my "stupid" questions, for challenging me every step of the way and for your unwavering support throughout. Prof Deb Hall for her continued support and guidance, for providing opportunities and encouragement to develop my understanding and knowledge. Dr Johanna Barry for your advice, guidance and enthusiasm in the project and statistics.

In addition to this, my thanks goes to the National Institute Hearing Research for funding my PhD and my colleagues at the NIHR Nottingham Hearing Biomedical Research Unit for their support and knowledge through the project. Specifically, I want to thank Ms Sandra Smith, for her infectious enthusiasm, life tips and for assisting me with my project.

I would also like to acknowledge the invaluable support and training in Rasch analysis given by Mr Mike Horton, without your help, I would still be lost in a sea of data trying come to grips with my analysis.

Finally, I would not have made it through this PhD without the help and support from my friends and family, you have distracted when I needed a break, provided nourishment and motivated to continue through the ups and downs.

In particular, I would like to thank my parents for their guidance, mathematical support, proof reading and patience and care throughout. I would like to expressly thank my younger sister, Victoria Fackrell, whose unwavering faith, support and encouragement has truly got me through this PhD, and for this I dedicate this thesis to you. You are truly one in a million.

PUBLICATIONS

Publications arising from this thesis

K. Fackrell, D.A. Hall, J.G. Barry, D.J. Hoare. (2016). Psychometric properties of the Tinnitus Functional Index (TFI); Assessment in a large UK research volunteer population. *Hearing Research*, 335, 220-235.

K. Fackrell, D.A. Hall, J.G. Barry, D.J. Hoare. (2014). Tools for tinnitus measurement: development and validity of questionnaires to assess handicap and treatment effects. In "Tinnitus: Causes, Treatment and Short & Long-Term Health Effects". F Signorelli and F Turjman (Eds). pp. 13-60. New York: Nova Science Publishers Inc.

Additional publications

E. Watts, K. Fackrell, J. Sheldrake, D.J. Hoare. Identifying problems associated with tinnitus (in prep). Ear and Hearing

K. Fackrell, C. Fearnley, M. Sereda, D.J. Hoare. (2015). Psychometric properties of the Hyperacusis Questionnaire (HQ) in a UK research population. *Biomed Research Internationa*l.

K. Fackrell, M. Edmundson-Jones, D.A. Hall. (2013). A controlled approached to the emotional dilution of the Stroop effect. *PloS one*, 8(11), e80141

K. Fackrell, D.J. Hoare, S. Smith, A. McCormack, D.A. Hall (2012). An evaluation of the content and quality of tinnitus information on websites preferred by General Practitioners. *BMC Medical Informatics and Decision Making*, 12;70.

S.K. El-Shunnar, D.J. Hoare, S. Smith, P.E. Gander, S. Kang, K. Fackrell, D.A. Hall (2011). Primary care for tinnitus: practice and opinion among GPs in England. *Journal of Evaluation in Clinical Practice*.

DECLARATION

I certify that this is my own work, except where indicated by referencing. No part of this thesis has been submitted elsewhere for any other degree or qualification.

Kathryn Fackrell Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under receiver operator Characteristic curve
BAI	Beck's Anxiety Inventory
BDI	Beck's Depression Inventory
BRU	Biomedical Research Unit
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
ES	Effect Size
GHTQ	Goebel-Hiller Tinnitus Questionnaire
ICC	Intra-class correlations
LoA	Limits of Agreement
MI	Modification Indices
MCID	Minimal Clinically Important Difference
MIC	Minimal Important Change
MLM	Maximum Likelihood method
NHS	National Health Service
NIHR	National Institute for Health Researech
PCA	Principal Components Analysis
PCQ	Perceived Change in tinnitus Question
PSI	Person Separation Index
QoL	Quality of Life
RMSEA	Root Mean Square Error of Approximation
ROC	Receiver Operator Characteristic
SAC	Scientific Advisory Committee of the Medical Outcomes Trust
SB-χ^2	Satorra-Bentler scaled Chi-square
SD	Standard Deviation
SDC	Smallest Detectable Change
SEM	Standard Error of Measurement
Stdx EPC	Standardised Expected Parameter Change
SRMR	Standardised Root Mean Square Residual

TFI	Tinnitus Functional Index
THI	Tinnitus Handicap Index
THQ	Tinnitus Handicap Questionnaire
TLI	Tucker-Lewis Index
TQ	Tinnitus Questionnaire
VIF	Variance Inflation Factor
WHOQOL-BREF	World Health Organisation Quality of Life-Bref

CHAPTER 1. INTRODUCTION

Clinicians rely on self-report questionnaires as a primary means of determining tinnitus severity and success of treatment (Meikle et al., 2007). Despite this, there is no single questionnaire assessment tool that is routinely used in clinical practice or research that optimally provides both a scale of tinnitus severity and an assessment of treatment-related outcomes. Over the last eight years, in an international collaboration led by clinicians in the USA, the Tinnitus Functional Index (TFI) was designed to be used as both a diagnostic tool and a sensitive measure to treatment-related change (Meikle et al., 2012). This thesis is concerned with evaluating the psychometric properties of the TFI for clinical and research use in the UK.

1.1. THESIS OVERVIEW

Chapter 1 describes the key literature on tinnitus. The first section provides a definition of tinnitus, conceptualising and categorising the different types of tinnitus and the symptoms associated with the levels of perceived tinnitus severity. The second section considers the challenges in assessing and quantifying tinnitus and the outcomes of tinnitus interventions. In this section, the role questionnaires currently fill within clinical and research settings is briefly discussed before a short introduction to Tinnitus Functional Index (TFI) (Meikle et al., 2012).

Chapter 2 provides comprehensive information on the methodology used for evaluating and validating the psychometric properties of questionnaires. Explicit quality criterion for developing and evaluating questionnaires are essential, therefore outlined within this chapter are four validation topics proposed by Terwee et al. (2007) as essential in questionnaire validation. The methods used to assess the psychometric properties are critically reviewed, highlighting the strengths and

limitations with the different approaches, before identifying the key methods that will be used within the PhD. Consideration is given to the challenges of disentangling the precise definition of acceptability criteria for all of the methods. An *a priori* acceptability criterion was established for all of the validation methods identified.

Chapter 3 provides an in-depth review of the development and validation steps of a select number of tinnitus questionnaires that are currently recommended for use in clinical practice and research, including the TFI. Each questionnaire is critically evaluated as a valid and reliable outcome measure using the quality criteria identified in Chapter 2. Throughout this chapter, information is provided on the content and psychometric properties of each questionnaire. A different version of this work has been published as a book chapter 'Tools for tinnitus measurement: development and validity of questionnaires to assess handicap and treatment effects' (Fackrell et al., 2014).

Chapter 4 presents a major prospective clinical study involving 255 tinnitus patients recruited from 11 NHS audiology clinics from across the UK. I collected questionnaire data (TFI, Tinnitus Handicap Inventory (THI), questions on tinnitus history, perceived tinnitus severity, treatment and perceived change in severity) on four separate occasions over 9 months to assess the reliability and validity of the TFI as a measurement tool, in particular its utility as an outcome measure.

Chapter 5 presents a retrospective analysis of a large research population of 294 participants drawn from the general public. This large dataset enables evaluation of the TFI factor structure and its ability to reliably measure the functional impact of tinnitus. Key results pertaining to the reliability of the questionnaire, structure and items, are highlighted and discussed in relation its use in a research population. A

different version of this work has been accepted for publication in Hearing Research (Fackrell et al., 2016).

Chapter 6 presents Rasch measurement theory using 540 participants from the previous studies (Chapter 4 and 5). Rasch analysis enables in-depth evaluations of the variability between person ability and item difficulty in all 25 items, each individual subscale and the proposed eight-factor second order model (providing a measure of the functional impact of tinnitus), (ii) the accuracy of the response options for each item, (iii) the item variance across groups, i.e. research and clinical populations, male and female, and finally (iv) to transform the raw data scores into a linear scores reflective of the model fit. Therefore, the Rasch measurement model provides a criterion for the structure of the responses, not just a statistical description.

Chapter 7 brings together all the key findings from the thesis, identifying the limitations and strengths of each study. I make a number of recommendations on the use of the TFI as the tool of choice for use in the diagnostic assessment of tinnitus in UK clinics and research. Consideration is also given to the practicality of the transformation scores identified in the Rasch analysis and the implications for use in clinical practice and research.

1.2. DEFINING TINNITUS

Tinnitus, derived from the Latin ‘tinnire’ meaning ‘to ring’, is the conscious experience of sound in the absence of a corresponding external auditory sound stimulation (McFadden, 1982). Tinnitus can be chronic and disabling and was, in the earliest medical references, described as “an extremely irksome discomfort, which leads to a profound sadness in affected individuals” (Itard, 1821; translation by Stephens (2000), p.443). This emotional impact can significantly differ between

individuals, complicating tinnitus assessment and management (Langguth et al., 2011).

Although almost all the population will have experienced a momentary sensation of 'phantom' sounds once in their lifetime (Zeman et al., 2012) it is estimated that in the UK alone, 10-15% of the general population are currently experiencing continuous persistent tinnitus (Davis & El Rafaie, 2000). Of this group, 20% find the experience sufficiently bothersome to seek guidance and treatment (Davis & El Rafaie, 2000). Tinnitus can affect anyone and occur at any age, but is more prevalent in older adults and in men (Davis & El Rafaie, 2000; Nondahl et al., 2011).

Tinnitus sounds are often described in the terms of 'ringing or buzzing in the ears', but these perceptual characteristics can vary between individuals (Andersson et al., 2005). Other descriptors include whistling, tinkling, clicking, roaring or tonal (a pool of undefined sound components) (Meikle & Taylor-Walsh, 1984; Stouffer & Tyler, 1990). These sounds can be perceived as either a continuous or an intermittent stream that varies in intensity, pitch and loudness. Location of the sound also varies, either presenting centrally in the head, in one or both ears. In general tinnitus is classified according to whether the source of the perceived noise has a physical sound source originating within the body (objective) (Lockwood et al., 2002) or can only be perceived by the individual and therefore lacks a specific origin (subjective) (Holmes & Padgham, 2011). Objective tinnitus accounts for ~5% of tinnitus case and is associated with abnormal functioning of the central auditory system, muscular or vascular abnormalities (Lockwood et al., 2002; Belli et al., 2012). Abnormal rhythmic muscle contractions or vibrations from turbulent blood flow pulsations in the middle ear can cause an audible 'beating' pulse (Lockwood et al., 2002; Henry et

al., 2005). In contrast, subjective tinnitus is unlikely to be explained by a single underlying pathological process. For the remainder of this report subjective tinnitus will be referred to simply as tinnitus.

The incidence of persistent spontaneous tinnitus is closely associated with pathogenesis of the auditory system (Baguley, 2002; Henry et al., 2005; Nondahl et al., 2011). The majority of tinnitus patients also have hearing impairments, such as hearing loss (Gopinath et al., 2010; Nondahl et al., 2011). Other risk factors include medical conditions such as obesity and arthritis (Nondahl et al., 2011), and acute intoxication from medications (Crummer & Hassan, 2004; Yorgason et al., 2006).

1.2.1. The multiple domains of tinnitus

Domains are defined as a cluster of symptoms, feelings or limitations that are theoretically similar. Tinnitus is associated with a wide variety of domains that impact on daily life such as sleep difficulties (Andersson et al., 1999; Miguel et al., 2014), concentration difficulties (Langguth et al., 2011; Pierce et al., 2012), cognitive impairments (Robinson et al., 2003; Stevens et al., 2007), impact on Quality of Life (QoL) (Kennedy et al., 2004; Zeman et al., 2014) and psychological well-being, such as stress, generalised anxiety and depression, (Hoffman et al., 2004; Gopinath et al., 2010; Nondahl et al., 2011). The degree to which tinnitus distress is perceived can depend on all of these domains.

In 1983, Tyler and Baker conducted an open-ended questionnaire study using a Nottingham tinnitus self-help association, where the responders listed the difficulties that resulted from their tinnitus, in order of importance or difficulty. Four domains of reported difficulties were observed: i) lifestyle (93% of cases), ii) general health (56%), iii) hearing (53%) and iv) emotional problems (69%). These four

domains are defined by fifteen of the most common difficulties attributed to tinnitus (Figure 1.1). Erlandsson and Holgers (2001) used the Nottingham Health Profile domains for subjective health status to predict tinnitus distress, and observed that, after the emotional domain (i.e. feelings of anxiety and depression), the sleep domain was the second highest predictor of tinnitus distress.

The importance given to these complaint categories have been documented by Sanchez and Stephens (1997, 2000). They assessed complaint responses of tinnitus clinic patients using the same questionnaire as Tyler and Baker (1983) and then reassessed the same patient responses 18 months to five years later. They found some similar complaint responses to Tyler and Baker (9/15 complaints) but with five general domains relating to difficulties with tinnitus (sleep, auditory, health, situational, and psychological). Sanchez and Stephens (1997; 2000) demonstrated that the largest proportions of difficulties were connected with psychological problems. The importance placed on these difficulties only slightly varied over time (> 2% variation; Figure 1.2).

Similarly, Kennedy et al. (2004) re-examined the difficulties identified by Tyler and Baker (1983) and redefined them into these general domains of tinnitus distress, but with the addition of a 'tinnitus-specific' domain, i.e. annoyance with tinnitus. In general, there appears to be even distribution of the difficulties identified by people experiencing tinnitus suggesting that tinnitus distress can be equally affected by a number of different conditions (Figure 1.2).

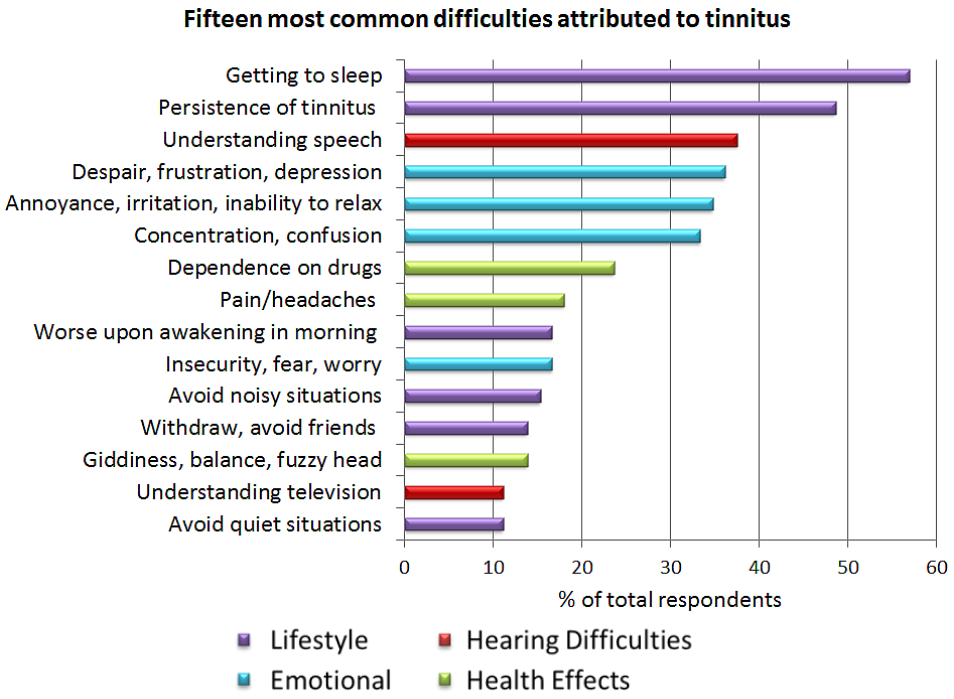

Figure 1.1. Difficulties reported by tinnitus patients in open-ended questionnaire (Tyler and Baker, 1983).

It is apparent that tinnitus is a highly heterogeneous condition with many possible co-morbid complaints that all impact on the distress perceived by the individuals (Langguth et al., 2011). A tinnitus questionnaire therefore needs to be broad in scope and cover multiple domains of tinnitus to be sensitive for general use in the tinnitus population.

1.3. THE MEASUREMENT OF TINNITUS

1.3.1. The importance of questionnaires

Evidence-based assessment and treatment of tinnitus is important (Department of Health, 2009). However, tinnitus is notoriously difficult to measure objectively because it is an experiential phenomenon and can significantly differ between individuals. Objective measures include matching the pitch and loudness of tinnitus to an external sound.

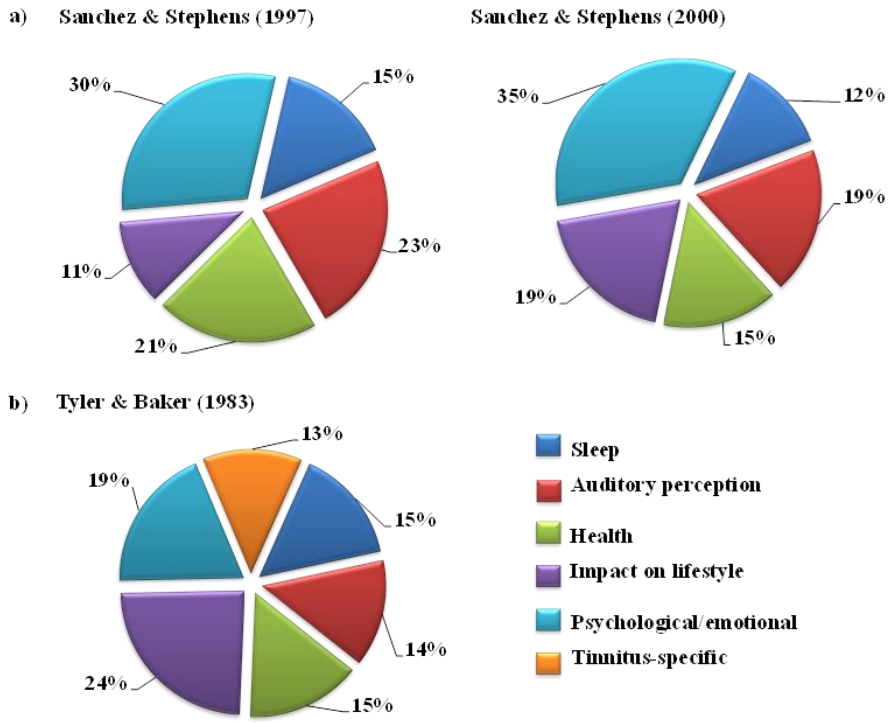

Figure 1.2. Complaint domains for tinnitus patients.
(a) The differences in complaint domains for tinnitus patients over five years. Reproduced from Sanchez and Stephens (2000). (b) Tinnitus difficulties reported by Tyler and Baker (1983) based on the six categories from Kennedy et al. (2004). Reproduced from Kennedy et al. (2004).

However, it is not possible to determine or predict tinnitus distress or clinical need based on these measures (Jakes et al., 1985b; Andersson et al., 2005; Zeman et al., 2011). Across individuals who report tinnitus of similar psychoacoustic characteristics (the same matched pitch and matched loudness level for example), there can be significant variability in self-reported handicap or the difficulties that are attributed to their tinnitus.

Self-report measures, especially questionnaires, are the primary way to quantify the severity of an individual's tinnitus and to evaluate the effect of a clinical intervention (Meikle et al., 2007). Although, it is not a requirement to include a standard questionnaire in clinical practice (Hesser, 2010), most clinicians (83%) in NHS audiology departments in England use validated questionnaires to assess

tinnitus in adults (Hoare et al., 2015), including those recommended by the Department of Health (2009); i.e. the Tinnitus Handicap Inventory (THI) (Newman et al., 1996) and Tinnitus Questionnaire (TQ) (Hallam, 1996; 2008) which were used by 68% and 14% of clinicians surveyed.

Questionnaires offer clinicians and researchers alike an approach to quantifying tinnitus by defining grades of tinnitus symptoms. This is vital for triaging patients effectively into the most appropriate interventions and for standardising selection criterion in research. Questions addressing different symptom domains can also form subscales, which, in turn help clinicians identify domains of concern and target interventions accordingly. Tinnitus questionnaires can be used as pre- and post-treatment outcome measures, providing evidence of changes in tinnitus severity. This is important within NHS services where there has been a move towards evidence-based commissioning; healthcare professionals need to demonstrate to third-party payers that their management approach is effective (NHS White Paper, 2010).

Tinnitus questionnaires are required to fill a dual role of measuring the functional impact of tinnitus and providing an assessment of treatment-related outcomes, but they fail to do both optimally. Some questionnaires are useful in the context of diagnostic assessment but inappropriate for use to sensitively measure change or vice versa (Meikle et al., 2008). For instance, the Tinnitus Handicap Questionnaire (THQ) was developed to evaluate treatment-related change, with less emphasis on its ability to discriminate between individuals (Kuk et al., 1990), whilst the THI was developed as a diagnostic tool and lacks the sensitivity to measure treatment-related changes (Meikle et al., 2007).

The current lack of a standardised validated questionnaire that is specifically developed for both purposes has led to difficulties in identifying and interpreting (1) the relative merits of the various interventions that are currently on offer or under investigation and (2) the most appropriate therapeutic approaches (Hesser, 2010). Consequently, the ability to confidently estimate the effectiveness of different tinnitus management approaches, compare across individuals or sub-populations between trials, and to confidently evaluate clinical audit across departments, is severely hindered by this lack of a standardised 'all purpose' questionnaire. The TFI (Meikle et al., 2012) has been identified as having the potential to address these problems. It was developed to provide (i) comprehensive coverage of the domains of tinnitus severity, (ii) reliable measurement of tinnitus severity that distinguishes between individuals and (iii) responsive measurement of change in tinnitus severity. The questionnaire underwent a systematic process of development to distil an initial item pool of 175 items through two prototypes (prototype 1 had 43 items, prototype 2 had 30 items) to arrive at a final questionnaire containing 25 items each mapping onto one of eight functional subscales. The final 25-item version has never been directly subjected to formal psychometric evaluation. The only assessment of validity and reliability was based on analysis of a subset of data collected for the 30-item prototype 2 of the questionnaire (Meikle et al., 2012).

1.4. AIMS AND OBJECTIVES

This PhD specifically aims to examine the ability of the TFI to:

(i) cover a broad range of problems and symptoms associated with tinnitus-related distress, in particular assess and confirm the reliability of the proposed eight-factor TFI structure reported by Meikle et al. (2012).

(ii) reliably measure tinnitus severity, distinguishing between individual differences in tinnitus-related distress from those whose tinnitus is ‘not a problem’ to those whose tinnitus is a ‘very big problem’.

(iii) responsively measure changes in tinnitus-related distress over long time intervals similar to those used in clinics and research, in particular assess the ability to measure small changes over time above measurement error

(iv) provide meaningful interpretations to the scores and the change in scores that would enhance clinical understanding of the scores and the importance associated with the change in scores.

CHAPTER 2. METHODOLOGY FOR EVALUATING PSYCHOMETRIC PROPERTIES OF A QUESTIONNAIRE

2.1. INTRODUCTION

Measurement in behavioural and social sciences is plagued with the problem that the attribute/trait/attitude in general cannot be directly measured because it is perceived by the individual alone. It is therefore important to assess whether patient-reported measures are reliable and valid measures of the underlying construct. Classical test theory plays a fundamental role is establishing reliability and validity, as does the hierarchical categories identified by Stevens (1946). Measurement instruments scales can be defined into one of four hierarchical categories: nominal, ordinal, interval or ratio. The majority of scales are assumed to employ interval-level measurement, such as zero to 100 scale and as such can be subjected to parametric tests and demonstrate approximate equality of intervals. A principal component of classical test theory is the concept of measurement error. The assumption is that every observed score is made up of the error in measurement and a "true" score that reflects the underlying construct. These need to be disentangled in order to reliably measure the degree to which tinnitus impacts on the different problem domains such as daily life, emotional well-being, concentration and sleep.

To be clinically useful any health-related questionnaire should ideally have the following properties: i) it should provide thorough assessment of the relevant presenting symptoms of the construct (i.e. the concept, attribute or variable that is the target of measurement) to support planning of treatments, ii) it should provide a reliable characterisation and quantification of individual differences in terms of the perceived severity (i.e. its diagnostic properties), and iii) it should have the capacity

to estimate responsiveness to changes in health status over time, to permit the assessment of the efficacy of different treatments and interventions (i.e. its evaluative properties) (Kirshner & Guyatt, 1985; Guyatt et al., 1992a; Jones & Kaplan, 2003; Hankins, 2008; Frei et al., 2011) (Table 2.1).

In order to identify individual differences and changes in scores on a given questionnaire, the item choice must be considered with respect to the stated aims and objectives of the questionnaire (Kirshner & Guyatt, 1985; Kennedy et al., 2004). Although there may be a tacit assumption that a diagnostic tool can also successfully serve as an outcome measure, if the item choice were predominantly based on the diagnostic properties, i.e. items show variability between individuals, then the questionnaire will not efficiently evaluate variability over time (Kirshner & Guyatt, 1985; Guyatt et al., 1992b; Meikle et al., 2007). This is also true of the choice of response scale. For example, coarse-grained, categorical units of measurement do not reliably detect the change in the construct, whilst fine-grained graduated units of measurement (defined by Stevens (1946) as an interval scale) are more sensitive to change and only measure the change between the category responses not the potential changes within the categories that can occur with categorical measurement units (Lipsey, 1983; Kirshner & Guyatt, 1985; Lipsey & Hurley, 2008).

It is challenging to encompass both discriminative and evaluative properties in a single measurement tool due to the contradictory nature of their needs (Meikle et al., 2007). Therefore it is important to be clear about the aims and elements behind the development of the questionnaire (Meikle et al., 2007; Terwee et al., 2007; Magasi et al., 2012). Measurement tools need to be psychometrically sound to ensure the integrity of study findings and clinical interpretations.

Table 2.1. A summary of the elements of diagnostic and evaluative properties

	Diagnostic properties	**Evaluative properties**
Item selection	Items should reliably reveal differences between individuals that are diagnostically relevant. Examined using content validity and response frequency.	Items should be sensitive to change tapping into areas related to health status change. Examined using content validity and response frequency.
Item reduction	Delete items with high inter-item correlations	Delete items that are insensitive to change to increase the likelihood of real treatment effects being observed.
Response options	The response scale should facilitate interpretation by not including too many options (i.e. 0 – 100). High resolution can decreases response reliability	The response scale should include sufficient graduations to show small changes. The amount of response options are proportional to the items sensitive to change. High resolution numerical scales (0-10) are recommended.
Reproducibility	High and stable variability between individuals	Small variability between replicate measures.
Validity	Cross-sectional construct validity	Longitudinal construct validity

Adapted from Kirshner and Guyatt (1985)

Therefore identifying the essential types of psychometric properties and the level of sufficient evidence is paramount.

The properties of validity, reliability, responsiveness and interpretability appear consistently across the literature and consequently form the basis (validation topics) for the methods outlined below (section 3.2). Although, there is some variation between the properties identified in the literature, for the most part, the majority are similar (Fitzpatrick et al., 1992; Lohr et al., 1996; Andresen, 2000; Frost et al., 2007; Mokkink et al., 2010a, 2010b; Scholtes et al., 2011; Reeve et al., 2013).

In 2007, Terwee et al. established a quality criterion framework using four overarching psychometric properties (validity, reliability, responsiveness and interpretation) to evaluate questionnaires. More recently, following Delphi (consensus) study conducted by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group (Mokkink et al., 2010a, 2010b, 2010c), this criterion was refined and updated to address the advancements in psychometric methodology (Terwee & Prinsen, 2014). Details are given in Tables 2.2 - 2.5. This framework is adopted throughout this thesis. Within this framework, the methodology was informed by evidence targeted at the particular psychometric property. The methodology and criteria for assessing validity, reliability, responsiveness and interpretation inform the work presented in Chapters 3, 4 and 5.

2.2. METHODS

2.2.1. Validity

2.2.1.1 Content validity

Content validity is the degree to which all facets of a questionnaire are relevant to the population of interest and comprehensively represent the important aspect of the underlying target construct of the questionnaire (Haynes et al., 1995; Frost et al., 2007; Terwee et al., 2007; Reeve et al., 2013). Facets include theoretical concept, item construction and content (questions), and recall period; all of which can affect the obtained data. Content validity is therefore fundamentally associated with evaluating aspects of questionnaire development and as such does not involve analysing any quantitative data from the questionnaire (Fitzpatrick, 1983; Streiner & Norman, 2008; Magasi et al., 2012).

Table 2.2. Essential criteria and statistical tests for evaluating the psychometric properties in over-arching concept of validity.

	Concept and definition	Statistical test	Essential criteria	Additional criteria
Validity	Content validity **"**The extent to which the most relevant and important aspects of a concept/population are accounted for by the tinnitus questionnaire"	—	Clearly defined evidence-based framework for aims, concepts, and items that reflect the population/construct being measured.*	—
	Structure validity#** "The degree to which the items measure a single underlying construct or multiple concepts within the questionnaire"	Exploratory factor analysis#	+ At least 50% of the variance explained + Factor loading > 0.40 + Adequate sample size ≥100	− Large unexplained variance (communalities < 0.5) − Large cross-loading if loading estimates > 0.3
		Confirmatory factor analysis**	+ S-B χ^2/df ≤0.2 p >0.05 + SRMR ≤ 0.07 + RMSEA < 0.06** + CFI & TLI > 0.95**	+ Factor loading ≥ 0.7 − Loading < 0.4 poor fit − Factor intercorrelations < 0.3 or > 0.85 poor fit − MIs >3.84 parameter cross-loading
	Construct validity Convergent "compare to other tinnitus questionnaires	Pearson's correlation Regression analysis	+ Hypotheses about expected direction and strength * + At least 75% of relationships in accordance to hypotheses*	Convergent Discriminant Excellent ≥ 0.60 Inadequate Acceptable 0.30 – 0.59 Inadequate ≤ 0.30 Excellent
	Discriminant "distinguish from other general health"	Kappa statistics		

S-B χ^2 = Satorra-Bentler Chi-square: SRMR = Standardised Root Mean Square Residual: RMSEA = Root Mean Square Error of Approximation: CFI = Comparative Fit Index: TLI = Tucker Lewis Index; + = ideal criteria for acceptable psychometric properties; – = indication of poor fit. * criteria from Terwee et al. (2007); ** newly introduced criteria by Terwee & Prinsen (2014);# Criteria first introduced by Uijen et al. (2012). Underline = methods used in empirical studies (Chapter 4 & 5)

Table 2.3. Essential criteria and statistical tests for the evaluating psychometric properties in over-arching concept of reliability.

	Concept and definition	Statistical test	Essential criteria		Additional criteria
Reliability	Internal consistency "The extent to which the items are inter-related or inter-correlated"	Cronbach's alpha (α) 95%CIs	> 0.95 0.80-0.95 0.70- 0.79 0.60- 0.69 < 0.60	Redundancy Excellent (+)* Acceptable (+)* Questionable Unacceptable	+ Reported 95% CIs within α >0.70 - 0.90 + Reported sample size > 30
		Inter-item correlations			+ Desirable range within 0.15 to 0.50
	Reliability "How well can patients be distinguished from each other, despite measurement errors?"	Intra-class correlations (ICC) 95% CIs*	+ Calculate ICC agreement if systematic error present* + Short time interval 1 to 3 weeks		+ Reported 95% CIs within 0.40 - > 0.70 + Identify unchanged' groups using global rating of health-status change
		Weighted Kappa coefficients*			
		Pearson's coefficients*	≥ 0.70 0.40 – 0.69 < 0.40	Excellent(strong)* Moderate Poor (weak)	+ ICC calculated in appropriate population
	Agreement "What is the extent of measurement error?" "How close are the individual scores from repeated measures?"	Limits of agreement (LoA)			+ Report LoA 95% CIs identify precision of LoA
		Standard error of measurement (SEM)	+ LoA with 95% agreement* + Reported SEM < LoA* + SDC comparable to LoA		+ SEM agreement reported if appropriate*
		Smallest Detectable Change (SDC)			+ SDC group reported*

ICC = Intra-class Correlation Coefficient; CI = Confidence Intervals; LoA = Limits of agreement; SEM = Standard error of measurement; SDC = Smallest detectable change; + = ideal criteria for acceptable psychometric properties. * criteria from Terwee et al. (2007). Underlined = methods used in empirical studies (Chapter 4 & 5).

Table 2.4. Essential criteria and statistical tests for the evaluating psychometric properties in over-arching concept of responsiveness.

	Approach and definition	Statistical test	Essential criteria	Additional criteria
Responsiveness "the ability to detect change in scores"	Floor and ceiling effects "The extent to which responses to the items are in the lowest and highest possible score options"	Item response distributions	+ ≤15% respondents achieve the lowest or highest possible score*	—
	Criterion based approach "Gold standard global rating of change as indicator of questionnaire ability to detect change"	Descriptive statistics and distribution of scores in global rating of change subgroups Spearman's rho correlations	+ *A priori* hypotheses of magnitude and direction of differences* + Spearman's coefficient ≥ 0.50** + At least 75% of *a priori* hypotheses confirmed**	—
		Receiver operator characteristic (ROC) curves	+ Area under the ROC curve; AUC ≥ 0.70*	+ Clearly define improved from unchanged
	Distribution-based approach "statistical characteristics as an indicator questionnaires ability to detect small change in scores, above error"	SEM SDC Effect size (ES)	+ Report SEM and SDC* + SDC/SEM *preferably* smaller than minimal important change (MIC) value*	+ *A priori* hypotheses of magnitude and direction of ES ≥0.20 Small effect 0.50 Medium effect ≥ 0.80 Large effect

ROC = Receiver Operator Characteristic; AUC = Area Under receiver operator Characteristic curve; SEM = Standard error of measurement; SDC = Smallest detectable change; ES = Effect size;; + = ideal criteria for acceptable psychometric properties. * criteria from Terwee et al. (2007); ** new criteria introduced by Terwee & Prinsen (2014). Underlined = methods used in empirical studies (Chapter 4 & 5)

Table 2.5. Essential criteria and statistical tests for providing interpretability of scores

	Concept and definition	Approach	Statistical test	Essential criteria
Interpretability	Interpretability of the scores "assign qualitative meaning to relevant subgroups of scores providing a grading system"	Distribution-based approach "statistical characteristics as indicator"	Quartile analysis	+ Means and SD for relevant subgroups* + Subgroups categories are distinct p> 0.05 + Integrate findings to identify clear distinct groups
		Criterion-based approach "mapping a gold standard s established cut-off criteria"	Descriptive statistics and distribution of scores in subgroups	
		Reference anchor of patient experience "classifying scores based on global rating of perceived problem"	ANOVA	
		Integrating approaches "the best possible range in each subgroup"	ROC curve analysis	+ Optimal threshold identified + AUC < 0.70
	Minimal Important change "assign qualitative meaning change in scores that reflects patient perceived improvement or worsening"	Criterion based approach "Using gold standard global rating of change to identify the minimal important change (MIC) score in the new tinnitus questionnaire""	Descriptive statistics and distribution of the differences in scores between time intervals	+ Identify difference scores between unchanged, improved or worsened groups*
			Spearman's rho correlations	+ Spearman's coefficient < 0.40
			ROC curve analysis	+ Identify optimal threshold value for improvements and worsening + AUC < 0.70 + Identify MIC values that are adjusted for baseline values
		Triangulation of estimates "Combining MIC estimates and distribution-based estimates from responsiveness and reliability"	Visual anchor-based MIC distribution graph	+ Identify an important change score (or range) that has external validity and accounts for error
			SEM	+ Prioritise MIC estimates*
			SDC/ LoA	+ MIC values *preferably* higher than SEM and SDC values
			ES	

ANOVA = analysis of variance; SD = standard deviations; ROC = Receiver Operator Characteristic; AUC = Area Under receiver operator Characteristic curve; MIC = Minimal Important Change; SEM = Standard error of measurement; SDC = Smallest detectable change; LoA = Limits of agreement; ES = Effect size; + = ideal criteria for acceptable psychometric properties. * criteria from Terwee et al. (2007). Underlined = methods used in empirical studies (Chapter 4 & 5)

Terwee et al. (2007) define five aspects of content validity evaluation which are used to evaluate content validity in my critical review of current recommended tinnitus questionnaires and the TFI (Chapter 3).

1) *Purpose of the measurement;* the intended purpose of the questionnaire should be clearly apparent (i.e. discriminative or evaluative).
2) *Target population;* the population in which the questionnaire was originally developed to measure.
3) *Concepts*; a clearly defined evidence-based conceptual framework to the items and concepts (domains/subscales) the questionnaire intends to measure. Patient and other expert opinions can ensure content is relevant and comprehensive (Magasi et al., 2012).
4) *Item selection and reduction;* items must be representative of the concept to which they measure and are relevant to the population of interest. Items can be sourced from theory, patient and clinician perspectives and previous questionnaires (Streiner & Norman, 2008; Reeve et al., 2013).
5) *Item interpretability;* item comprehension and the time of reference or time period. The readability of items should not be beyond the ability of 12 year olds to prevent misunderstandings and missing data. A time period should be clearly stated and justified (Reeve et al., 2013).

2.2.1.2 Structural validity and confirmatory factor analysis

Structural validity refers to the degree to which the items within the questionnaire are an adequate reflection of the construct being measured and whether all the items measure a single or multiple underlying constructs or a multi-dimensional construct, so that any composite questionnaire score is meaningful (Scholtes et al., 2011; Uijen et al., 2012; Mokkink et al., 2012). The structure of the questionnaire can be

determined using Confirmatory Factor Analysis (CFA) when an *a priori* hypothesis of the structure does exist. The TFI has a proposed eight factor structure (see Chapters 4 and 5). When no *a priori* hypothesis of the structure exists, it can be determined using Exploratory Factor Analysis (and Principal Components Analysis) (see Chapter 3).

For the TFI, CFA was used to verify the second-order factor (i,e, functional of impact of tinnitus), the eight first-order factors (intrusiveness, sense of control, cognition, sleep, auditory, relaxation, QoL and emotional), and the relationship between those constructs and the observed variables (Figure 2.1). CFA therefore provides a measure of the extent to which they reproduce the covariance in the observed variables-factor pattern (covariance matrix) (Brown, 2006; Brown & Moore, 2012). The parameters of the modelled factor solution are therefore specified in advance. These include the second-order factor (i.e. functional impact of tinnitus), number of first-order factors (i.e. eight), the observed variable-factor patterns (which items load onto which factor and whether these are constrained to only load onto one factor) and any anticipated covariance between factors/observed variables and any unique (error) variance.

Estimation methods

In normally distributed data, Weighted Least Squares estimation is used for ordinal data (categorical), whilst maximum likelihood parameter estimation is used for interval data (continuous). To adjust for non-normality in the data, the robust weighted least squares estimation (ordinal) or the maximum likelihood estimation method with adjusted Satorra-Bentler Chi-square (S-B χ^2; interval) are used to provide more robust parameter estimates and goodness of fit indices (Satorra & Bentler, 1994; Hu & Bentler, 1998; Bentler, 2006).

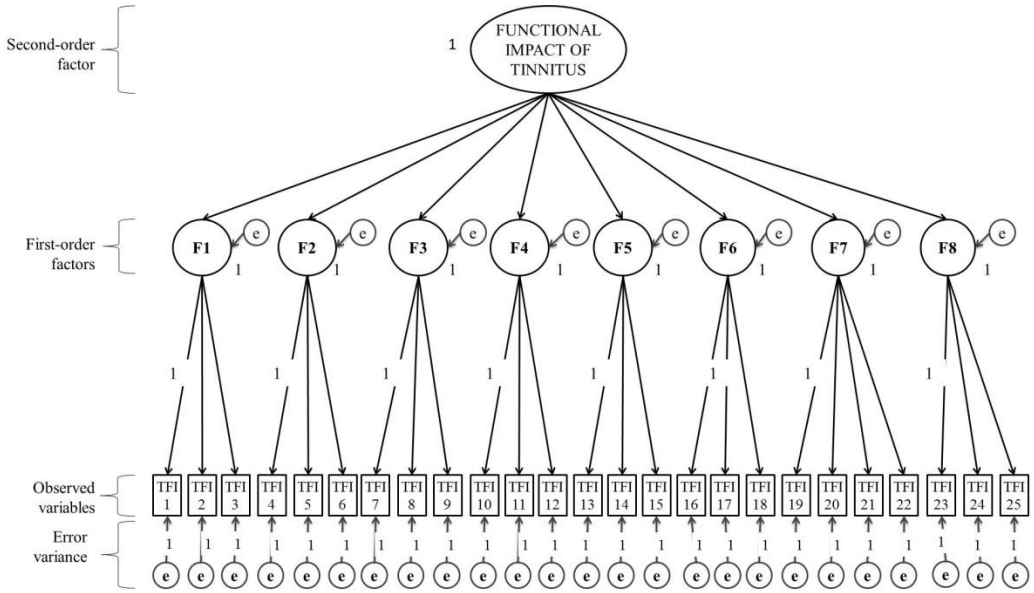

Figure 2.1. The *a priori* TFI factor structure.
The model represents the proposed relationships between the observed variables (items i.e. TFI 1), the first order factors (F1 to F8) and the second-order factor (Functional impact of tinnitus). Unidirectional black arrows (→) represent the direct effects of the second-order latent construct onto the first-order factors; (ii) and the direct effects of the first-order constructs onto the observed measures; (iii) 25 observed variables: Variance is fixed at 1 on second-order factor and the first item on each factor. Unidirectional grey arrows (→) represent the residual variance (e) associated with each variable. F1 = Intrusiveness; F2 = Sense of control; F3 = Cognition, F4 = Sleep; F5 = Auditory; F6 = Relaxation; F7 = Quality of life; F8 = Emotional; 1 = fixed variance; e = residual variance (error and uniqueness terms).

The TFI has an 11-point response scale which is considered as a continuous/interval scale (Muthén & Muthén, 2012), therefore the methods and criteria for establishing the fit of the model for interval data described below were used in Chapters 4 and 5.

To examine the TFI structure, the model is first estimated with just the first-order factors, allowing for examination of the correlations among the factors to be freely estimated and to assess model fit at this level. A second-order analysis should only be specified if the first-order factors correlate with each other and the model fit is deemed acceptable based on criteria.

The first-order factor covariance indicates the magnitude and pattern in which the factors are related to one another and the degree of overlap in content. Although a degree of overlap is expected between factors that are purported to be measuring the

same underlying construct, i.e. second-order factor (functional impact of tinnitus), highly correlated factors (> 0.85) indicate that they are not measuring distinct constructs from each other (poor discriminant validity). In contrast, weakly correlated factors (< 0.30) indicate that they are highly distinct from each other, and potentially measuring an alternative underlying construct. Ideally the correlations should be approximately 0.6 to indicate a single second-order factor (Brown, 2006; Brown & Moore, 2012).

The criteria for determining whether the eight-factor solution provides a good fit to the data is established through (i) goodness of fit indices, (ii) standardised parameter estimates (β; factor loadings) and (iii) misspecification parameter estimates.

(i) Goodness of fit criteria

Goodness of fit is determined using absolute fit indices and approximation fit indices. The absolute fit indices, S-B χ^2 (Satorra & Bentler, 1994) and Standardised Root Mean Square Residual (SRMR; Hu & Bentler, 1998; Bentler, 2006), both indicate the degree to which the discrepancies between the implied correlations (predicted by the model) and observed covariances deviate from the expected distribution values. The S-B χ^2 is assessed relative to the degrees of freedom, and this estimate has a critical ratio cut-off of ≤ 2.0. Alongside this, a large S-B χ^2 with $p < 0.05$ and SRMR that exceeds 0.07 (ideally it should be less than 0.06) taken together indicate poor fit and that the model should be rejected. Some caution is needed though in the interpretation of a significant S-B χ^2 value, since it is strongly influenced by sample size and variability in the data (Hu & Bentler, 1998; Brown, 2006). Approximation fit indices are used to assess additional parameters in the

model. Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and Comparative Fit Index (CFI; Bentler, 1990) assess the model fit to baseline assumptions. Values for both should exceed 0.90, and preferably exceed 0.95 (Hu & Bentler, 1999). Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) measure the discrepancy per degree of freedom in the model. Ideally, RMSEA should be less than 0.05, but values up to 0.08 are considered reasonable when the SRMR value is ≤ 0.06. RMSEA confidence intervals should also fall within the desired criteria (Hu & Bentler, 1998, 1999; Brown, 2006). These estimates reflect the recommendations from Terwee and Prinsen (2014).

It is important to note that although the majority of Structural Equation Model investigators refer to the specific *threshold-values* published by Hu & Bentler (1998, 1999) as the "gold standard" of "acceptable fit" in the approximation fit indices, this is not the case. Hu & Bentler (1998) cautioned against relying solely on the specific designated cut-off values as these thresholds-values may not reasonably work with different types of data, sample size and number of estimators. Although Terwee and Prinsen (2014) only provide recommendations based on the above fit statistics, consideration should be given to the factor loading and theoretical rational before making any adjustments to the model and in turn the questionnaire (Brown, 2006; Byrne, 2012; Fabrigar et al., 2010).

(ii) Standardised parameter estimates (factor loadings)

Factor loading estimates provide an indication of the magnitude and pattern of the relationship between the latent constructs (factors) and the observed variables (items). Factor loadings exceeding 0.7 indicate that the majority of the shared variance was explained by the latent construct. Loadings below 0.4 are associated with measurement error or poor explained variance and were taken to indicate a

potential source of poor model fit (Floyd & Widaman, 1995; Brown & Moore, 2012).

(iii) Misspecification parameter estimates

The Modification Index (MI) and Expected Parameter Change (EPC) identify any misspecification in the parameters of the model. Large modification indices exceeding 3.84 indicate that if a parameter was freely estimated, rather than fixed or constrained, the overall model fit would significantly improve (Brown & Moore, 2012). The EPC value indicates the approximate direction or magnitude by the parameter would change in subsequent analysis if adjustments were made. Together, these estimates, supported by conceptual foundations, are used to decide which parameter should be adjusted (MacCallum et al., 1992; Brown & Moore, 2012).

2.2.1.3 Construct validity

Construct validity refers to the extent to which a questionnaire actually measures the target construct it purports to measure (Westen & Rosenthal, 2003; DeVon et al., 2007). Construct validity is dependent on theory-based inferences about the meaning of test scores in relation to the hypothetical constructs (Cronbach & Meehl, 1955). For example, the extent to which observed associations with other measures are consistent with theoretically derived hypotheses concerning the relationship between the concepts being measured (Westen & Rosenthal, 2003; Reeve et al., 2013). The emphasis for construct validity is on clear specification of the theoretical construct, the identification of informative tests to measure the underlying theoretical construct and most importantly the hypothesised relations among the constructs (Cronbach & Meehl, 1955; Smith, 2005). Construct validity evaluations include two components: convergent and discriminant.

Convergent validity refers to the extent to which the construct of a new questionnaire corresponds with other questionnaire constructs that are theoretically similar. An observation of relatively strong relationship (high correlation coefficients) with measures of theoretically similar constructs is indicative of convergent validity. In contrast, discriminant validity assesses the extent to which the underlying construct of the new questionnaire can differentiate between constructs that are theoretically independent. Relatively weak relationships (low correlations) between theoretically distinct constructs suggest good discriminant validity (DeVon et al., 2007; Streiner & Norman, 2008).

Typically, construct validity is examined using bivariate correlations to compare two measures (Pearson's correlation coefficients and Spearman's Rho correlations) but methods such as multi-trait-multimethod matrix (Campbell & Fiske, 1959), multivariate regression, and Structural Equation Model techniques (such as CFA) can be used (Westen & Rosenthal, 2003). Bivariate correlations, partial correlations and multivariate regression analysis are used to assess convergent and discriminant validity in the empirical studies (see Chapters 4 and 5).

In the guidelines for evaluating assessment tools, Andresen (2000) provides criteria for the correlation coefficients used in convergent validity: ≥ 0.60 indicate acceptable, whilst ≤ 0.30 indicate inadequate evidence of convergent validity, but does not provide similar criteria for discriminant validity. For the purpose of this thesis, the criteria for discriminant validity coefficients are the reverse of the criteria for convergent, e.g. ≤ 0.30 indicates acceptable discriminant validity. Given that construct validity is based on inferences, the main recommended criterion is that the direction and magnitude of the relationships are predefined, to reduce potential retrospective bias. Terwee et al. (2007) state that for acceptable construct validity at

least 75% of the results should correspond with the *a priori* hypotheses. They also suggest that the sample size should be at least 50 participants.

2.2.2. Reliability

Reliability is concerned with identifying the magnitude of the measurement error inherent in each and every measurement and the ability of measurement instruments to yield consistent results (Streiner & Norman, 2008; Bartlett & Frost, 2008; de Vet et al., 2011; Gregory, 2014). Variations in measurement (error) are complex, varied and are related to random error that occurs because of person performance for instance or systematic errors that occur because of the measurement instrument (Streiner & Norman, 2008; Scholtes et al., 2011; de Vet et al., 2011). Of the two, systematic measurement error is particularly important to identify as the magnitude of measurement differences can inflate reliability coefficients (Bland & Altman, 1996a, 1996b; Hankins, 2008). The psychometric approaches to understanding reliability are complex and have been made all the more difficult to understand through the use of different terminology across authors for referring to the same approach. Here, the term "reliability" is used as an umbrella term for four measurement properties: internal consistency, standard error of measurement, reliability and agreement. It is important to note that the reliability estimates are not fixed properties of the questionnaire; these estimates reflect the characteristics of the scores and consequently the population in which they were measured (Feldt & Brennan, 1989). Therefore the population of interest should be considered when reviewing and evaluating questionnaires for use in clinical practice and research.

2.2.2.1 Internal consistency

All measurements with multiple items need to establish internal consistency (Bland & Altman, 1997). There is misunderstanding in the concept of internal consistency (Hogan et al., 2000). Internal consistency is often referred to as the extent to which all items in a questionnaire are measuring the same underlying construct (Streiner, 2003; Kottner & Streiner, 2010; Tavakol & Dennick, 2011). Although reasonably accurate, this does not clearly distinguish homogeneity from internal consistency. Homogeneity refers to the extent to which items are measuring a unidimensional structure. Internal consistency measures the extent to which the items are inter-related or inter-correlated and assesses the error variance associated with persons and items (Cronbach, 1951; Cortina, 1993; Clark & Watson, 1995). Whilst the two concepts are linked, they do not necessarily go hand-in-hand. Internal consistency is necessary for homogeneity but it does not provide evidence of dimensionality. The less the items relate, the more unique the item content, the higher the error variance (Cronbach, 1951; Cortina, 1993; Schmitt, 1996; Streiner, 2003). To facilitate interpretation of the item responses and the composite questionnaire score, it is desirable to have a high degree of internal consistency.

The most widely used measure of internal consistency is the coefficient alpha or otherwise known as Cronbach's alpha (α) (Cronbach, 1951). Cronbach's alpha is a generalisation of Kuder-Richardson formula 20 estimate (KR-20; Kuder & Richardson, 1937), designed to measures the inter-relatedness of polytomous items rather than dichotomous items (i.e. KR-20) (Cronbach, 1951; Streiner, 2003; Streiner & Norman, 2008). Cronbach's alpha estimates the proportion of variance explained by the average correlations among all of the items. The data is essentially splits in two in every possible way, calculating the correlation coefficient for each split,

therefore providing the average of all possible splits similar to KR-20. Cronbach's alpha is expressed as a coefficient between 0 and 1, with larger estimates (α>0.70) indicating highly related items with a large proportion of the variance attributed to a general factor with little item-specific variance (uniqueness) (Cronbach, 1951; Cortina, 1993; Streiner, 2003). That being said, coefficient alpha estimates need to be interpreted with caution. They are susceptible to the heterogeneity of the population, the number of items and presence of more than one trait being measured (Cortina, 1993; Clark & Watson, 1995; Schmitt, 1996; Streiner, 2003; Streiner & Norman, 2008; de Vet et al., 2011).

Cronbach's alpha is calculated based on the variance in the total scores, and as a consequence, the alpha estimates will be larger with a population that is varied (heterogeneous) than similar (homogeneous) since the variance will be larger (de Vet et al., 2011; Streiner, 2003). Furthermore, alpha estimates in short scales lack sufficient power to make meaningful interpretations about the relationships between the items (Schmitt, 1996), whilst in longer scales estimates are relatively invariant (Cortina, 1993; Peterson, 1994). Moreover, the dimensionality of the scale can inflate estimates. Although these estimates do not provide a measure of dimensionality, they do, fundamentally, assume homogeneity of the scale, and reliability will be overestimated in the presence of more than one underlying trait being measured (multidimensionality) (Cortina, 1993; Schmitt, 1996; Shevlin et al., 2000; Streiner, 2003; Kottner & Streiner, 2010).

If multidimensionality is suspected, factor analysis techniques should be conducted before alpha estimates are calculated for each component alone (subscale) (if they have a sufficient number of items >3 items) (Clark & Watson, 1995; Tavakol & Dennick, 2011; Agbo, 2014). Due to the complex nature of tinnitus, where there

are a wide variety of symptom domains, there is always a degree of heterogeneity among the items scales when the scale is designed to provide a composite score for an overall underlying construct (e.g. functional impact of tinnitus). Therefore as long as every item is an indicator of the whole, it can be expected that there are correlations between the items and the scale as a whole should be assessed (Streiner, 2003; Kottner & Streiner, 2010). Although, given the number of items in the scale, the alpha estimates for the whole should still be interpreted with caution. Nevertheless, alpha is a useful estimate of internal consistency and will be reported in the following chapters (Cortina, 1993; Schmitt, 1996; Streiner, 2003).

To enable interpretations of alpha estimates, the inter-item correlations (average and range) and alpha confidence intervals are reported (Clark & Watson, 1995; Charter, 2003; Iacobucci & Duhachek, 2003; Keszei et al., 2010).

Inter-item correlations show the strength of the association between all pairs of items. These correlations indicate the inter-relatedness of the item content, identifying items that are unrelated to the others or items that strongly related to each other forming a homogeneous structure or items that only related to certain other items to form separate factors. The average inter-item correlations, preferably within 0.30 to 0.60, provide information on the total variance, and can reflect the breadth of the construct being measured. For example, broader higher-order constructs such as tinnitus are expected to have lower average values. The range provides information on the breadth of the inter-relatedness, with a narrower range, preferably between 0.15 to 0.50, indicating less item uniqueness and high communality (Cortina, 1993; Clark & Watson, 1995). This is particularly important to assess since a wide range can indicate the presence of distinct dimensions which are known to inflate alpha estimates (Cortina, 1993; Shevlin et al., 2000).

Confidence intervals (CI) provide useful information about the precision of the alpha estimate and are more robust under varying conditions, such as test length and small sample size, than the single coefficient estimate (Iacobucci & Duhachek, 2003). The CI gives the range in which the true point of the parameter estimate should plausibly be located, accounting for likely measurement error that occurs (Charter, 1997; Cumming & Finch, 2005; Bland, 2015). For the coefficient alpha (α), the upper and lower confidence intervals limits (95%) for a given test with *k* items and a sample size of *n* are calculated as;

$$Lower\ limit = 1 - (1 - \alpha)F_{L\ (0.975),[df_1,df_2]} \quad (3.1)$$

$$Upper\ limit = 1 - (1 - \ \alpha)F_{U\ (1-F_{0.975})[df_2,df_1]} \quad (3.2)$$

In which *F* represents the F-ratio value for 95% CI in lower (L) and upper (U) limits, df_1= (n - 1), and df_2 = (n - 1)(k - 1) (Charter, 1997; Fan & Thompson, 2001; Iacobucci & Duhachek, 2003; Streiner & Norman, 2008).

In terms of sample size requirements for alpha estimates, large sample sizes from 250 to 500 are recommended for more precise coefficient estimates from 0.7 to 0.9 (95% CIs) (Iacobucci & Duhachek, 2003; Streiner & Norman, 2008).

Recommended alpha estimates that indicate a good questionnaire tool range anywhere between 0.70 to as high as 0.95 (Bland & Altman, 1997; Streiner & Norman, 2008; Gregory, 2014). Peterson (1994) found in his review of behavioural research that in the majority of cases a score α=0.7 was considered acceptable, whilst a score α<0.5 was poor. Nunnally (1978) recommended different coefficient alpha scores depending on the purpose of the tool. For example, for decisions at group level (research), α≥0.80 are recommended, and for decisions at an individual level (clinical practice), alpha estimates should be at least 0.95. These latter estimates have

however largely been discredited as extreme and more likely to represent large item sets and unnecessary redundancy (Peterson, 1994; Charter, 2003; Streiner, 2003; DeVon et al., 2007; Terwee et al., 2007; Streiner & Norman, 2008; Tavakol & Dennick, 2011).

The general consensus, and chosen criterion for this thesis, for acceptable internal consistency is that Cronbach's coefficient (α) estimates (and CIs) should be within 0.7 and 0.95 and the inter-item correlations should fall within the range of 0.15 to 0.50.

2.2.2.2 Reliability and agreement

The terms reliability and agreement parameters are often used interchangeably to define whether results are consistent in test-retest situations. However, they are distinct properties of measurement. Reliability compares the degree to which a measurement tool can distinguish people from each other despite measurement error, whilst agreement relates to the measurement error and the degree to which scores are identical or "agree" (described in section 2.2.2.4). Although conceptually different, both measure variance over time using the same tinnitus questionnaire assessing the same subjects (test-retest reliability/agreement). The following sections provide the methods and criterion to evaluate test-retest reliability/agreement.

An important assumption for test-retest reliability and agreement is that no real change in the underlying condition will occur between administrations, therefore the time interval length is particularly important to clearly identify. Time intervals are dependent on recall bias, the stability of the characteristic, condition or attribute being measured, and the target population. For example since tinnitus has a variable natural history, a shorter time interval should be used, but not too short as to enable

recall of materials and original answers. Time interval ranges of 2 days to 2 weeks have shown no significant differences in reliability or agreement estimates (Marx et al., 2003). Therefore, for tinnitus, an interval range of 1 to 3 weeks is considered appropriate (Marx et al., 2003; DeVon et al., 2007; Terwee et al., 2007; Keszei et al., 2010; Scholtes et al., 2011; Mokkink et al., 2012). Longer time intervals for reliability and agreement can be supported by the use of a global rating of health-status change (improvement) to identify individuals who have experienced "no change" (an anchor-based method discussed in section 2.2.3).

The sample size required for test-retest reliability is somewhat controversial, with the recommended number ranging from <50 to >400 (Nunnally, 1978; Charter, 2003; Shoukri et al., 2004; Terwee et al., 2007; Streiner & Norman, 2008). Here, it is generally taken that 50 participants is sufficient (see Terwee et al., 2007; Streiner & Norman, 2008).

2.2.2.3 Test-retest reliability

The amount of variability between individuals' scores and the amount of measurement error are both known to affect the ability to discriminate between individuals. For example, large measurement error in comparison to the variability between individuals reduces the ability to discriminate between individuals. There is too much measurement error to distinguish whether the differences between individuals are a product of error or genuine differences in true score. Therefore, the magnitude of the reliability coefficient is related to the variability of scores between individuals (true score) ($\sigma^2_{s)}$ and total variance, i.e. the variability between individuals and measurement error (σ^2_{err}), such that:

$$Reliability = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{err}^2} \quad (3.3)$$

Therefore reliability is expressed the proportion of variance that can be attributed to the true score. The formula is the basis for all intra-class correlations (ICCs), the primary recommended method for test-retest reliability.

ICCs are the ratio of all variances ranging from 0 (unreliable measure) to 1 (excellent reliability). They are sensitive to the occurrence of both random error and systematic bias between administrations. ICCs account for a systematic change in the mean and, as consequence, require all the observations to be close to identical. The ICC formula (3.3) is a measure of the consistency of the scores between the administrations ($ICC_{consistency}$) and although it does account for random error, including large differences in variance of the means, it does not necessarily identify potential systematic differences due to time. The following ICC formula is calculated in a way in which systematic bias between time can be accounted for.

$$ICC_{agreement} = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{time}^2+\sigma_{err}^2} \quad (3.4)$$

The additional term of variability between time points (σ^2_{time}) is now included in the denominator. This ICC is confusingly named "agreement" of the scores between the administrations, even though it is a reliability parameter. It is not a measure of test-retest agreement (2.2.2.4).

ICCs are sample dependent, in that a homogenous sample (similar scorers) will have less variability between scores and the magnitude of the ICC will therefore be smaller, indicating less reliability to differentiate (Griffiths & Murrells, 2010; de'Vet et al., 2011; Altman & Bland, 2013; Bland, 2015). ICC estimates should be calculated in the specific population for which the questionnaire is intended.

To determine the accuracy of ICC estimates, it is recommended that CIs for the ICCs are reported (Shrout & Fleiss, 1979; Atkinson & Nevill, 1998; DeVon et

al., 2007; Bartlett & Frost, 2008; Streiner & Norman, 2008; Kottner et al., 2011). This involves Fisherz_R transformations which identifies the upper and lower limits of standard error as ICC estimates. Initially, the ICC estimate and standard error are *z*-transformed to remove skewness from the standard error before returning the identified limits back into ICC coefficients.

Interpreting ICCs is reasonably straightforward. The ICCs estimate the proportion of variance in observed scores that can be attributed to the true score, generally ranging from 0 to 1. Therefore, an ICC estimate of 0.86 for example indicates that approximately 86% of the variability in the observed score can be attributed to the true score (Weir, 2005; Streiner & Norman, 2008; Bland, 2015). Establishing acceptable estimates for test-retest reliability is complicated by the fact that once again there are many suggestions in the literature regarding the interpretation of the estimates. The minimum standards for test-retest estimates range anywhere from 0.60 to 0.95, depending on the purpose of the tool, although the practicality of achieving these high scores has been questioned (Nunnally, 1978; Shoukri et al., 2004; Terwee et al., 2007; Streiner & Norman, 2008; Kottner et al., 2011; de Vet et al., 2011). Based on recommendations by Fleiss et al. (2003), an ICC estimate <0.40 should be considered poor, between 0.40 and 0.74 moderate, and ≥0.75 excellent (see also Andresen, 2000). Terwee et al. (2007) recommend that an ICC estimate >0.70 indicates high test-retest reliability.

In Chapters 4 and 5, ICC estimates and CIs were reported for assessments of test-retest reliability. The criteria standards set by Fleiss et al. (2003) and Andresen (2000) were applied to both, with one exception that the upper limit reflects the recommendation from Terwee et al. (2007).

Others methods for test-retest reliability include Pearson's correlation coefficients (denoted by *r*) and Kappa coefficient (Terwee et al., 2007; Streiner & Norman, 2008; de Vet et al., 2011). Pearson's correlations measure the relationship between variables using regression-based analysis whereby it measures the extent of which the scores can be fitted by a straight regression line, irrespective of the slope of the line. A "perfect" correlation does not require a 45^{o} line. It just measures how closely the scores follow the line, reflecting any linear function between two tests. As such, it fails to account for systematic differences that can be inherent in measurement over time, i.e. a systematic large change in mean for all subjects at test two. Therefore, although Pearson's can account for random error, if systematic errors are present, the coefficient correlation will overestimate reliability and should be interpreted with caution (Till, 1989; Atkinson & Nevill, 1998; Andresen, 2000; Streiner & Norman, 2008; de Vet et al., 2011). Pearson's correlations were not used to evaluate reliability of the TFI (Chapters 4 and 5), but were accepted as evidence of reliability in the critical review (Chapter 3). Kappa coefficients are reliability parameters intended for categorical variables. The kappa coefficient (also known as Cohen's kappa) examines the proportion of agreement whilst adjusting for chance between two observations (i.e. abnormal/normal) and therefore is not relevant within this thesis because all the questionnaires being used or evaluated have more than two response categories.

2.2.2.4 Test-retest agreement

The underlying assumption is that the difference between repeated measures should be zero without any interventions (or any changes to underlying condition). The preferred measurement parameters for evaluating agreement are Bland & Altman (1986, 2010) Limits of Agreement (LoA), the standard error of measurement (SEM),

and Smallest Detectable Change (SDC) (de Vet et al., 2006b, 2006a; Terwee et al., 2007, 2009). For other references see (Streiner, 2003; de Vet et al., 2006a, 2006b; Terwee et al., 2007, 2009; Bland & Altman, 2010; Mokkink et al., 2010b; de Vet et al., 2011; Gregory, 2014). All these methods were used in Chapters 3, 4 and 5.

2.2.2.4.1 Limits of Agreement (LoA)

The seminal work on LoA by Bland and Altman (1986) explicitly separates bias (i.e. consistent tendency for one method to exceed the other) and random error from the observed scores between two repeated measures (Bland & Altman, 1986; Streiner & Norman, 2008; Bland & Altman, 2010; Griffiths & Murrells, 2010). The limits essentially indicate the extent to which scores can vary within people who experience a stable underlying construct. The LoA between two measures are summarised by calculating the mean difference between the two measures (systematic bias) and standard deviation of those differences (random error). The mean differences in scores are initially plotted against the subject mean to visually inspect the direction and magnitude of the data around the zero line. The assumption is that if there was complete agreement between the scores, the mean difference between the scores of two measures would be zero and assuming that the difference scores are normally distributed then 95% of differences would be within ±2 standard deviations of the zero difference score. LoA were calculated as

$$LoA = \bar{d} \pm 1.96 \times SD_{diff} \quad (3.5)$$

where $\bar{d}$ represents the mean difference in scores between the two administrations, the ±1.96 represents two standard deviations, whilst the SD_{diff} represents the mean difference in standard deviation. To identify the SD of the difference (and total variance) across multiple measures per individual, Bland and

Altman (2007) recommend calculating a one-way ANOVA using the differences scores as a response. The standard error (SE) and 95% CIs for the lower and upper limits of agreement estimates were calculated as;

$$SE\ of\ \bar{d} = \sqrt{(SD_{diff}^2/n)} \tag{3.6}$$

$$SE\ of\ \bar{d}\ \pm SD_{diff} = \sqrt{(3\ \times SD_{diff}^2/n)} \tag{3.7}$$

$$LoA\ CI_{95} = LoA\ \pm (1.96 \times SE\ of\ \bar{d}\ \pm SD_{diff}) \tag{3.8}$$

where *n* is the sample size. This provides an indication of how precise the LoA estimates are (Bland & Altman, 2010). The limits estimates provide the degree of measurement error and any change in score greater than or equal to the value would represent real change in 95% of people. 95% agreement was taken as an indication of high test-retest agreement (Bland & Altman, 2010).

Finally, an important assumption of the LoA is that since the variation between subjects has been removed leaving only measurement error, the differences should be normally distributed, and not affected by increasing mean values (Bland & Altman, 1999, 2010). The LoA values represent the entire range in scores (Bland & Altman, 2010). There should be no clear relationship between the variability and the magnitude and this can be tested using rank correlation coefficients (Kendall's τ). If no relationship is present (coefficient is close to zero), then the differences are assumed to be normally distributed. If there is a relationship, in order to satisfy the assumption of constant SD_{diff}, a logarithmic transformation of the data is recommended before calculating the limits of agreement and then antilog the data to provide interpretation of the values with the above criteria would still apply (Bland & Altman, 1996a, 1996b).

2.2.2.4.2 Standard Error of Measurement (SEM)

Measurement error is expressed as the standard error of measurement (SEM) and provides an absolute measure of the precision of individual scores within a test (Weir, 2005; Streiner & Norman, 2008). SEM is not only an integral parameter in reliability assessments but is also integral for identifying error in minimal important change scores since this knowledge enhances the clinical relevance of the outcome (section 2.2.4) (Cella et al., 2002; Crosby et al., 2003, 2004; Wyrwich, 2004; Terwee et al., 2009; Scholtes et al., 2011; Wyrwich et al., 2013). SEM is regarded as being relatively constant and so independent of the variability in the population. With the possible exception of extreme scores, the same SEM should apply for all scores across the scale (Wyrwich et al., 1999; Weir, 2005; de Vet et al., 2011; Gregory, 2014).

The most common way of calculating SEM is directly related to reliability estimates and is expressed as:

$$SEM = \sigma\sqrt{1 - r} \tag{3.9}$$

where σ is the standard deviation of the observed scores and r is the reliability of the measurement (Stratford & Goldsmith, 1997; Wyrwich, 2004; Frost et al., 2007; Streiner & Norman, 2008; de Vet et al., 2011; Gregory, 2014). The reliability (r) is expressed as either the ICC estimate or alpha estimate for a given questionnaire from another study (Wyrwich et al., 1999; de Vet et al., 2011). The ICC calculated in one study is applied to the standard deviation from a different population. The variability in the population in which the ICC originates influences the magnitude of the ICC and therefore if the heterogeneity of new population was not similar (or the systematic bias was not accounted for), the obtained SEM value is

liable to either under- or over-estimate the degree of measurement error (Weir, 2005; de Vet et al., 2006b, 2011). Additionally, the SEM is the extent of measurement error between repeated measures. Therefore the use of alpha estimates seems redundant given that the estimates are based on a single measure (de Vet et al., 2011). The following two SEM formulas are based on the random error variance (the mean square error term) one of which can be identified from the $ICC_{consistency}$ formula (3.3), the other makes use of the information provided in the limits of agreement (3.5);

$$SEM_{consistency} = \sqrt{\sigma^2_{err}} \quad (3.10)$$

$$SEM_{consistency} = SD_{diff}/\sqrt{2} \quad (3.11)$$

where σ^2_{err} is the random error and SD_{diff} is the SD of the mean differences, and $\sqrt{2}$ is included to account for the error in both of the measurements. These formulas are independent of the specific ICC and therefore the estimates are not susceptible to problems mentioned above, they are consistent across different studies. Nevertheless, neither formula accounts for the systematic bias in the measurement. The final formula includes the variability between time points (σ^2_{time}) and the variability caused by random error (Weir, 2005; de Vet et al., 2006b; Terwee et al., 2009);

$$SEM_{agreement} = \sqrt{(\sigma^2_{time} + \sigma^2_{err})} \quad (3.12)$$

This final formula is recommended for use where possible as systematic bias, including the variability between time, is considered by some as part of measurement error (Weir, 2005; Terwee et al., 2007, 2009). Otherwise, either of the formulas for $SEM_{consistency}$ (3.10 - 3.11) are acceptable to identify measurement error (de Vet et al., 2006b; Terwee et al., 2007). Unlike other reliability parameters such as ICCs, SEM

is expressed in same units of measurement as original questionnaire scores and is therefore are reasonably easy to interpret (Stratford & Goldsmith, 1997; Weir, 2005; Streiner & Norman, 2008; de Vet et al., 2011). Essentially, the SEM value is dependent on the score range of the questionnaire of interest, but the general assumption is that larger SEM scores equal lower reliability.

2.2.2.4.3 Smallest Detectable Change (SDC)

Smallest Detectable Change (SDC) is derived from the SEM and, similar to LoA, it represents the smallest change in score that is beyond measurement error and must be overcome for a change to be considered real (de Vet et al., 2006a; Terwee et al., 2007, 2009; Scholtes et al., 2011; de Vet et al., 2011). It is calculated as;

$$SDC_{ind} = 1.96 \times \sqrt{2} \times SEM \quad (3.13)$$

This represents the SDC at an individual level (SDC_{ind}) that can distinguish measurement error. The SDC score should be comparable to the LoA score to be deemed an acceptable score. Alternatively, Terwee et al. (2009) have proposed multiplying the SEM estimate by 4 to identify the smallest change as this accounts for the variability in individual scores over time and both Type I and Type II errors. This score is expected to be similar to or slightly higher than the estimates from SDC_{ind} (both are reported in Chapter 4).

For group-wise assessments, the measurement error is reduced since averaging the results gives a more reliable interpretation of the amount of fluctuations in the scores. To identify the SDC in a group (SDC_{group}), the SDC for individuals (SDC_{ind}) that is calculated using formula above (3.13) should be divided by √n (Terwee et al., 2007; de Vet et al., 2011). The criterion for an acceptable SDC_{group} estimate has not been identified. Therefore, the results of this calculation are treated with caution and

no inferences are made. No similar group-wise adjustments have been proposed for the LoA.

2.2.3. Responsiveness

Responsiveness refers to the questionnaire instruments ability to detect a change in scores, under circumstances in which the questionnaire is intended for use (Lipsey, 1983; Scientific Advisory Committee of the Medical Outcomes Trust, 2002; Terwee et al., 2007; Revicki et al., 2008; de Vet et al., 2011; Reeve et al., 2013). To be considered a responsive measure for outcome assessment, the questionnaire needs to reliably identify small changes in scores that truly reflect changes in the construct of interest (Lipsey, 1983; Jacobson et al., 1999; Lipsey & Hurley, 2008). As a consequence, reliability estimates, SEM and SDC, are a necessary component of responsiveness (section 2.2.3). These parameters fall under distribution-based methods. Additional parameters are recommended to assess the ability of the item and scale to assess change (Terwee et al., 2007; de Vet et al., 2011). These include item response distributions, which identify the number of items exhibiting floor and ceiling effects, and a criterion-based approach using anchor-based methods which identifies a global rating of perceived health-status change and receiver operator characteristic (ROC) analysis.

2.2.3.1 Identifying the items ability for change

Floor and ceiling effects compromise the responsiveness of a questionnaire to meaningful changes (Terwee et al., 2007; Lipsey & Hurley, 2008). They are assessed at item level by examining the frequency of the responses to the lowest and highest possible scores, respectively. Floor effects limit detection of reductions in the functional impact of tinnitus. Ceiling effects limit detection of increases in the

functional impact of tinnitus. Terwee et al. (2007) identifies potentially problematic items as those rated at the lowest or highest possible response option (e.g. 0 or 10 on a 10-point scale) by more than 15% of respondents.

2.2.3.2 Identifying the scales ability for change

The criterion-based approach uses a "gold standard" measurement as an indicator of adequate changes in scores in the new questionnaire measure being evaluated. It examines the degree to which changes in scores on the gold standard are suitably reflected in the change in scores of the new questionnaire (de Vet et al., 2011).

Although some work has been done to establish responsiveness of tinnitus questionnaires (see Chapter 3), there is no "gold standard" that is reliably responsive to small changes. In cases such as this, anchor-based methods are often used as a comparison measure of change (de Vet et al., 2011; Wyrwich et al., 2013). Anchor-based methods are essentially external indicators that provide an interpretable measure of the optimal points of change or what patients consider an important change which can be used to map changes in scores. They can also determine the perceived level of the problem at baseline (see section 2.2.4.1). See Crosby et al. (2003) for other anchor-based methods. The most commonly reported anchor-based method for assessing change in scores is the global rating of perceived health-status change first introduced by Jaeschke et al. (1989). Participants would rate how much change they perceive in health status over a particular period of time or between two time points, such as baseline and end of study, on a 7- or 15-point scale (Jaeschke et al., 1989; Revicki et al., 2008; Wyrwich et al., 2013). In Chapter 4, participants rated their change in tinnitus on a 7-point scale (much improved (3), moderately improved (2), slightly improved (1), no change (0), slightly worse (-1), moderately worse (-2), and much worse (-3)).

To evaluate responsiveness using this approach, the mean change scores (SD) for the TFI between the baseline and follow-up assessments are calculated for each anchor rating group, and then inspected to ascertain the magnitude and direction of the changes in each group. To reduce risk of bias, a pre-defined hypothesis should be established about the expected (magnitude and direction) differences between groups. The change scores are then correlated with the anchor ratings using Spearman's rank correlation coefficient to ascertain the strength of the relationship between the scores. There is no apparent criterion for an acceptable magnitude of the correlation coefficient, and given that measurement error increases with the number of measurements, then the correlations coefficients are not expected to be as high as those mentioned previously in this chapter (section 2.2.2).

The final stage of assessing responsiveness is to use another anchor-based method, the ROC curve analysis, to establish whether the change scores based on global ratings of change can discriminate 'improvement' from 'no change' and 'worsening' of symptoms.

Based on external criteria, the anchor groups are normally divided into two classifications: (i) 'no change' versus 'improvement', and (ii) 'no change' versus 'worsening'. ROC curve analysis combines information on sensitivity (true positive rate) and specificity (true negative rate) to detect the threshold value used to classify individuals as 'improved' (Riddle et al., 1998; Eng, 2005). Sensitivity would refer to the proportion of improved (or worsened) participants, according to the anchor, that are correctly classified according to their TFI score as improving (or worsening). In contrast, specificity would refer to the proportion of participants correctly classified by the TFI as showing no change (Eng, 2005; Uslu et al., 2008; de Vet et al., 2011).

ROC curve analysis provides a means to evaluate the sensitivity and specificity over the range of values.

The ROC curve plots the sensitivity (y axis) vs 1 – specificity (x axis; which flips the graph horizontally) with the Area Under the receiver operator Characteristic curve (AUC) representing the TFI's ability to discriminate between people who have improved (or worsened) from those who are unchanged. When presented with random scores, the AUC would be equal to the probability that the TFI will correctly identify improvement (Eng, 2005; Zou et al., 2007). If the ROC curve was a 45˚diagonal line from 0,0 to 1,1, then the AUC would be 0.5 indicating that there was 50% probability that the questionnaire will be unable to identify people who have improved from those who do not, i.e. poor discrimination. A more prominent curve is therefore equivalent to a more accurate test, and AUC values of above 0.7 are deemed appropriate for identifying changes in the scores (Eng, 2005; Zou et al., 2007).

ROC analysis is based on two classifications, therefore two separate ROC curves were calculated to examine the TFI's ability to identify improving and worsening symptoms. It is important to highlight that this analysis also plays an integral part in the identification of a minimal important change score (section 2.2.4).

Distribution-based methods are often used to identify and interpret change in scores (Wright & Young, 1997; Wyrwich et al., 2002; Crosby et al., 2004; Wyrwich, 2004; Eton et al., 2004; Yost et al., 2011). Distribution-based methods are based on statistical properties of the sample and include methods such as the SEM (described in section 2.2.2.4), effect size (ES) and a paired t-test. See Crosby et al. (2003) and Husted et al. (2000) for other methods. Some recommend ES as a measure of responsiveness (Husted et al., 2000; Scientific Advisory Committee of the Medical

Outcomes Trust, 2002; Revicki et al., 2008), while others have recommend SEM (Terwee et al., 2007; Mokkink et al., 2010c). The SEM would provide much needed information on the precision of the TFI measurement. But conventional standards, such as one SEM is equivalent to a minimal important change score, should be interpreted with caution. The other methods provide only information on the magnitude of the score or the significance of the change. They do not determine the validity of the change score and/or account for any error in measurement (Mokkink et al., 2010b, 2010c; de Vet et al., 2011; Mokkink et al., 2012). One way in which this problem with ES could be elevated is to define *a priori* hypotheses about the expected magnitude of the differences between the groups and use a comparison instrument or global rating scale.

During the development of the TFI, ES was calculated in this way and so it is conducted here for comparative purposes (Chapter 4). Using Lipsey's criterion group approach (Lipsey, 1983), in which individual TFI scores are stratified based on their global ratings of change, the ES was calculated for those groups expected to differ and compared to *a priori* hypotheses (Chapter 4). The ES is calculated as;

$$ES = \frac{x_1 - x_0}{\sqrt{\frac{\sum(x_0 - \bar{x}_0)^2}{n-1}}} \quad (3.14)$$

where X_0 refers to pre-test score and X_1 is the post-test score, divided by the SD of the pre-test scores (Crosby et al., 2003). The standard criterion for the ES estimates are: ≥0.20 is a small effect, 0.5 is a medium effect, and ≥ 0.80 is a large effect (Cohen, 1988). ES estimates can be used as evidence of identifying change if the direction and magnitude follow the expected pattern, but it is not recommended for use as a standalone evidence of change (Lipsey & Cordray, 2000; de Vet et al., 2011; Meikle et al., 2012; Mokkink et al., 2012).

2.2.4. Interpretability

TFI scores should be easily understood to enable clinicians and researchers to clearly identify and quantify the functional impact of tinnitus (Scientific Advisory Committee of the Medical Outcomes Trust, 2002; Mokkink et al., 2012). Despite it being an important property of a questionnaire, interpretability can be somewhat overlooked. Attention tends to focus on methods for identifying minimal changes in scores that are statistically and clinically relevant (section 2.2.4.2) (Wyrwich et al., 1999; Crosby et al., 2004; de Vet et al., 2006b; Terwee et al., 2007, 2009; Mokkink et al., 2012). Guidance for the methods of interpretation is limited, but tend to highlight the distribution of scores and descriptive statistics (median, means, SD and range) from "relevant" population groups, including a normative population (Terwee et al., 2007; Reeve et al., 2013). This can provide clinicians with insights on what the different scores mean for different patient subgroups compared to each other and to the normative population (Lohr et al., 1996; Scientific Advisory Committee of the Medical Outcomes Trust, 2002; Terwee et al., 2007). Terwee et al. (2007) suggest using groups based on clinical diagnosis to aid interpretation of the scores, but do not provide information on how to use these comparisons to apply qualitative meaning to the scores. It would be beneficial, especially in tinnitus, to be able to interpret scores according to categories that define the different levels of impact as this enables clinicians and researchers alike to identify and quantify tinnitus severity.

2.2.4.1 Identifying a grading system

In order to facilitate interpretation of TFI scores and define grades of tinnitus impact, I used quartile analysis and anchor- based methods with the aim of cross validating and defining a final score range for each category of tinnitus impact. Quartile analysis is a distribution-based method in which patient population data is divided

into categories based on the distribution of the total score at baseline assessment (Lohr et al., 1996; Newman et al., 1998). These scores do not provide qualitative meaning related to patient experience as they are based merely on the statistical properties of the score.

Anchor-based methods are used to assign participants into distinct grades that *do* have qualitative meaning related to patient experience (Hays & Woolley, 2000; Crosby et al., 2003; Yost & Eton, 2005; Revicki et al., 2008). The criterion anchor-based approach involves mapping established cut-off criteria from an established questionnaire to the new questionnaire. Essentially this method would stratify individuals from their TFI score using their corresponding scores on an established tinnitus questionnaire, such as the THI. Descriptive statistics would then be calculated for each grading and differences in scores between grades would be compared using one-way ANOVA to establish whether the grades were distinct. To address the patient experience, responses to a global rating of perceived problem can be used to stratify participants and their questionnaire scores into distinct grades. Descriptive statistics and ANOVA can be calculated to establish the distinctive score range within each grade. The grading scheme can then be assigned descriptors that correspond to those used in the global question, e.g. ‘small problem’ to ‘very big problem’ (see Chapter 4 for further details).

ROC curve analysis was used to identify the scores that best discriminate between the participants in each adjacent category (Riddle et al., 1998; Kaye & Darke, 2002; Eng, 2005; Copay et al., 2007; Uslu et al., 2008; Ward et al., 2014).

In the case of anchor-based methods, the TFI would function as the diagnostic tool and the anchor would be the gold standard. ROC curve analysis synthesises this information into sensitivity and specificity for detecting the

difference in categories, i.e. ‘no problem’ versus ‘mild’ tinnitus, or ‘mild’ versus ‘moderate’ tinnitus (Riddle et al., 1998; Eng, 2005). Sensitivity is equivalent to the probability that participants are correctly classified according to their TFI score as experiencing “mild” tinnitus (positive cases), whilst specificity refers to the probability that participants are correctly classified as not experiencing tinnitus, i.e. “no problem” (negative cases) (Eng, 2005; Uslu et al., 2008). AUC scores should be above 0.5 and values of more than 0.7 are desirable for establishing independent grades (Eng, 2005; Zou et al., 2007). ROC curve analysis provides a range of scores in which an optimal threshold (cut-off) was identified. In this case, rather than a balance between sensitivity and specificity, sensitivity is prioritised above specificity for the optimal threshold since it is more important as a diagnostic tool to identify the greater tinnitus symptomatology. The threshold provides the cut-off value for the range in each diagnostic category. For each set of adjacent diagnostic categories separate ROC curves were calculated (Chapter 4).

2.2.4.2 Interpretation of change in scores

It is important to not only identify the questionnaires ability to detect small changes, but to also provide clinically meaningful interpretations to those changes in scores (Jaeschke et al., 1989; de Vet et al., 2006b; Terwee et al., 2009; Revicki et al., 2008; de Vet et al., 2011; Wyrwich et al., 2013). For example, a detectable change in TFI score is possible without any corresponding change in the patient experience of their tinnitus, the patient may not feel better. To address this, the concept of the minimal clinically important difference (MCID) was proposed by Jaeschke et al. (1989) and defined as “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management.” (p. 409). It is the

process of identifying the smallest change in scores that is considered important from the perspective of a patient. MCID is also referred to as minimal important difference (Revicki et al., 2008), minimal clinically important improvement (Ward et al., 2014) or minimal important change (MIC, Terwee et al., 2007). For the sake of clarity, the former term used by Terwee et al. (2007) is adopted throughout the thesis.

The most widely used approach to identify what is a meaningful change is to classify patients according to the amount of change experienced using their responses to a global rating of health-status change. The mean change in scores (difference) on the TFI between baseline and the follow-up points are calculated corresponding to each rating change category. It is recommended to triangulate the results of multiple methods for determining MIC, including relating the distribution-based methods identified in the previous sections (particularly the SEM and smallest detectable change). This can help to converge on a small range of values and include an assessment of measurement error (Crosby et al., 2004; Revicki et al., 2008).

The basis of the first method comes from the methods used by (Jaeschke et al., 1989) where the MIC was identified as the mean change scores (follow-up – baseline) for the pooled 'almost the same', 'a little' and 'somewhat' improved and 'deteriorated' groups. However, there are some limitations with this approach. The "almost the same" group is fundamentally indicating "unchanged" health-status (Juniper et al., 1994; Beaton et al., 2002; Revicki et al., 2008; Wyrwich et al., 2013). There was an explicit assumption that the mean change scores represent the same change for improvement as deterioration which might not the case. These concepts are fundamentally measuring different aspects of change which may not be equally important. It has been shown that larger change is needed to feel worse than better (Cella et al., 2002).

To identify the MIC values, the mean change scores for each category were plotted and then the difference score between the mean change scores for "no change" and "slightly improved" groups and for the "no change" and "slight worse" groups at each time interval were calculated (Chapter 4) (Redelmeier et al., 1996; Terwee et al., 2009; Meikle et al., 2012). The difference scores and MIC values for the 'much-to-moderately' improvements/worsening are also calculated and reported.

ROC curve analyses were used to identify the TFI global scores that best identify improvements (i.e. slight, much to moderate) and worsening (slight, much to moderate) from no change based on global rating of change categories (section 2.2.3.2 for detail on ROC). The AUC is expected to be reasonably high, above 0.7 to successfully discriminant change. In this case, the balance between sensitivity and specificity (optimal threshold) is employed as the value for identifying the MIC score (Zou et al., 2007).

It is recognised that the MIC scores are influenced by the magnitude of the baseline values (Crosby et al., 2004; Ward et al., 2014). To account for this, the MIC values were calculated for each diagnostic category/subgroup (i.e. the "small problem" baseline score range) corresponding to the ratings of change category, such as improved (section 2.2.4.1 for grading). Optimal thresholds were calculated for each diagnostic category.

The MIC and optimal threshold estimates identified were triangulated with the distribution-based methods, SEM and smallest detectable change to identify a single important change score or at narrow range of values that has both external validity and accounts for the variability. To identify the most appropriate important change score, it is recommended to visually examine the score distributions and prioritise the estimates identified using the anchor-based techniques (Terwee et al.,

2007; de Vet et al., 2007; Revicki et al., 2008). For direct comparisons between all methods and change groups, estimates are visually depicted using a unique technique identified by de Vet et al. (2007), called the visual anchor-based MIC distribution (Figure 2.2).

Initially, the distributions of the change scores on the TFI for the improved, unchanged and worsened groups are plotted. To ease comparisons across the distributions, proportional frequencies were used for the scores to make the curves independent of sample size. The anchor-based and distribution-based estimates are plotted with these distributions to clearly illustrate the extent of the differences (distance) between these estimates. The form of the curve indicates the extent of the relationship between the global change anchor and questionnaire, a flatter curve equates to weak relationship and indicates that there might be some misclassification. The preferred important change cut-off point should desirably account for both patient experience and measurement error, but priority can be placed on the MIC. Either way the choice should be justified (de Vet et al., 2007, 2011).

2.3. SUMMARY

A summary of the methods and criteria that are used in this thesis are provided in Tables 2.2 – 2.5. The essential criteria are informed by the criteria identified by Terwee and colleagues (2007, 2014) and were used in the critical review described in Chapter 3. The essential and additional criteria were both applied as standards for statistical analysis in Chapters 4 and 5.

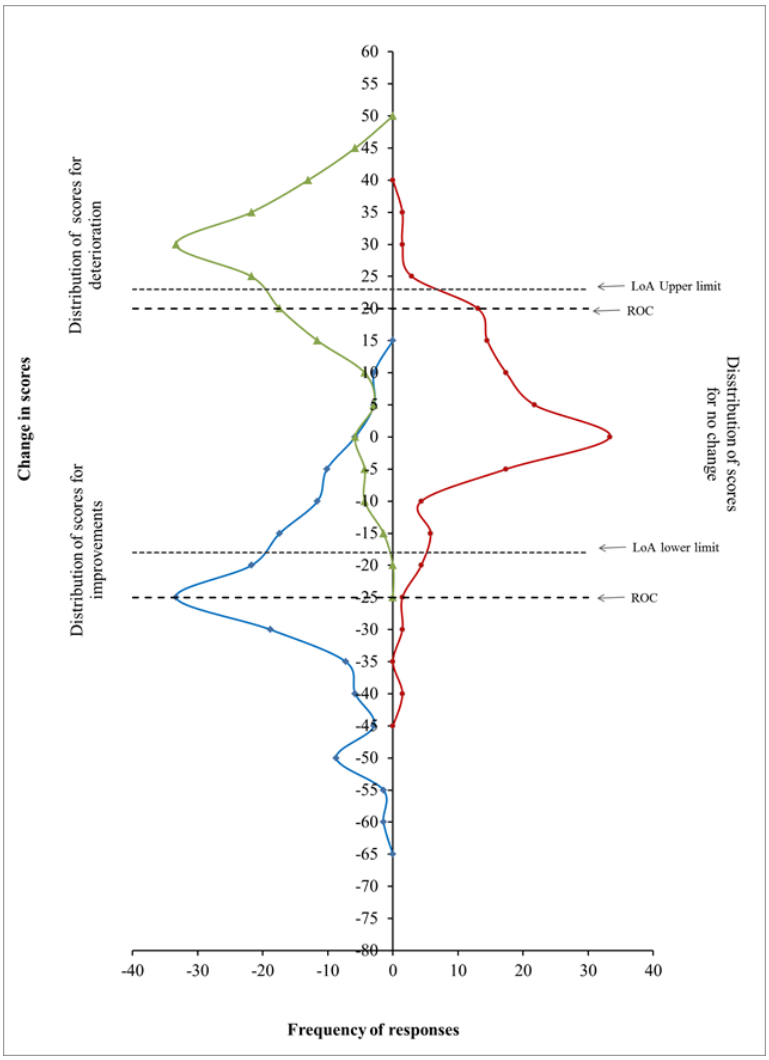

Figure 2.2. Example of visual-anchor-based MIC distribution.
Distributions (expressed in percents) of the changes in scores for patients who reported improvements (blue), no change (red) and worsening (green). Horizontal dashed lines represent optimal value from Receiver Operating Characteristic (ROC) analysis. Horizontal dotted lines represent Limits of Agreement estimates.

CHAPTER 3. PSYCHOMETRIC PROPERTIES OF FIVE STANDARD TINNITUS QUESTIONNAIRES: A REVIEW

3.1. INTRODUCTION

There are numerous questionnaires in existence which reputedly measure the range of psychosocial effects and aspects of tinnitus that affect everyday life (Newman et al., 2014). This Chapter reviews five commonly used questionnaires from the perspective of their development and psychometric evaluation. The questionnaires are: Tinnitus Questionnaire (TQ; (Hallam et al., 1988; Hiller & Goebel, 1992)), Tinnitus Handicap Questionnaire (THQ; Kuk et al., 1990), Tinnitus Reaction Questionnaire (TRQ; Wilson et al., 1991), Tinnitus Handicap Inventory (THI; Newman et al., 1996), and the Tinnitus Functional Index (TFI; Meikle et al., 2012). These have been identified in practice guidelines and expert opinions (Department of Health, 2009; Langguth et al., 2011; Landgrebe et al., 2012; Tunkel et al., 2014).

The quality criteria relating to the psychometric properties are those identified in Chapter 2 (see Tables 2.2 – 2.5). Table 3.1 provides an overview of the characteristics of the five questionnaires.

During development of the TFI, reliability and validity testing (i.e. factor analysis, effect size and convergent/discriminant validity) were conducted on prototypes 1 and 2. Evidence for the validation of the 25-item TFI comes from a post-hoc analysis of the 30-item prototype 2. In this chapter, the evidence provided is based on this re-analysis and from subsequent evaluations conducted in the USA (Henry et al., 2015) and Belgium (Rabau et al., 2014).

.

Table 3.1. Tinnitus questionnaires characteristics

Questionnaire	No of items Response options Total range	Subscales	Translations	Original development Internal consistency	n	Population
Tinnitus Questionnaire[a]	52 items True, partly true, not true 0 – 82	1. Emotional and cognitive distress 2. Intrusiveness 3. Auditory Perceptual difficulties 4. Sleep disturbance 5. Somatic complaints	German, Dutch/French, Chinese	0.94 0.87 0.89 0.81 0.75	100	Neuro-otology outpatients (UK) Royal National Throat Nose and Ear Hospital Clinic (London)
Tinnitus Handicap Questionnaire[b]	27 items 0 – 100 0 – 100	1. Physical, emotional & social effects 2. Hearing and communication ability 3. Individual perception of tinnitus	Dutch, French	0.94 0.88 0.47	275	Iowa University Hospital Iowa VA Medical Centre The House Ear Institute
Tinnitus Reaction Questionnaire[c]	26 items Not at all (0) to almost all the time (4) 0 – 104	No clear subscales	French	N/A	156	VA Hospital General Audiology Research
Tinnitus Handicap Inventory[d]	25 items Yes(4) Sometimes(2) No(0) 0 – 100)	1. Functional 2. Emotional 3. Catastrophic	Hungarian, Chinese-Cantonese, Chinese-Mandarin, Italian, German, Danish, Turkish, Japanese	0.86 0.87 0.68	150 196*	Henry Ford Hospital (USA) VA medical Centre (USA) Audiology Clinical/ Surgery (UK)
Tinnitus Functional Index[e]	25 items 0 – 10 0 – 100	1. Intrusiveness 2. Sense of Control 3. Cognition 4. Sleep 5. Auditory 6. Relaxation 7. Quality of Life 8. Emotional	Dutch	0.85 0.82 0.96 0.97 0.97 0.96 0.93 0.94	336	Two VA Medical Centres Hearing and Speech Institute University Tinnitus Clinic Tinnitus Management Clinic

a: Hallam et al., 1988; b: Kuk et al., 1990; c: Wilson et al., 1991; d: Newman et al. 1996; e: Meikle et al., 2012. * Sample size from factor analysis study (Baguley & Andersson, 2003)

3.2. VALIDITY OF THE FIVE QUESTIONNAIRES

3.2.1. Content validity

Aspects of content validity considered (i) measurement aims, (ii) target population, (iii) concept, (iv) item selection and reduction, and (v) item interpretability (Terwee et al., 2007).

3.2.1.1 Measurement aims

The TQ and THI aim to be diagnostic measures of tinnitus severity (Hallam et al., 1988; Newman et al., 1996). Nevertheless, both are used internationally as outcome measures for the effectiveness of therapeutic interventions. They were not designed for this purpose. Response options and items were not selected to be specifically responsive to treatment-related changes. These questionnaires consequently lack the resolution to detect small changes in scores (Lipsey, 1983; Kirshner & Guyatt, 1985; Meikle et al., 2008; Lipsey & Hurley, 2008).

The THQ was developed to be a diagnostic measure of handicap due to tinnitus, but the need for effectively evaluating treatment-related change was also considered. The THQ has high resolution response options (0-100), allowing for small changes in scores to be accounted (Kuk et al., 1990).

The TRQ was developed to measure psychological distress in relation to tinnitus and the changes that occur before and after treatment (Wilson et al., 1991). However, it is not obvious how these aims influenced the development of the scale, for example a 5-point scale was chosen even though it potentially does not have a sufficient resolution to detect small changes.

The TFI was developed for measuring the functional impact of tinnitus on daily life, discriminating people who are bothered by their tinnitus from those who are not,

and for measuring the changes over time or after clinical intervention (Meikle et al., 2012). Both of the discriminative and evaluative components were considered throughout the development.

3.2.1.2 Target population

All five questionnaires clearly define the population of interest as people experiencing tinnitus and intended to measure the construct of tinnitus handicap/distress. The majority were originally developed with patients presenting at audiology clinics within hospitals (Table 3.1). However, item development and evaluations of the THQ, THI, TRQ and TFI, used data collected from Veterans' Affairs (VA) hospitals and so are not representative of the wider clinical population. Those recruited from the VA sites tended to be male and experienced a range of co-morbidities, such as Post-Traumatic Stress Disorder (PTSD) that could influence responses to the TFI items. Although the TQ development did not include VA hospital patients, it important to note that there is very little published information available on its validation and development. Apart from Hallam et al. (1988), the only information available is in the 2008 TQ manual, which provides limited information (Hallam et al., 1988; Hallam, 1996, 2008). Only the TRQ developers recruited from the general public.

Questionnaires have been translated into different languages, including some psychometric evaluation in the target country. Notably, the majority of the reliability and validity assessments for the TQ have been conducted on the German translation in clinical populations (GHTQ; (Hiller & Goebel, 1992; Goebel & Hiller, 1998).

3.2.1.3 Concept

All questionnaires were based on the clinical experience of the developers or previous qualitative patient data identifying potential domains of tinnitus (e.g. Tyler & Baker, 1983; Jakes et al., 1985). No direct evidence from patient perspective was included.

Tinnitus Questionnaire

The work conducted by Jakes et al. (1985) investigating tinnitus complaints and tinnitus loudness provided the theoretical basis for the TQ. In particular, the general tinnitus complaint dimensions of emotional distress and intrusiveness, and the specific tinnitus complaint dimensions of sleep disturbance, medication use and interference with auditory entertainment identified informed the development, although it is not apparent how these dimensions informed subsequent item selection.

Tinnitus Handicap Questionnaire

Kuk et al. (1990) used the four patient-reported domains of hearing, lifestyle, health and emotion identified by Tyler and Baker (1983). They also included one domain reflecting "others' reaction to tinnitus". It is unclear how this additional domain was derived. No other information was provided.

Tinnitus Reaction Questionnaire

Based on evidence suggesting links between tinnitus and stress, depression and anxiety, Wilson et al. (1991) identified psychological distress associated with tinnitus as an individual component that could specifically be relevant for evaluating psychological interventions.

Tinnitus Handicap Inventory

During item selection, Newman et al. (1996) proposed a three-domain model reflecting the different aspects of tinnitus impact on everyday function; (i) mental, social and physical functioning, (ii) emotional response to tinnitus, and (iii) the desperation associated with tinnitus (catastrophic). It appears that it is largely based on the developers' personal experience and opinion.

Tinnitus Functional Index

Only the TFI used expert opinions (tinnitus researchers and clinicians), beyond those of the developers, to create a conceptual framework that was relevant to the target population and comprehensively covered the construct. Patient perspective, however, was not considered. The TFI was intended to be broad in scope comprehensively covering multiple domains that impact on daily functioning.

Thirteen domains were initially proposed by the 17 judges and four developers; (1) Emotional distress, (2) Social distress, (3) Unpleasantness, intrusiveness of tinnitus, (4) Persistence of tinnitus, (5) Interference with work activities, (6) Interference with leisure activities, (7) Disturbance of sleep and rest, (8) Disturbance of relaxation, (9) Auditory perceptual difficulties, (10) Somatic and physical complaints, (11) Cognitive interference, (12) Reduced Quality of Life, (13) Reduced sense of control.

3.2.1.4 Item selection and reduction

Tinnitus Questionnaire

Item selection is only referenced once. The 52 items were based on adverse effects and complaints reported by tinnitus patients (Hallam et al., 1988; Hallam, 2008). There is no reference to item reduction, even though some items could be considered

redundant or of limited clinical relevance. For example, 11 of the 52 items are only used as baseline information and do not contribute to the total and subscale scores (Newman & Sandridge, 2004; Hiller & Goebel, 2004; Hallam, 2008).

Tinnitus Handicap Questionnaire

Kuk et al. (1990) generated 87 items based on the difficulties identified by Tyler & Baker (1983). The 87 items were reduced to the final 27-items through examination of response frequency distributions, inter-item correlations, and item-total correlations. Items were eliminated if they showed high inter-item correlations, or they showed low item-total correlations. The specific criterion cut-off values were not reported. In addition, some items were eliminated if they scored either 0 or 100 in over 50% of the cases, but this criterion was applied inconsistently.

Tinnitus Reaction Questionnaire

Wilson et al. (1991) briefly refer to the domains reported by Tyler & Baker (1983) as the principal foundation for item selection, supplemented with experienced gained from patient interviews (i.e. Ireland et al., 1985). The developers did not appear to conduct any item reduction.

Tinnitus Handicap Inventory

It appears the original 45-items were derived from "case histories of patients with tinnitus", from Tyler and Baker (1983), and adapted from the Hearing Handicap Inventory for Elderly (HHIE, (Ventry & Weinstein, 1982)) and the Dizziness Handicap Inventory (DHI, (Jacobson & Newman, 1990)). No details of the procedure for item development were provided. The 45 items were reduced to the final 25-item version using response frequency distributions, item-total correlations, and content validity. Items were eliminated if high endorsement rates were found

(i.e. 85% of patients selected the same response), or if item-total correlations were ≤0.50. There is little clarity on the items removed and on the decision behind the frequency distribution and item-total elimination criteria.

Tinnitus Functional Index

Meikle et al. (2012) reports that 175 items were sourced from nine widely used tinnitus questionnaires. Items that were ambiguous (i.e. referring to multiple subtopics), overly negative or duplications were excluded. On the consensus of 17 experts, items were assigned to the initial 13 domains and rated for expected responsiveness. This formed the selection criterion for prototype 1.

Of 175 items, 35 were judged most relevant. Eight items were added in order to maintain a minimum of three items on each domain making 43 items in prototype 1, with a 0 – 10 response scale. Thirteen items were eliminated after examination of response frequency distributions, effect size and Principal Components Analysis (PCA) making 30 items for prototype 2. The same examinations were conducted on prototype 2, but none of these findings contributed to the removal of the final five items. This decision was based on "careful examination" of the items (content validity), with no further information provided.

3.2.1.5 Item interpretability

None of the questionnaires appear to use jargon terms or double-barrelled items but none were assessed for readability at the time of conception/development. Atcherson et al. (2011) subsequently examined readability of the THI, THQ, TRQ and TQ in the United States. It was not apparent which version of the TQ was investigated. The THI and THQ requires a reading ability equivalent to 13–14 year olds, while the TRQ requires a reading ability equivalent to 11–12 years, within the recommended

criteria (Terwee et al., 2007). There is no estimate of readability for the TFI. Only the TRQ and TFI included a time of reference in which patients are instructed to recall their problems over the past week. The remaining questionnaires leave the decision on the time of reference to the clinician, researcher or patient.

3.2.2. Construct validity

Evidence of convergent and discriminant validity were examined, in particular with respect to any predefined hypotheses about the expected correlations.

3.2.2.1 Convergent validity

All five questionnaires have shown high convergent validity with each other. However, for the TQ, THQ and TRQ, not all the predictions about the strength of these correlations were hypothesised *a priori* (Wilson & Henry, 1998; Baguley et al., 2000)or they were too conservative in their predictions (Robinson et al., 2003) (Table 3.2). These findings do not meet the criterion established by Terwee et al. (2007) and so cannot be considered as good evidence of convergent validity. In contrast, the THI and TFI show good evidence of convergent validity. The THI showed an expected strong correlation with the THQ (r = 0.78) (Newman et al., 1996), and the TFI showed an expected extremely strong correlation with the THI (r = 0.86).

3.2.2.2 Discriminant validity

The TQ, THQ, TRQ and THI have all shown acceptable discriminant validity with a range of measures. Robinson et al. (2003) appropriately predicted that these questionnaires would demonstrate moderate correlations with standardised measures of depressive symptoms, well-being, and internal attention and focus, see Table 3.2.

Table 3.2. Pearson's correlation coefficients for the four tinnitus questionnaires and other general health measures as reported by Robinson et al. (2003).

	TQ	THQ	TRQ	THI	HRSD	BDI	QWBS	MSPQ
TQ								
THQ	0.75							
TRQ	0.82	0.79						
THI	0.89	0.77	0.88					
HRSD	0.48	0.57	0.52	0.49				
BDI	0.51	0.62	0.66	0.59	0.71			
QWBS	-0.37	- 0.48	- 0.39	- 0.37	- 0.49	- 0.59		
MSPQ	0.46	0.37	0.53	0.54	0.58	0.58	- 0.39	

TQ = Tinnitus Questionnaire (Hallam et al., 1988); THQ = Tinnitus Handicap Questionnaire (Kuk et al., 1990); TRQ = Tinnitus Reaction Questionnaire (Wilson et al., 1991); THI = Tinnitus Handicap Inventory (Newman et al., 1996); HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960); BDI = Beck's Depression Inventory (Beck et al., 1997); QWRS = Quality of Well-Being scale (Kaplan et al., 1996); MSPQ = Modified Somatic Perception Questionnaire (Main, 1983).

The only exception was the THI, THQ and TRQ which showed moderately strong correlations with the BDI ($r = 0.59$, $r = 0.62$ and $r = 0.66$, respectively).

Other evidence suggests that the TRQ is sensitive to generalised emotional distress (Wilson et al., 1991; Andersson et al., 2003), but as no prior assumptions were made the information for the TRQ is defined as indeterminate. Recent evidence has shown that there is an overlap in item content between the THI and the BDI, specifically 9 out of the 25 items on the THI are significantly associated with large beta values in the regression model with the BDI reflecting the similar properties of depressive symptoms such as concentration, sleep and emotional problems (Ooms et al., 2011; Zeman et al., 2014). Although neither study provided *a priori* assumptions about the strength of relationships, Ooms et al. (2011) did anticipate which items would overlap and that the somatic subscale in BDI would most likely be associated

with the THI. Therefore, given the consistency of this evidence, the THI is defined as having indeterminate discriminant validity with measures of depressive symptoms.

The TFI 25-item prototype 2 was, as predicted, moderately associated with the BDI-primary care with distinctly lower correlations than those with the THI (r = 0.56) (Meikle et al., 2012). This is modest evidence of discriminant validity.

3.2.3. Structural validity

Structural validity was generally examined using the data reduction technique of Principal Components Analysis (PCA). The TFI used a factor analysis estimation method known as Principal Axis Factoring. CFA was not conducted on any of the questionnaires.

Tinnitus Questionnaire

Apart from Hallam et al. (1988), the only evidence available is in the 2008 TQ manual (Hallam, 2008). The 1996 TQ manual (Hallam, 1996) where the revisions of TQ internal structure are reported, is out of print.

Hallam et al. (1988) first reported a three-factor solution (based on 34 of the 51 items); i) emotional distress, (ii) auditory difficulties and (iii) sleep disturbance, which together explained 80% of the variance. In 1996, Hallam revisited the factor structure. PCA with orthogonal rotations revealed a six-factor solution explaining 55% of the variance, but only five factors were considered reliable (Table 3.1). The information provided on the 1996 revalidation, in particular the factor loadings, is at times contradictory and vague. Moreover, it is important to highlight that this PCA analysis was based on all 52 TQ items, not on the 41 items that contribute to calculate the total score. This could have a major impact on the resulting five-factor structure.

Tinnitus Handicap Questionnaire

Following exploratory factor analysis using PCA, Kuk et al. (1990) identified three factors for the 27-item THQ (Table 3.1). Altogether these explained 57.6% of the variance; factor 1 (15 items: social, emotional and physical functioning) explained the majority of the variance (42.6%), whilst factor 2 (8 items: auditory and social functioning) and factor 3 (4 items; Individual perception of tinnitus) only explained 9.4%, and 5.6% of the variance, respectively. The structural validity is questionable. Closer inspection of the factor loadings indicated that the majority of items loaded onto factor 1 and that there was unreported cross-loading between items on factor 1 with items on factor 2 and factor 3. For five items there was also a large amount of unexplained variance (communalities > 0.5), including three of the items on factor 3 which also resulted in extremely low item-total correlation scores (r = 0.15). Instead of removing these items, the developers chose to maintain them even though they would undermine the global score.

Tinnitus Reaction Questionnaire

PCA and orthogonal rotation revealed a four-factor solution explaining 66.4% of the variance (Wilson et al., 1991; Table 3.1); factor 1: General distress (50%), factor 2: Interference (7.9%), factor 3: Severity (4.6%) and factor 4: Avoidance (3.9%). Again, the structural validity is questionable. The majority of items heavily loaded onto factor 1 and many others cross-load onto factors 2 and 3. Wilson et al. (1991) also inspected a two-factor solution, but not all the statistical findings are reported. Evidence for variance explained, potential cross-loading or redundant items cannot be reviewed. Although Wilson et al. (1991) recognised that the full scale was homogeneous, they still recommended use of the scores in the two factor solution.

Tinnitus Handicap Inventory

Newman et al. (1996) predefined three domains; (i) functioning (12 items), (ii) emotional (8 items), and (iii) catastrophic (5 items). However, the structure was not subjected to statistical analysis at the time. In 2003, Baguley and Andersson investigated the three-factor structure using PCA with oblique rotations. The item-total correlations were high (r = 0.60) indicating the THI is tapping into a narrow construct. Baguley and Andersson (2003) observed that the structural validity is questionable. For example, most items loaded onto the Factor 1 (19 items), possibly indicating a unidimensional structure. Factors 2 and 3 each only consisted of 5 items with 4 items from factor 3 cross-loading onto factor 1. Altogether these factors explain 52.8% of variance.

So far, confirmatory factor analysis has only been conducted on the German translation of the THI, where the original 3-factor structure (with error covariance between two pairs of items) was found to have reasonably good fit, and better than the fit for a unidimensional structure (Kleinstäuber et al., 2015).

Tinnitus Functional Index

Meikle et al. (2012) used a thorough approach to identify the factor structure of the TFI and the items that should be maintained. PCA followed by Principal Axis Factoring (assessing both orthogonal and oblique rotations) revealed an eight-factor solution for prototype 1 which was replicated for the 30-item prototype 2 (79.5% of variance explained). The 25-item prototype 2 data was not subjected to confirmatory factor analysis; but Principal Axis Factoring with oblique rotations. Unsurprisingly, the same eight factors were clearly identified for both all the participants and a subset of only those that reported problems with their tinnitus (79.5% variance explained;

Table 3.2). Three items (Q2, Q21, Q22) had low loading estimates with their designated factor, less than the recommended value of 0.4. Two of which, from the QoL factor, also appeared to load on another factor (Cognitive). This does call into question the usefulness of these items as not only do they contribute less to the overall score but also are potentially confounding the results for the subscale. The remaining 22 items loaded onto their designated factor, demonstrated by the high loading values. However, it should be noted that all the items associated with the Sleep factor had negative loading values indicating that these items might not be directly related to the overall construct. This is not apparent in the results of the intercorrelations between the factors, in which Sleep moderately correlated with the other factors. The Auditory factor, on the other hand, had noticeably weaker correlations with all the other factors, indicating that it is potentially measuring a different construct. Meikle et al. (2012) recognised this and proposed that the TFI has two possible structures, either "a general tinnitus severity factor underlying all eight subscales...[or] a general tinnitus severity factor underlying seven of the eight subscales, with the Auditory subscale representing an underlying specific factor" (p.20).

3.3. RELIABILITY OF THE FIVE QUESTIONNAIRES

This was investigated by examining internal consistency, and reliability and agreement.

3.3.1. Internal consistency

Extremely high Cronbach's alpha estimates were observed for the TRQ ($\alpha = 0.96$; (Wilson et al., 1991)) and the TFI total scores ($\alpha = 0.97$) Table 3.1 (Meikle et al., 2012)). These extremely high scores exceed the criteria for good internal consistency

($\alpha < 0.95$) and indicate potential redundant items (Schmitt, 1996; Streiner, 2003; Tavakol & Dennick, 2011). Alpha estimates can be inflated by a multidimensional structure, which could be the case for the TFI (Shevlin et al., 2000; Kottner & Streiner, 2010). The alpha estimates for the TQ, THI and THQ total scores are consistently all just within the criterion ($\alpha = 0.95$, $\alpha = 0.93$, and $\alpha = 0.94$, respectively) (Kuk et al., 1990; Newman et al., 1996; Hallam, 2008). This suggests that items in each questionnaire are reliably measuring the same underlying construct. Consistent with this were the item-total correlations which indicated that in general the majority of the questions in each questionnaire related to the overall construct, except for the items in the 'Individual perception of tinnitus' subscale of the THQ (factor 3), which again indicated poor consistency. High values could be a reflection of the complex nature of tinnitus handicap and inter-relations between the domains.

In terms of the questionnaire subscales, all except two show good to acceptable Cronbach's alpha estimates and fall within the acceptable criterion range (Table 3.1). Two subscales dropped below the criterion ($\alpha > .70$) for poor internal consistency. The THQ's 'Individual perception of tinnitus' subscale (Factor 3: 4 items) had an unacceptably low alpha value ($\alpha = 0.47$), whilst the THI's catastrophic subscale (5 items) had a questionable alpha value ($\alpha = 0.68$). Low scores could be due to the small number of items in the subscales, but given the other evidence against the validity of these subscales, it is more likely that this indicates a heterogeneous construct with poor inter-relatedness between the items (Tavakol & Dennick, 2011). If this is the case, the subscales would not be adding value to the overall construct of tinnitus handicap and could in fact be redundant (Cortina, 1993; Tavakol & Dennick, 2011).

3.3.2. Reliability and agreement

Tinnitus Questionnaire

Evidence of reliability is only available for the German translation of the TQ (GHTQ; (Hiller & Goebel, 1992; Goebel & Hiller, 1998)). Extremely high test-retest reliability was observed for the total (r = .94) and subscale scores (r =.86 to r = .93) for a 3 day interval (n: 60; Hiller et al., 1994). Although these scores provide good evidence of the GHTQ to reliably distinguish people with tinnitus from each other, they do not necessarily reflect the reliability of the original English TQ. There is no apparent evidence on agreement.

Tinnitus Handicap Questionnaire

Extremely high test-retest reliability has been observed for the THQ total score (Pearson's r = 0.89), and two of the subscale scores (functioning: r = 0.89 and hearing: r = 0.90) with 30 to 40 day interval between tests (Newman et al., 1995). The variability in these scores is stable across time. Changes can therefore be attributed to treatment-related effects rather than measurement error. In contrast, the 'individual perception of tinnitus' subscale only showed moderate correlations with the other subscales (r = 0.50). This would provide additional evidence against the use of this subscale.

Newman et al. (1995) calculated the limits of agreement by multiplying the standard deviation of the mean change in scores of repeated measures by ±2 standard deviations of the change scores. The levels of agreement for the THQ were ±20.0 for the total score, ±23.3 for factor 1, ±25.8 for factor 2 and ±44.8 for factor 3. Newman et al. (1995) found that for only two participants, out of a total of 32, the difference in test-retest scores were not within the agreement values for the THQ total and

factor 1 and 2 scores. Factor 3 showed a large amount of variability and did not show the same agreement (three outliers). Therefore, it would appear the THQ (total and factors 1 and 2) showed agreement for 95% of the observed differences. Consequently, this indicates acceptable agreement between the individual scores, the mean difference in scores was not included in the limits of agreement which could impact on the percentage agreement observed, and therefore there is some uncertainty in the evidence for agreement. The SEM consistency using reliability estimates was reported for each factor and the total scores, ranging from 7.4 to 15.5, with the largest error observed for factor 3. However, the authors did not provide the information on which reliability estimates were used to calculate the SEM.

Tinnitus Reaction Questionnaire

Wilson et al. (1991) provide evidence of extremely high test-retest reliability (r = 0.88) over 3 days to 3 week time interval (n: 43 research population only), suggesting that despite measurement error the TRQ scores can be used to reliably discriminants between individuals. Although it should be noted that the sample size was below the recommended size ($n \geq 50$) and the estimates may not be representative of a larger clinical population. There is no apparent evidence on test-retest agreement or the SEM.

Tinnitus Handicap Inventory

Newman et al. (1998) observed high test-retest reliability for the THI total score (r = 0.92) and subscales scores (r = 0.84 to r = 0.94) with a 3 week interval between tests. The THI shows stability over short time intervals.

To assess agreement, Newman et al. (1998) calculated the limits of agreement for total and subscales. The levels of agreement were ±19.5 for the THI total score

and ±8.1, ±9.8 and ±5.8 for the functional, emotional and catastrophic subscale scores, respectively. Newman et al. (1995) demonstrated that all participants, except one, test-retest difference scores were within the agreement values for the THQ total and subscale scores. It would appear the THI showed agreement for 95% of the observed differences. The total scores were not shown in a plot so could not be examined and similar to the THQ, the mean difference scores were not included in the limits of agreement. However, the SEM consistency was reported and indicated a mean score is unlikely to deviate more than 7.0 points between tests (Newman et al., 1998). But again there was no clarification on which reliability estimates were used in the calculation. Therefore, some of the evidence for the agreement was not reported, creating uncertainty. Additionally, the sample size for both reliability and agreement (n=29) was below the recommended size.

Tinnitus Functional Index

Meikle et al. (2012) observed excellent test-retest reliability for the 25-item subset prototype 2 score (r = 0.78) with a 7 to 30 day time interval. The eight subscales had moderate to excellent reliability, ranging from r = 0.63 (QoL) to r = 0.90 (Auditory). This would suggest the TFI scores differentiate between individuals and shows stability over short time intervals. However, the sample size (n=37) was lower than recommended. Meikle et al. (2012) did not report which correlation coefficient was conducted and there were no measures of test-retest agreement.

3.4. RESPONSIVENESS OF THE FIVE QUESTIONNAIRES

Only the THQ and TFI response scales used interval measurement units. These are more sensitive to detecting changes than the categorical units of the TQ, THI and TRQ. The THQ's 0 – 100 item response scale has however been criticised for being

too large a range and unwieldy for patients to complete (Newman & Sandridge, 2004).

3.4.1. Floor or ceiling effects

No information was available on response distributions for the TQ. Response frequency distributions were investigated during development of the THQ, TRQ, THI and TFI. For the THQ and THI this was only used to detect which items to remove in the initial development and the analysis was not repeated after developing the questionnaire.

Tinnitus Reaction Questionnaire

Wilson et al. (1991) investigated the percentage of responses in the five response options for the individual items. Twenty-two items were identified as having floor and ceiling effects. Nine items were reported to have extreme floor effects with 50% - 83% of the participants selecting the "not at all" category. Thirteen items reported had more than 20% of participants endorsing the two highest options (a good deal of time, almost all of the time). The authors attributed these effects to reflecting the participant's feelings. However, the ceiling effects could limit the detection of worsening tinnitus. More concerning are the extreme floor effects which could severely limit the detection of individual improvements in tinnitus, and as consequence limit the ability of the TRQ to detect change and be a reliable measure of outcome.

Tinnitus Functional Index

Meikle et al. (2012) found that four of the 25 items (1, 4, 6, 18) showed "mild ceiling effects" (p.16). These items were endorsed for the most severe response of 10 by 25 to 34% of participants. Meikle et al. (2012) were not concerned with these "mild

ceiling effects" since the main aim of the measure is to evaluate improvements after treatments (p.16). In terms of the criteria established by Terwee et al. (2007), this is hardly a mild ceiling effect. In fact, following Terwee et al.'s criterion (2007), a further five items (3, 10, 11, 12, 24) showed ceiling effects with 17-18% of participants selecting the severe response (10), and four items (19, 21, 22, 25) showed floor effects with 16-27% of participants selecting the least severe response (0). What is concerning is that these effects are limiting both the detection of worsening tinnitus and improvement of tinnitus, therefore reducing the chances of the TFI being responsive to treatment-related changes. Three of the four items on the QoL subscale showed floor effects. In fact, one issue to consider in this analysis is the population that completed prototype 2. The majority of the participants from the two tinnitus clinics experienced more severe tinnitus, with relatively few reporting mild tinnitus, which could have contributed to the high levels of ceiling effects and therefore the results may not be representative of the wider clinical population. This needs re-investigation.

3.4.2. The ability to detect changes in scores

The GHTQ, THI, THQ and TRQ have been used to demonstrate treatment effects across a wide range of interventions and as a consequence all might be assumed to have the ability to detect changes over time and between different subgroups (Hoare et al., 2011; Henry et al., 2015).

Unfortunately, the concept of responsiveness is still evolving, and the majority of the methods now applied were not routinely included in the development of questionnaires in the late 80s to early 90s. There is no gold standard for tinnitus. Only the GHTQ, THI and TFI scores have been examined using a global ratings of perceived change in tinnitus to identify the ability to detect change and a minimal

important change score. Only the THQ and THI report evidence for the degree of variability over time and the amount by which scores would have to change to be considered true changes in scores above that variability and measurement error (3.3.2).

Tinnitus Questionnaire

For no versions of the TQ (Hallam et al., 1988; Hallam, 1996, 2008), has any analysis been conducted to compare changes in the TQ with global ratings of perceived change (Newman & Sandridge, 2004). However, Adamchic et al. (2012a) have provided evidence using the GHTQ and Clinical Global Impression Improvement (CGI-I) ratings (Table 3.3). Moderate correlations were reported between the GHTQ scores and CGI-I scores (r = 0.52), consistent with the expectation that the correlation would be more than 0.3. This indicates that the GHTQ scores are reflecting the changes that are perceived by patients. The SEM indicated minimal measurement error (-4.7 points) and the mean change scores and ES (Cohen's d) for each rating group followed an expected pattern, in which larger difference scores and ES were observed for the minimally better group (Mean: -6.7; d = 0.41) than the no change group (Mean -0.33; d = 0.02). However, against recommendations, there were no *a priori* predictions on the size of the effect for each group. ROC curves calculated between the minimally better/minimally worse and the no change groups indicated that the GHTQ can distinguish improved participants from those who did not change (AUC = 0.79, Optimal score = -5), but is less reliable at distinguishing worsening from no change (AUC = 0.6) (Adamchic et al., 2012a). Thus while there is evidence that the GHTQ can reliably detect changes, this would need to be re-examined in a UK population using the TQ.

Table 3.3. Clinical Global Questions to determine patients' judgement about perceived treatment-related change

a) RESET1 Clinical Global Impression	b) TRI Clinical Global Impression	c) TFI Global Question
Verbal rating of their tinnitus loudness and annoyance for each ear where the tinnitus was perceived as compared to baseline.	Rate the total improvement of their tinnitus complaints compared to before the beginning of treatment.	All things considered, how is your overall tinnitus condition now, compared to your first visit to this clinic?
1. Much better	1. Very much better	1. Much improved
2. Somewhat better	2. Much better	2. Moderately improved
3. No change	3. Minimally better	3. Slightly improved
4. Somewhat worse	4. No change	4. No change
5. Much worse	5. Minimally worse	5. Slightly worse
	6. Much worse	6. Moderately worse
	7. Very much worse	8. Much worse.

RESET: Randomized Evaluation of Sound Evoked Treatment of Tinnitus trial study; TRI: Tinnitus Research Initiative database

Tinnitus Handicap Inventory

Zeman et al. (2011) classified THI scores based on the CGI-I ratings (Table 3.3). Again a moderate correlation was reported between the THI scores and CGI-I scores (r = 0.46). However, no *a priori* predictions were made on the expected magnitude of the correlations.

The mean change scores for the different rating categories were significantly different from each other and followed the expected pattern. Medium ES was reported for the minimally better group (d = 0.74), whilst a small ES was reported for the no change group (d = 0.26). Again no *a priori* hypotheses were made on the expected magnitude of the effects. The SEM of 7 reported by Newman et al. (1998) would indicate the ES could be measurement error. However this SEM was calculated on different population and therefore may not necessarily represent measurement error in a different population.

There is some evidence that THI does have the ability to detect changes in scores, but with some degree of uncertainty.

Tinnitus Functional Index

Meikle et al. (2012) categorised the mean change scores using five ratings on Global Perception of Change at 3 and 6 months (Table 3.3). However, no correlations were examined between the rating groups and the TFI (25-item prototype 2). The TFI mean change scores followed a logical progression through the 'much-to-moderate improved' to 'moderate-to-much worse' groups in both 3 and 6 months, with the 'much-to-moderately improved' group mean change scores being noticeably larger than the score for the 'no change' group on both occasions (Figure 3.1).

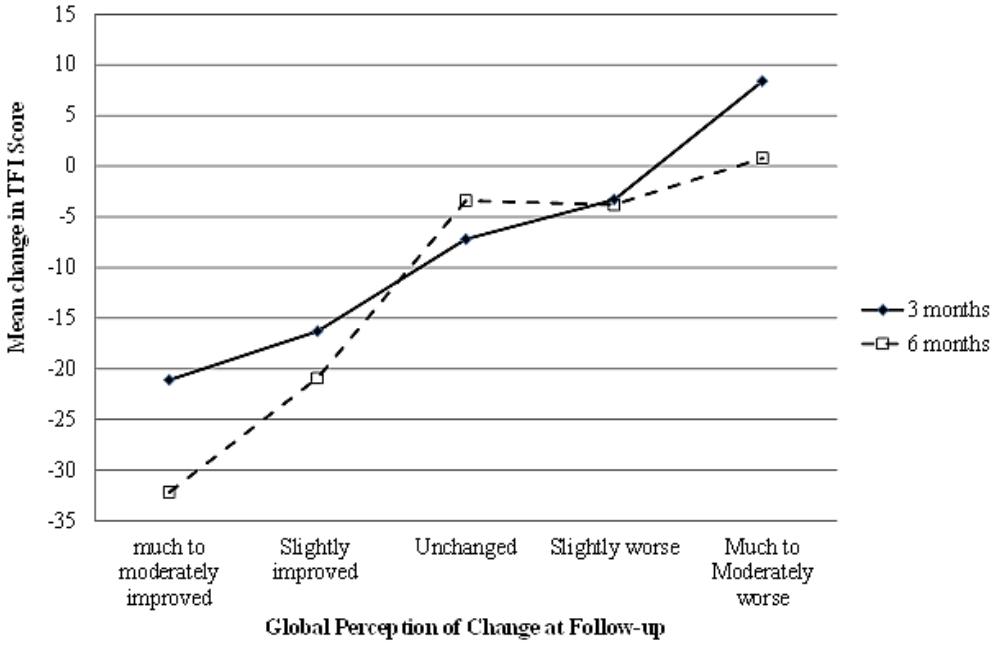

Figure 3.1. Overall mean TFI change scores at 3 and 6 month follow-ups corresponding to responses on Global Perception of Change score.
A minimal clinically important change score of -13 points was determined based on the difference between the unchanged and much-to-moderately improved groups at 3 and 6 months. Adapted from Meikle et al. (2012).

Meikle et al. (2012) also examined ES and did explicitly state the expected direction of the ES for the different ratings of change groups. In general, the results were as expected. ES for the TFI change scores in the improved groups were all positive with the much improved group resulting in a larger effect (d = 1.01) than the slightly improved group (d = 0.74).

ES for the no change group was slightly more variable than predicted, but was reasonably close to zero at both three and six months. ES for the slightly worse groups were not negative as predicted; they were very similar to the no change group on both occasions (d = 0.14), whilst the moderate to much worse groups were below zero at three and six months. Therefore, the evidence does suggest that the TFI has the ability to detect change. However, this evidence is indeterminate. Despite recognising the importance of responsiveness, Meikle et al. (2012) did not report

measurement error or the correlations, and therefore the change identified may not represent patients perception of change (strength of the relationship between the perceived change and TFI scores could not be identified) or "true change" above error as the ES cannot be disentangled from measurement error.

3.5. INTERPRETABILITY OF THE FIVE QUESTIONNAIRES

3.5.1. Interpreting the scores and grading tinnitus severity

It is important to be able to reliably grade tinnitus symptoms if treatment needs or efficacy are to be established. For the original English TQ and the TRQ, no grading systems have been developed.

Tinnitus Handicap Questionnaire

Kuk et al. (1990) have not assigned qualitative meanings to the scores, although they have suggested comparing mean THQ scores to their published normative data (n = 275). For example, a mean score of 60 would indicate the individual's tinnitus was more severe than 80% of the tinnitus patients in this sample (Figure 3.2). Although helpful in determining individual severity relative to others, it does not provide clinical interpretations of the scores (Kuk et al., 1990; Newman & Sandridge, 2004). Alternatively, a score of over 22 points has been suggested as a lower boundary for bothersome tinnitus (Sullivan et al., 1993), although there does not appear to be reliable empirical evidence to support this.

Tinnitus Handicap Inventory

Based on quartiles analysis of test-retest reliability data (Newman et al., 1998), an initial grading system was developed with four categories defining no handicap, mild, moderate, and severe handicap (Table 3.4).

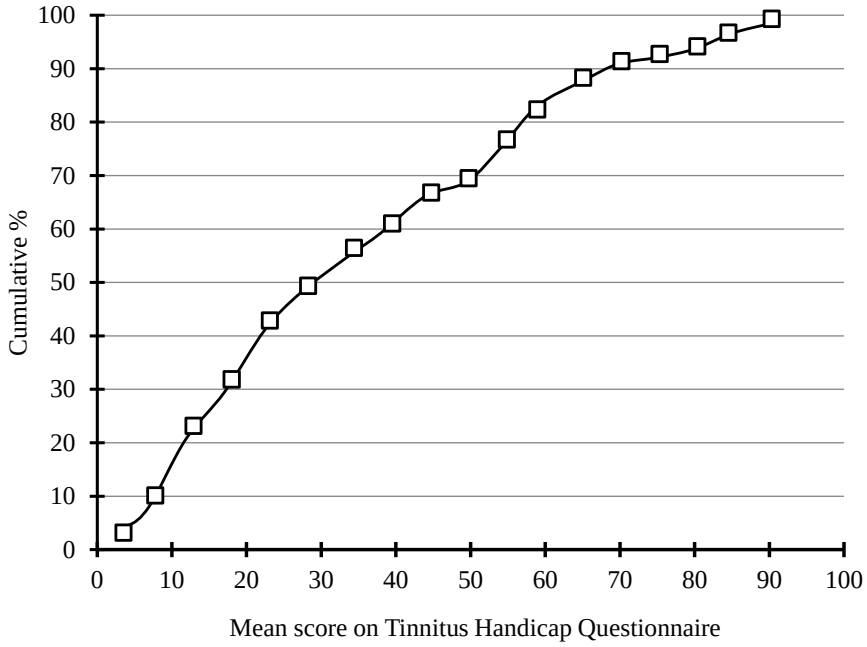

Figure 3.2 Cumulative distribution of total scores on the Tinnitus Handicap Questionnaire (THQ).
Data obtained from 275 patient responses (Kuk et al., 1990). Adapted from Kuk et al. (1990).

However, no qualitative evidence for the category interpretations was provided and given the small sample size (n: 290) the categories may not be representative of normative data. The grading system was extended into a five categories by a UK working group in 2001 (McCombe et al., 2001) (Table 3.4). Each category was provided with a detailed description of the meaning behind the scores and the last category range from the initial grading system was subdivided into scores representing severe and catastrophic tinnitus handicap. This recommendation was based on the expert opinions only. No empirical evidence has been provided on the validity of the definitions or the boundaries of the scores. This prompts questions about the reliability of these categories to provide clinicians with valid meanings behind the scores.

Table 3.4. Grading systems providing qualitative meanings to the quantitative scores

Questionnaire	Score	Tinnitus handicap	Description of symptom severity
THQ[a]	>22	Bothersome tinnitus	
THI[b]	0 – 16	Slight	"Only heard in a quiet environment, very easily masked. No interference with sleep or daily activities".
	18 – 36	Mild	"Easily masked by environmental sounds and easily forgotten with activities. May occasionally interfere with sleep but not daily activities".
	38 – 56	Moderate	"May be noticed, even in the presence of background or environmental noise, although daily activities may still be performed. Less noticeable when concentrating. Not infrequently interferes with sleep and quiet activities".
	58 – 76	Severe	"Almost always heard, rarely, if ever, masked. Leads to disturbed sleep pattern and can interfere with abilities to carry out normal daily activities. Quiet activities affected adversely. There should be documentary evidence of the complaint having been brought to the general medical practitioner. Hearing loss is likely to be present but its presence is not essential".
	78 – 100	Catastrophic	"All tinnitus symptoms at level of severe or worse. Should be documented evidence of medical consultation. Hearing loss is likely to be present but its presence is not essential. Associated psychological problems are likely to be found in hospital or general practitioner records".
TFI[c]	< 25	Mild	
	25 – 50	Significant problems	
	> 50	Severe	

a. data from Sullivan et al., 1993, b. data from Newman et al. (1998), c. 25-item prototype 2 data from Meikle et al. (2012).

Tinnitus Functional Index

A preliminary grading system was developed using the 25-item prototype 2 data. The mean scores were initially categorised based on individual responses to the global question asking 'How much of a problem is your tinnitus?' on a five-point scale ('not a problem' (1), 'a small problem' (2), 'a moderate problem' (3), 'a big problem' (4), 'a very big problem' (5)). Response distributions and the modal range of the total scores corresponding to the five categories were examined. From this inspection of the scores, the authors classified the TFI scores into three categories representing mild problems, significant problems, and severe tinnitus, with descriptions indicating the intervention needed (Table 3.4). However, there is very little clarity on the process in which the final categories were identified. For example, they only refer to the modal categories as "support" for classifying the TFI into three categories, but it is not apparent from the modal range for five original categories why the authors choose the ranges within the three grades. No empirical evidence was provided on the qualitative interpretation of the scores.

3.5.2. Interpreting changes in scores and identifying Minimal Important Change

Tinnitus Questionnaire

No data have been provided to determine a minimal clinically important improvement score for the English version of the TQ. Therefore, until this is rectified, I would advise not to rely on the TQ for clinical audit.

Minimal important change scores for the GHTQ have been proposed (Goebel et al., 2006; Adamchic et al., 2012a). Having categorised treatment effects on the GHTQ, Goebel et al. (2006) proposed a reduction of 6-14 points was indicative of "responders" and ≥15 points of clear improvement ("winners"). Adamchic et al.

(2012a) identified four potential improvement scores using the analysis discussed in section 3.4.2. The ROC curve estimate of a reduction in TQ score of 5 was considered the most representative of improvement as it was similar to the score received by patients and above error (Adamchic et al., 2012a). Although this improvement score of 5 has been utilised as representation of minimal important improvement in the TQ scores, no analysis has been conducted to confirm this score for the English version of the TQ or to define a minimal important change score for the UK population.

Tinnitus Handicap Questionnaire

Despite claims to effectively evaluate treatment-related change, Kuk et al. (1990) report no information on a minimal important change score. An important change score was identified using the test-retest reliability data (Newman et al., 1995). Having calculated 95% CI for the difference scores (20.9) and SEM consistency for the total THQ (7.4), Newman et al. (1995) conclude that a reduction of ≥21 points in the mean total THQ score between pre- and post-intervention would need to occur for the change to be considered clinically relevant. However, this value may be compromised by lack of consistency and reproducibility of the items in factor 3.

Tinnitus Reaction Questionnaire

Wilson et al. (1991) do not provide evidence for a minimal clinically important improvement score and one has not been determined since.

Tinnitus Handicap Inventory

Minimal important change improvement scores have been proposed for the THI, even though it was not developed as an outcome tool and will always receive criticism for its inherent lack of sensitivity to change. The first related to test-retest

reliability data (Newman et al., 1998). Using the 95% CI and the SEM, a reduction in THI score of 20 points was classified as being clinically meaningful. However, this score is dependent on the intake assessment score being >20 points.

Zeman et al. (2011) proposed a reduction of 7 points as a minimal important change score for improvement, calculated using the ES (and 95% CI) separating the minimally better and no change groups (d = 0.5) and the pooled SD of the difference scores (SD = 14). This is differs from Newman et al. (1998), but does not consider measurement error or the variability in scores between assessments. Essentially it only identifies the magnitude of the change.

Tinnitus Functional Index

Mean changes on the TFI global score were categorised into five groups based on global perception of change ratings at three and six months (Table 3.3). The difference between the change scores in the much-to-moderately improved group and the unchanged group at three months (14 points) and the slightly improved group and the unchanged group at six months (17 points) were examined in relation to the half of SD value of the initial total scores at baseline ($SD_{/2}$ = 12.5). From this, Meikle et al. (2012) proposed a reduction of 13 points to represent an improvement over time and to be meaningful to patients (Figure 3.1). However, this does not take into consideration the measurement error. Further work is needed.

3.6. SUMMARY

The five tinnitus questionnaires have different strengths and weaknesses. Table 3.5 summarises the evidence reviewed and reported in this section according to eleven quality criteria for good measurement properties, arranged under the four key categories (validity; reliability; responsiveness and interpretability).

Table 3.5. Summary of the critical evaluation of the psychometric properties of the five tinnitus questionnaires.

	VALIDITY				RELIABILITY			RESPONSIVENESS		INTERPRETABILITY	
			Construct validity								
	Content	Structural	Convergent	Discriminant	Internal consistency	Reliability	Agreement	Floor or ceiling effects	Ability to detect changes in scores	Grading system	MIC
TQ	?	?	?	+	+	?	0	0	?	?	?
THQ	+	+	?	+	+	+	?	0	?	?	?
TRQ	?	?	?	?	–	+	0	–	?	0	0
THI	?	?	+	?	+	+	?	0	+	+	?
TFI	+	+	+	+	–	+	0	–	?	?	+

Rating: + = positive, 0 = no information available, – = poor, ? = indeterminate rating (doubtful design or method, limited information), where two symbols are given this indicates that several different validation criteria were used.

The TRQ meets only one of the criteria. The TQ meets only two of the criteria. The evidence was unclear for the British version of the TQ and in most cases based on a German translation (GHTQ). The THQ and THI each meet five criteria.

The TFI come out "best", meeting six of the quality criteria. The TFI did not meet the criteria for internal consistency, scoring above the predefined $\alpha > 0.95$, or for floor and ceiling effects; a large number of items were rated at 0 or 10 by more than 15% of patients. Although test-retest reliability was conducted, no information so far has been provided on absolute agreement. The information on responsiveness and the grading system was indeterminate, due to lack of clarity and not using the most appropriate method.

However, it should be noted that the order of the items between prototype 2 and the final TFI changed, this could cause item order effects, where the meaning of the items change dependent on the order they are completed. Therefore, prototype 2 data may not be the "best" representation of the TFI. The final 25-item version of the TFI has not been subjected to formal evaluation.

CHAPTER 4. UK VALIDATION OF THE TFI IN A LARGE CLINICAL COHORT

4.1. INTRODUCTION

The TFI was developed to specifically provide a valid measure of the impact of tinnitus and be responsive to treatment-related changes, but the reliability and validity of the TFI needs to be examined within the population in which the questionnaire is intended for use. Just because a questionnaire has been found to be reliable measure in one population, it does not mean that same questionnaire will be reliable in a different population. In order to provide empirical evidence on the validity and reliability of the TFI and establish whether the TFI is appropriate for use as a clinical assessment tool in a UK clinical population, the psychometric properties of the TFI were evaluated.

4.2. AIM AND HYPOTHESIS

This study assessed the reliability of the eight-factor TFI structure reported by Meikle et al. (2012), and whether the TFI provides (i) comprehensive cover of the broad range of symptoms associated with tinnitus severity, (ii) a reliable measure the functional impact of tinnitus, distinguishing between individual differences in tinnitus-related distress, (iii) a responsive measure of change in symptoms, and (iv) interpretations to the scores and the change in scores.

4.3. METHODS

This was a prospective multi-site, longitudinal questionnaire validation study recruiting first-time tinnitus patients from audiology clinics across the UK to complete questionnaires at four time intervals over a 9 month period.

4.3.1. Approvals

This study was conducted in accordance with the permissions granted by the Cornwall and Plymouth Research Ethics Committee and the Sponsor (Nottingham University Hospitals NHS Trust). Permissions were obtained from Research and Development departments within each NHS trust and the equivalent trusts in Scotland, Ireland and Wales. A Research and Development contact was identified within each hospital trust to obtain the required approvals for each audiology clinic. This included information on the feasibility of the study, for example the amount of allocated time the clinicians would require to complete all the required study documentation and the expected expenses. Recruitment only commenced at each site once these approvals were obtained for the site. The first site was approved in October 2013 and last site was approved in March 2014. Written informed consent was obtained for all participants by the Principal Investigator at each site who had the required training to consent participants. All data were anonymised and stored in accordance with the Data Protection Act.

4.3.2. Participants

A sample size of 250 participants was required to conduct the intended psychometrics. This estimate was based on the data requirements needed for the CFA using the initial data (T0). To reliably assess CFA model fit, in particular the χ^2 test and RMSEA, it is recommendation that the sample size is above 200 (MacCallum et al., 1996). Although, a sample size of 5-10 participants per estimator parameter is often advocated in the literature (Nunnally, 1978; Floyd & Widaman, 1995; Schreiber et al., 2006), which would suggest more than 290 participants needed to be recruited (5:1 ratio for 53 estimated parameters in the TFI model), MacCallum et al. (1996) results indicate that models with large degrees of freedom would require a

smaller sample size than that indicated above to provide sufficient power to test model fit. Therefore, given that the degrees of freedom for the model are more than 200 (*df* 267), a sample size of 250 participants is more than satisfactory to effectively test model fit and allow for missing data. This would also account for the sample size requirements for assessing internal consistency (~250), as discussed in 2.2.2.1 In general, for the analysis of follow-up data, in particular reliability and agreement, responsiveness and interpretability analysis, a sample size of ≥ 50 is recommended for each element of the analysis, for example 50 participants in the improved group and the no change groups to assess responsiveness (see 2.2.3 for methods). Based on the work by Vernon et al. (1992) investigating unreturned postal questionnaires, a dropout rate of approximately 38% was estimated for the follow-up data collection. Therefore, the remaining sample size based on 250 participants would be sufficient to conduct the desired follow-up analysis (2.2.2 – 2.2.4).

4.3.2.1 Inclusion criteria

- Men and women ≥18 years
- Reporting tinnitus
- "First time" tinnitus patients
- Sufficient command of English language to read, understand and complete questionnaires
- Able and willing to give informed consent

4.3.2.2 Exclusion criteria

- Unable to independently complete questionnaires

Although the patients were recruited as "first time" tinnitus patients, this does not mean that their experience with tinnitus was new. The term "first time" in this case

means that none of the participants had been treated or attended a clinic in past 6 months before their appointment.

4.3.2.3 Withdrawal criteria

Participants were withdrawn from participation in the study if they no longer wished to participate in the study, or if, after receiving two reminders they failed to return the complete follow-up questionnaire packs. During the consent process, participants are informed that if they withdraw then their data collected to that point would not be erased and would be used in the final analyses. In the event of a participant withdrawing this information was reiterated.

4.3.3. Settings and recruitment

NIHR Nottingham Hearing BRU led the validation study co-ordinating with 12 additional sites (NHS audiology clinics) in the recruitment of new tinnitus patients. A single identified individual from the clinical care team at each NHS site was responsible for identifying participants, consenting, and collecting the first visit (T0) data.

4.3.3.1 Identification of sites

Email invitations to become a recruitment site were sent to 24 audiologists/hearing therapists from NHS audiology clinics (n=21) and independent sector clinics (n=3) from around the UK. These clinicians were identified through the British Society of Audiology clinical network, known tinnitus experts in the field, and clinicians with previous involvement with tinnitus team members and projects conducted within Nottingham Hearing BRU. Twenty-one audiologists replied to the initial email. The study co-ordinator contacted each clinician to discuss the intended study protocol and enquire about the sites procedures for booking patient appointment and sending

information before appointments. The information gained from each site was incorporated into the study protocol to ensure that the study followed current clinical procedures. Each site reviewed the study protocol before beginning the ethics and research and development approval process. At this point, 14 NHS sites and 2 independent sector clinics agreed to participate. Three NHS site withdraw during the ethics procedures due to time constraints within the clinics, such as expected workloads for the dates of the study.

4.3.3.2 Independent sector sites

Two independent sector clinics were approved to participate; (i) The Tinnitus Clinic, London and (ii) The Clitheroe Therapies Clinic, Lancashire. The Tinnitus Clinic agreed to recruit 20 new tinnitus patients, whilst Clitheroe agreed to recruit 5 tinnitus patients. However, both sites deviated from the study protocol. These deviations included failure to respond to information requests from the study co-ordinator on study progress (i.e. progress reports on recruitment and number of invitations sent), changing study exclusion criteria without informing the study co-ordinator, (i.e. excluding participants with severe tinnitus). No participants were recruited by either site during this time. Following discussions with the study sponsor, the decision was made to close these sites and closure was completed in March 2014.

4.3.3.3 NHS Audiology sites

Eleven NHS clinics were initially established as recruitment sites (Table 4.1). Each site was expected to recruit 20 participants, except for Cambridge, who specified a target of 25. Recruitment start dates for the sites were dependent on the date of the approvals which varied across the sites. Recruitment was anticipated to be complete by June 2014 for all sites.

Table 4.1. Number of participants providing initial and follow-up data at each NHS audiology site.

Procedure model	NHS Audiology sites	Follow up questionnaires			
		Initial data	3 months	6 months	9 months
B	Aintree University Hospital NHS Foundation Trust, Liverpool	20	14	12	12
B	Belfast Health and Social Care Trust, Belfast	20	16	12	10
A	Brighton & Sussex University Hospitals NHS Trust, Brighton	15	10	9	8
A	Cambridge University Hospitals NHS Foundation Trust, Cambridge	26	25	24	23
A	Cardiff & Vale University Health Board, NHS Wales, Cardiff	20	11	9	9
A	Central Manchester University Hospitals NHS Foundation Trust, Manchester	23	16	15	14
A	Countess of Chester, Chester	10	7	6	6
A	Doncaster and Bassetlaw Hospitals NHS Foundation Trust, Doncaster	41	30	26	25
A	NHS Fife, Kirkcaldy	20	18	17	14
A	Nottingham University Hospitals NHS Trust, Nottingham	19	13	11	11
A	Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich	20	19	17	16
A	Sherwood Forest Hospital NHS Foundation Trust, Mansfield	21	19	18	18
	Total number of participants (% of total dropout)	255	198 (22%)	176 (31%)	166 (35%)

This ensured that the final 9 month follow-up questionnaires for all participants from all the sites were completed by June 2015. The recruitment strategy was monitored monthly and revised accordingly. Table 4.1 provides an overview of the recruitment numbers at each site.

Following the closure of the two independent sector sites, in March 2014, Doncaster agreed to increase the recruitment target to 40 participants and an additional NHS site, the Countess of Chester hospital, was identified and approved to recruit 10 participants within a restricted time frame (March 2014 - May 2014). In total, 12 NHS audiology clinics from the UK recruited participants to the study over an 8-month period.

Recruitment stopped when 250 participants had been consented on to the study. However, due to the procedure of sending out invitations and questionnaire packs with appointment letters, if participants brought the completed questionnaire packs to the clinical appointment then they were included in the study even if the site had reached the target. As consequence some sites recruited more than anticipated whilst others were asked to stop before reaching the target. Recruitment started October 2013 with the final participant recruited in June 2014. A total of 255 participants were recruited.

4.3.4. Procedures

The study design and procedure was adapted from Meikle et al. (2012), in particular the number of questionnaire packs and the time frame between the questionnaires being completed (Baseline (T0), 3 (T1), 6 (T2) and 9 months (T3)) (Figure 4.1). This allows for backwards comparisons to the original study findings in particular, the proposed grading system and the minimal important change score.

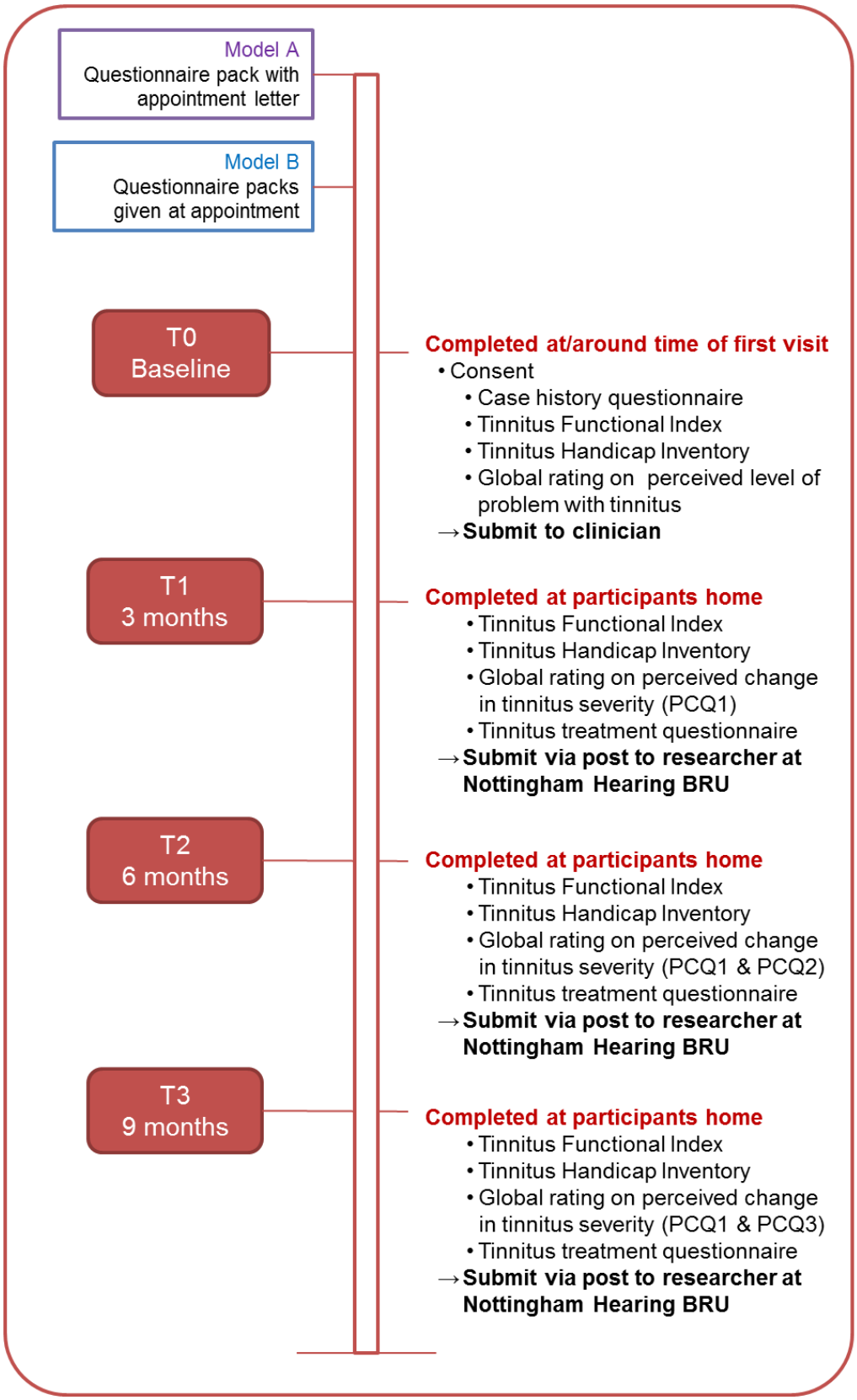

Figure 4.1. Flow diagram of project timeline for model A and model B.

The study required a commitment of four sessions over a 9-month period, during which each participant completed a set of questionnaires which included two tinnitus questionnaires (TFI and THI), a questionnaire on tinnitus history and questions on the global ratings of perceived tinnitus problem (baseline only) and global ratings of perceived change in tinnitus, both of which were adapted from the original study (section 4.3.4). At five of the sites, the THI was completed as part of standard procedure at the first clinical assessment appointment. Different sets of questionnaires were completed at the four time intervals, except for the two tinnitus questionnaires which were completed at every time (Figure 4.1).

4.3.4.1 Baseline collection procedure

At each site, questionnaires were administered following one of two models which allowed for sufficient flexibility to fit in with the different patient appointment booking procedures and information sent before appointments between the sites. For Model A, the questionnaire packs were sent to the identified tinnitus patients with their appointment letters and participants were asked to complete T0 questionnaire packs on the day of their assessment appointment and return the pack to the clinician at the appointment. At the assessment appointment, the clinician discussed the information provided, answered any queries and gained written consent, ensuring that the participant understood all aspects of what the study involved.

For Model B questionnaire packs were given to tinnitus patients at the assessment appointment and after discussing the information provided and gaining written consent, participants were asked to complete T0 questionnaire packs within 48 hours of the appointment and return the pack to the clinician from their designated clinic. Although, for this model, the participants could have directly returned T0 questionnaires to the study coordinator, returning questionnaires directly to clinicians

has been shown to increase return rate and compliance (Edwards et al., 2002, 2009). Clinicians from all the sites sent the completed T0 questionnaire packs to the main study site. Each participant was provided with a unique identifier to ensure anonymity of the data and all questionnaire data was uploaded onto a secure excel database.

4.3.4.2 Follow up collection procedure

The collection of all follow-up data was managed by the study co-ordinator (myself). Participants were mailed followed-up questionnaire packs with prepaid return envelopes at 3 months, 6 months and 9 months from their initial assessment appointment date. Packs were mailed two weeks before they were due to be completed, giving enough time for participants to complete and return the questionnaires.

Participants who did not return follow-up questionnaire packs (T1, T2 and T3) within two weeks were contacted by their preferred stated method (email, post or telephone), to ensure that the questionnaire pack was received, and to remind participants to return the pack. A second reminder was sent one week after the first reminder if questionnaire packs had still not been received. This procedure has again been shown to increase return rate and compliance (Edwards et al., 2002, 2009).

4.3.5. Measures

4.3.5.1 Baseline Demographics (T0)

To collect baseline characteristics at T0, participants completed a 10-item case history questionnaire on age, gender, tinnitus onset, characteristics, and duration, and self-reported hearing difficulties.

4.3.5.2 Global rating on perceived level of problem with tinnitus (T0)

Participants completed a single question asking "How much of a problem is your tinnitus?" Participants choose one of five response options (1 to 5) to indicate their tinnitus currently; 1 (not a problem), 2 (a small problem), 3 (a moderate problem), 4 (a big problem), and 5 (a very big problem). This question was used in the original TFI development procedure.

4.3.5.3 TFI (T0, T1, T2, T3)

Participants rated each item according to how they have felt over the past week. Each item is rated on an 11-point scale, with descriptors at either end of the scale. The mean global score reflects the sum of all responses, weighted to give a global score out of 100. Higher scores reflect greater impact on daily functioning. The procedure for scoring the TFI followed the instructions provided by Meikle et al. (2012) such that items with two responses circled (i.e. 4-5 circled) were scored with the higher value, items with more than one value on the scale circled (i.e. 3 and 7 circled) were scored as an average of the two values, and the overall global TFI score was only calculated if the respondent completed at least 19 items.

4.3.5.4 THI (T0, T1, T2, T3)

Participants rated each THI item on a categorical 3-point scale (yes (4)/no (0)/sometimes (2)). The mean global score reflects the sum of all responses with a maximum score of 100 indicating the greatest impact on everyday function. For the purposes of analysis here the THI was considered unidimensional and the subscales scores were not calculated. Newman et al. (1996) did not provide any guidelines on how to account for missing values in the calculation of the total score. A decision was made to calculate the global score for any questionnaires missing 3 items or

fewer. The global scores are classified based on THI grading system to quantify the level of tinnitus severity (see Table 3.4).

4.3.5.5 Global rating on Perceived Change in tinnitus Question (PCQ)

Participants completed a couple of questions focused on the extent to which they perceive their tinnitus severity has changed within a two different time frames. These questions were adapted from the original TFI development study:

> At T1, T2, and T3, participants were asked "All things considered, how is your overall tinnitus condition now, compared to 3 months ago?"
>
> At T2, participants were asked "All things considered, how is your overall tinnitus condition now, compared to 6 months ago?"
>
> At T3, participants were asked "All things considered, how is your overall tinnitus condition now, compared to 9 months ago?"

Participants rate each question on a 7-point response scale with descriptors (prompts) of the extent of change; 3 (much improved), 2 (moderately improved), 1 (slightly improved), 0 (no change), -1 (slightly worse), -2 (moderately worse) and -3 (much worse).

4.3.6. Analysis

The methods applied to the questionnaire data collected from the clinics are described in detail in Chapter 2. Listed below are the specific methods used and specifications that apply to this dataset in particular. CFA was performed in Mplus 7 (Muthén & Muthén, 2012), whilst validity, reliability, responsiveness and interpretability analyses were calculated in SPSS (v.21.0) and Microsoft Excel.

4.3.6.1 Confirmatory factor analysis

To establish the fit of the eight-factor structure devised by Meikle et al. (2012) in our clinical population, CFA was conducted on TFI data collected at baseline. Missing data was less than 7% and was identified as "missing completely at random". Listwise deletion is considered an effective approach to deal with small amounts of missing data (Schafer & Graham, 2002) and avoids any problems associated with estimating data, such as over-estimating factor estimates (Tabachnick & Fidell, 2013). Therefore, only those who completed all 25 items on the TFI were used for the analysis of the factor structure and so after list-wise deletion the effective sample size was 239.

As a first step the TFI item data were screened for outliers, linearity and multicollinearity since non-normality of data can have adverse effects on the CFA. There was no evidence of univariate outliers using standardised *z*-scores distribution or boxplots. Mahalanobis distance statistic indicated that there were eleven multivariate outliers with the greatest distance from the rest of the data points (Mahalanobis d-squared: 81.5 to 55.0, $p \leq 0.001$). From the boxplots for the items (section 4.4.7), it is apparent there is some evidence of skewness, however kurtosis and skewness did not exceed the recommended cut-off points (Curran et al., 1996). This indicates some non-normality in the distribution of the data, requiring control and therefore, an adjusted estimation method was applied to the current dataset; maximum likelihood parameter estimation with Chi-square adjusted for non-normality in the data (Satorra-Bentler scaled Chi-square (see 2.2.1.2)).

Initially the model was estimated with just the eight first-order factors, allowing for examination of the correlations among the factors, and then the model is estimated with the second-order factor. The proposed eight-factor structure provides

the basis for the model estimated parameters in the CFA. The first-order factor model was defined by three major properties: (i) latent constructs corresponded to the TFI subscales which were freely estimated to correlate with each other (8 estimated parameters); (ii) observed variables corresponded to the 25 items which were constrained to zero loadings on the other factors (no designated cross-loading: 25 estimated parameters); (iii) residual variance (error/uniqueness terms) associated with each observed variable were assumed to be uncorrelated and random (constrained to zero; 25 estimated parameters) (Figure 4.2).

The second-order factor model included the additional latent construct corresponding to the functional impact of tinnitus. It incorporated an additional property parameter: (iv) the variance of the second-order factor was fixed at 1 as it was assumed that the variance in the first-order factors could be completely explained by the relationship to the second-order factor (Figure 2.1).

4.3.6.2 Construct validity – Convergent validity

To evaluate convergent validity the TFI global and subscale scores were compared to THI global scores over the four time intervals. High convergent validity (<0.75) is expected between the two global scores (Meikle et al., 2012). In terms of the subscales, the THI mainly focuses on elements of tinnitus handicap in relation to psychological and emotional distress and impact on lifestyle. Therefore it was predicted that the eight TFI subscales, QoL and Emotional, would have the strongest correlations (<0.60) with the THI. Convergent validity was examined using Pearson's correlations and pairwise deletion to ensure the largest sample sizes for each comparison.

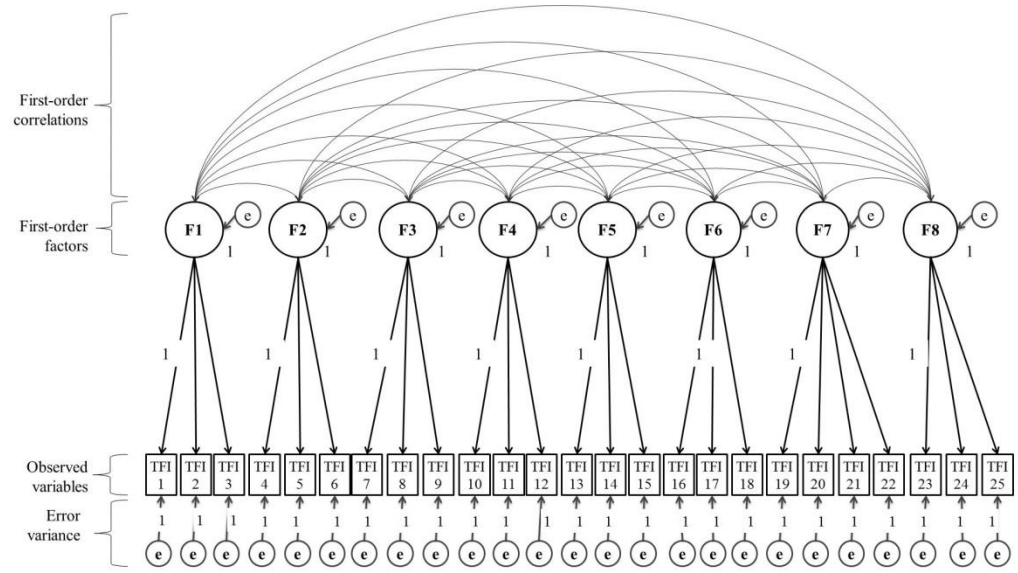

Figure 4.2. First-order TFI factor model.
The model represents the proposed relationships between the observed variables (items i.e. TFI 1) and the first-order factors (F1 to F8). Bidirectional curved arrows (↷) = covariance between first-order factors. Unidirectional black arrows (→) represent the direct effects of the first-order constructs onto the observed variables (items): Variance is fixed at 1 on the first item on each factor. Unidirectional grey arrows (→) represent the residual variance (e) associated with each variable . F1: Intrusiveness; F2: Sense of control; F3: Cognition; F4: Sleep; F5: Auditory; F6: Relaxation; F7: Quality of life; F8: Emotional. 1 = fixed variance; e = error

4.3.6.3 Reliability - Internal consistency

Cronbach's alpha and inter-item correlations were calculated for the items in the TFI global and subscales on data from all four time intervals. The follow-up data were used to examine the consistency of the alpha estimates. For comparative purposes, the alpha estimates for the THI items were also calculated. Cronbach's alpha can only be calculated on complete data, listwise-deletion is automatically conducted.

4.3.6.4 Reliability - Test-retest reliability and agreement

It was not possible to conduct a test-retest situation before the first appointment. As an alternative, the data from the "no change" responses to the two global rating of perceived health-status change questions was used to assess reliability (degree of variability) and level of agreement in scores over time (2.2.2). Only participant data

from those identified as "no change" at both 3 and 6 months corresponding to perceived change since baseline and/or change over the past 3 months was used. The sample size for the 9 month data of the "no change" group of participants for both questions was below the recommended size required to conduct the analysis (n = 29). The scores at 3 and 6 months were inspected for extreme outliers, participants with changes in scores of above 70 were removed as a score this high was considered to be inconsistent with their perceived "no change" since a large change such as this would correspond to change from severe tinnitus to mild tinnitus or vice versa. Separate analysis was conducted on the no change response data from responses to the global ratings of change compared to baseline ("baseline comparison" group) and the compared to 3 months ago ("3 month comparison" group).

$ICC_{agreement}$, limits of agreement and SDC were calculated for the TFI global and subscale scores for the separate groups. To account for the total shared variance over the three time intervals, a one way ANOVA was conducted for each analysis to identify the SD of the difference. The SD_{diff} was used to calculate the limits of agreement and SEM for consistency (SEM_{con}). The SEM for agreement (SEM_{agree}) was also calculated. In terms of the TFI subscales, it was felt that considering these subscales are recommended as standalone measures then the limits of agreement and smallest detectable change estimates should be examined to assess the degree of measurement error and variability in the change scores.

4.3.6.5 Responsiveness – floor and ceiling effects

Response frequency distributions for baseline item level data were examined to detect the presence of floor and ceiling effects. Considering that potentially more floor effects would be observed in the follow-up data due to improvements, the data

were used to confirm the consistency of the floor and ceiling effects observed in the baseline. Missing data for each item are reported.

4.3.6.6 Responsiveness – detecting changes in scores

Similar to reliability and agreement, separate analyses were conducted on the change scores stratified by responses to the global ratings of change compared to baseline ("baseline comparison" group) and the compared to 3 months ago ("3 month comparison" group). For the subscale analysis, the global ratings of change categories were collapsed to three categories defining improvement (1), no change (0) and worsening (-1) to ease interpretation. ROC analysis was conducted on global TFI change scores only. For ES calculations, only the global ratings of perceived change compared to baseline were used to stratify the TFI global and subscales scores, and for comparative purposes the THI global scores at 3, 6 and 9 months. To ensure large sample sizes in the ES analysis the rating of change groups were collapsed into three categories defined as improved, no change and worsened. All data from 3, 6 and 9 months were used. There were no missing data for the global change question.

It was predicted that there would be a notable differences in the TFI change scores for different levels of perceived change for both the baseline and 3-month comparisons with higher mean change scores in the worsened groups than any of the others and lowest mean change scores in the improved groups. Additionally, notable differences were expected in mean change scores between the 'no change' group and the 'improved' and 'worsened' groups. The 'much-to-moderately' improved and worsened groups would be expected to show the largest difference in change scores from the 'no change' group across all time intervals. The TFI global and subscale changes scores for each time interval were also expected to moderately correlate with

the perceived change responses (< 0.5). Based on Meikle et al. (2012), it was predicted that the ES for the 'improved' groups would have medium to large positive values, the 'worsened' groups would have small to medium negative values and the 'no change' group would be close to zero.

4.3.6.7 Interpretability – grading system

To provide meaning to the global TFI scores and identify grades of symptom severity, quartile analysis, distribution of the TFI scores and ROC analysis were conducted on the baseline global TFI data. Anchor-based approaches were used to classify the responses to the global TFI, the THI grading system and the perceived level of problem identified using the global rating of perceived problem at baseline. Only four of the five response categories for global rating of perceived problem question were employed to classify the global TFI scores. The "no problem" category was not endorsed by any participant and therefore was not utilised. No missing data was reported for the global TFI scores, but two participants did not complete the global rating of perceived problems and therefore their responses were not included in the examination of distribution within these categories.

4.3.6.8 Interpretability – interpreting changes in the scores and identifying a minimal important change score

To identify minimal important change and optimal ROC value for the global TFI, the change scores from 3, 6, and 9 months were examined within the five ratings of perceived change (much-to-moderately improved/worse, slightly improved/worse and no change). To examine the effects of baseline values on the minimal important change, the global TFI scores were classified within the problems grades identified in the previous analysis and then further sub-divided into the perceived change ratings. The response categories for the ratings of perceived change were collapsed into three

categories (improved, no change and worsened) to ensure sample size was sufficient for the analysis. The global TFI scores at 3 months (largest sample size) were used in the visual anchor-based MIC distribution plot. The estimates identified were then plotted with the SEM and smallest detectable change estimates to identify the range in important change scores and recommend a minimal important change score.

4.4. RESULTS

4.4.1. Participants

4.4.1.1 Baseline characteristics

A total of 255 tinnitus patients (male: 149 (59%), female: 105 (41%)) were recruited from 12 NHS audiology clinics. The average age for the participants was 53.6 years with age range of 18 to 84 years. Table 4.2 summarises the characteristics identified with the tinnitus case history questionnaire.

Just under 50% of participants had experienced tinnitus for less than 2 years, 30% reported tinnitus duration between 3 to 10 years, and the remainder reported experiencing tinnitus for more than 11 years. Descriptors of tinnitus sounds included whistling, buzzing, ringing, hissing, clicking, cracking, whooshing, and old TV static. The majority of participants felt their tinnitus had gradually appeared, was present constantly or most of the time, and did not know the cause.

4.4.1.2 Follow-up completion rates

Follow-up questionnaires were completed by 198 (78%) participants at 3 months (T1), 176 (69%) participants at 6 months (T2), and 166 (65%) participants at 9 months (T3). The largest dropout was at T1, in which 57 participants (22%) did not return the questionnaire packs.

Table 4.2. Tinnitus characteristics of participants at baseline

Initial data	(n = 255)	
	n	(%)
When did you first experience your tinnitus?		
less than 1 yr	60	(24)
1 to 2 yr	64	(25)
3 to 5 yr	43	(17)
6 to 10 yr	30	(12)
11 to 20 yr	15	(6)
20+ yr	33	(13)
Missing	10	(4)
How did you perceive the beginning?		
Abrupt	100	(39)
Gradual	154	(60)
Missing	1	(1)
Was the onset of your tinnitus related to:		
Change in hearing	28	(11)
Whiplash	2	(1)
Stress	26	(10)
Head trauma	8	(3)
Loud sound	37	(15)
Don’t know	113	(44)
Other cause	38	(15)
Missing	3	(1)
Does your tinnitus seem to pulsate?		
Yes, with heart beat	31	(12)
Yes, different from heart beat	42	(16)
No	176	(69)
Missing	6	(2)
Where do you perceive your tinnitus?		
Right ear	41	(16)
Left ear	48	(19)
Both ears, worse in left	53	(21)
Both ears, worse in right	41	(16)
Both ears, equally	57	(22)
Inside the head	15	(6)
Elsewhere	0	
Missing	0	
About how often does your tinnitus seem to be present?		
Present occasionally	4	(2)
Present some of the time	21	(8)
Present most of the time	64	(25)
Present always constant	164	(64)
Missing	2	(1)

Dropout reduced at each follow-up, with 10% dropout at T2 and <5% dropout at T3. The dropout did not exceed the expectation of 38%. In terms of reminders, the majority of participants were contacted through email or post (<80%). Telephone reminders were found to be less effective as questionnaire packs were not returned without an alternative reminder method, i.e. post. Compliance following the reminders was reasonably high at each of the follow-ups. For example, at T1 (3 months pack), 209 reminders were sent to 141 participants and 60% of those participants returned the questionnaires, at T2 (6 months), of the 91 participants sent reminders (125), 76% returned the questionnaires, whilst at T3 (9 months), 87% of the 78 participants sent reminders returned the questionnaires (Figure 4.3).

4.4.1.3 Treatment tried since baseline

At 3 months, 139 of the 198 (70%) participants reported having tried different treatments for their tinnitus with hearing aids, tinnitus maskers and portable-sound generating devices being the most commonly applied treatment (Figure 4.4). Fewer participants reported treatments at 6 (n=31) and 9 months (n=22), with hearing aids and medications for sleep and relation training being the most common. Over the 9 months, 45% of participants reported having tried more than one treatment, with 24% reporting using three or more treatments. Additional treatments that were listed include yoga, mindfulness, hypnosis and the radio playing.

4.4.2. Missing item data

From all four time intervals, of the possible 19,875 item values, only 16% of the data were incomplete before the participants were contacted to provide a response (8% of the participants provided a response). Item 3 (n=31) and item 22 (n=25) had noticeably larger amounts of missing data compared to the other items (n=0 to 10).

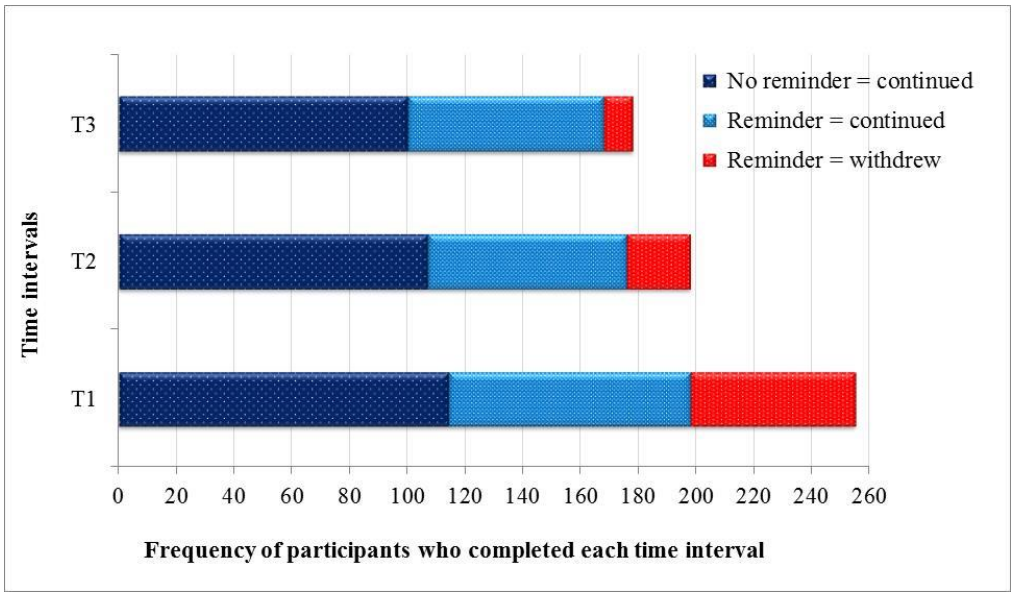

Figure 4.3. Compliance rates following reminders

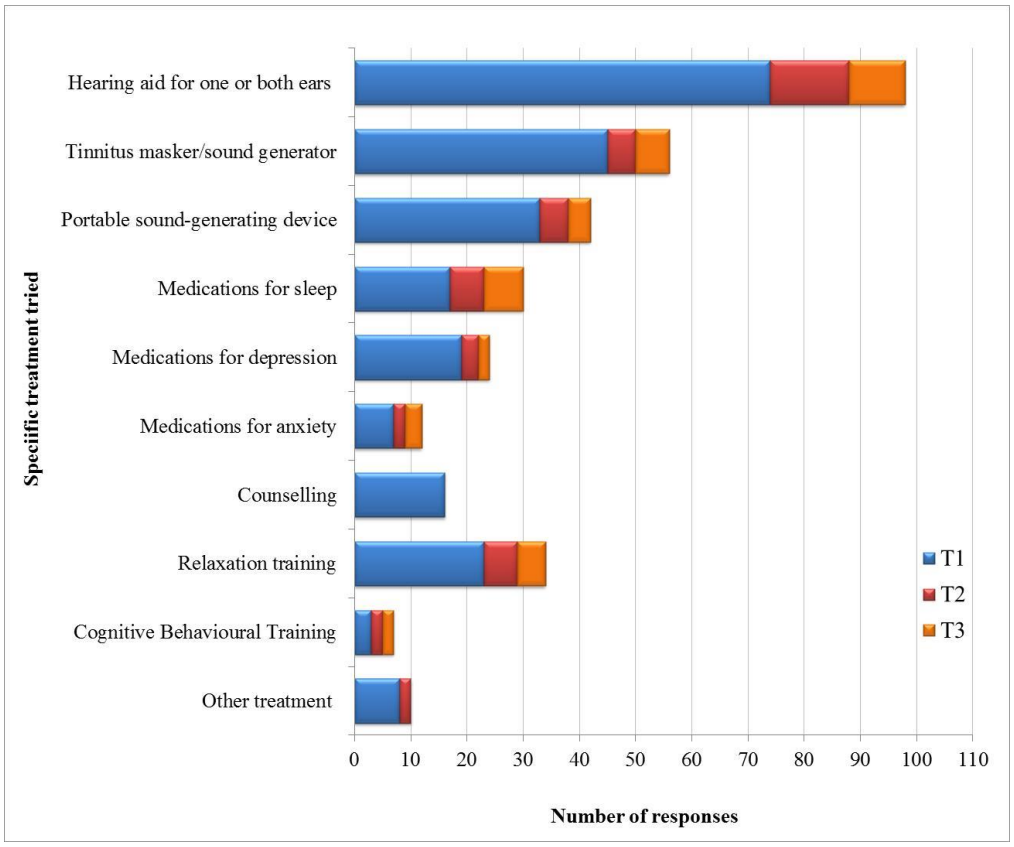

Figure 4.4. Number of specified treatments tried over the nine months.

4.4.3. Inspection of the distribution of the scores

Descriptive statistics for the TFI global and subscales scores and the THI global score over the four time intervals are presented in Tables 4.3. The mean scores on the TFI and THI were moderate (~ 50/100 in each case). The TFI global scores were 53 (at baseline) to 43 (at 9 months). Similarly, the THI mean scores reduced by 9 points over the 9 months. Interestingly, the biggest decrease in the TFI and THI global scores occurred in the first 3 months.

In terms of the global question rating their perceived level of problem at baseline, around 50% of participants defined themselves as having a moderate problem with tinnitus, consistent with the TFI and THI mean scores at baseline (Table 4.4). Over 35% of participants identified with having a "big to very big problem" with their tinnitus, whilst fewer reported a small problem with tinnitus and none reported "no problem" with tinnitus, which is hardly unexpected given that all the participants were recruited from their first visit to the audiology clinics about their tinnitus.

For the global rating of perceived change compared to baseline and 3 months ago, the highest percentage of responses were observed for the "no change" category at each time interval (Table 4.5). Fewer than 15% of participants perceived their change as getting worse at baseline (slightly, moderately, and much worse categories), none identified with the "much worse" category at 3 months, but the percentage of participants endorsing these categories slowing increased at 6 and 9 months for both questions.

Table 4.3. Descriptive statistics for the TFI global and subscale and THI global scores

	QP1					QP2					QP3					QP4				
			Quartiles					Quartiles					Quartiles					Quartiles		
Scale	n (missing)	Mean (SD)	25%	50%	75%	n (missing)	Mean (SD)	25%	50%	75%	n (missing)	Mean (SD)	25%	50%	75%	n (missing)	Mean (SD)	25%	50%	75%
TFI	255(0)	52.7±21.7	36.4	52.0	67.6	196(0)	44.7±22.4	27.3	42.2	60.3	175(0)	43.0±23.7	24.8	39.6	60.0	165(0)	42.9±25.5	21.0	39.6	61.6
INTR	251(4)	62.3±22.0	43.3	63.3	80.0	191(5)	52.3±23.8	33.3	50.0	70.0	163(12)	50.7±25.2	30.0	50.0	70.0	157(8)	48.1±25.8	26.7	43.3	70.0
SOC	251(4)	64.5±21.7	53.3	66.7	80.0	196(0)	54.4±24.6	36.7	53.3	73.3	173(2)	51.0±25.7	30.0	50.0	70.0	164(1)	52.1±27.4	30.0	53.3	73.3
COG	255(0)	47.1±26.7	23.3	50.0	70.0	193(3)	41.0±26.1	20.0	40.0	60.0	175(0)	39.3±27.1	16.7	36.7	60.0	165(0)	38.2±28.3	13.3	36.7	60.0
SLP	253(2)	55.6±31.9	28.3	63.3	80.0	196(0)	45.2±30.6	16.7	46.7	70.0	175(0)	42.4±31.1	16.7	40.0	66.7	164(1)	40.8±33.2	10.0	40.0	70.0
AUD	254(1)	42.6±30.7	13.3	40.0	70.0	194(2)	40.7±28.4	16.7	40.0	63.3	175(0)	40.7±28.7	13.3	40.0	63.3	165(0)	44.2±30.6	16.7	46.7	70.0
REL	254(1)	64.4±27.8	42.5	73.3	86.7	195(1)	53.6±26.7	30.0	53.3	76.7	173(1)	51.4±28.3	30.0	50.0	73.3	163(2)	50.9±29.4	23.3	50.0	76.7
QOL	255(0)	39.9±29.5	12.5	37.5	62.5	196(0)	33.7±27.3	10.0	28.8	55.0	175(0)	33.8±27.8	10.0	25.0	55.0	165(0)	34.2±29.0	10.0	27.5	56.3
EMO	255(0)	49.4±30.4	20.0	46.7	76.7	195(1)	39.9±29.6	13.3	33.3	63.3	175(0)	37.7±30.0	10.0	33.3	60.0	165(0)	37.3±30.9	10.0	33.3	60.0
THI	255(0)	46.1±23.8	26.0	44.0	62.0	195(1)	39.9±22.5	22.0	36.0	56.0	175(0)	38.2±23.6	18.0	34.0	52.0	165(0)	37.2±23.5	18.0	32.0	53.0

SD = Standard deviation.

Table 4.4. Frequency of responses to the global rating on perceived level of problem with tinnitus

	Initial data (n = 255)	
	n	(%)
How much of a problem is your tinnitus		
Not a problem	0	(0)
A small problem	36	(14)
A moderate problem	119	(47)
A big problem	63	(25)
A very big problem	35	(14)
Missing	2	(1)

In contrast, the number of participants identified as "improved" was more consistent across the time intervals, in general over 35% of participants identified themselves as improved, with the highest number in the "slightly improved" category at 3 months, consistent with the observation of TFI mean score reduction at this point, and the highest number in the "much improved" category at 6 months.

4.4.4. Confirming the eight-factor structure of the TFI

4.4.4.1 First-order model analysis

Correlations between the first-order factors ranged from very weak (r = 0.16) to extremely strong (r = 0.88), but most were strong, with 70% above 0.60 (Table 4.6). Notably, the Sense of Control factor showed an exceptionally strong correlation with the Intrusiveness factor indicating potential overlap in content. On the other hand, the Auditory factor only weakly correlated (<0.6) with all, except for the QoL factor. In fact, the Auditory factor appears to be completely unrelated to the Sleep, Relaxation and Emotional factors with correlations of 0.16, 0.23 and 0.27, respectively.

Table 4.5. Frequency of responses to the global rating on Perceived Change in tinnitus questions

How is overall tinnitus condition now	compared to baseline						compared to 3 months ago					
	3mths		6 mths		9 mths		3 mths		6 mths		9 mths	
	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
Much improved	8	(4)	19	(11)	14	(8)	8	(4)	11	(6)	10	(6)
Moderately improved	22	(11)	15	(9)	21	(13)	22	(11)	17	(10)	12	(7)
Slightly improved	39	(20)	33	(19)	24	(14)	39	(20)	28	(16)	26	(16)
No change	101	(51)	67	(38)	48	(29)	101	(51)	82	(47)	64	(39)
Slightly worse	23	(12)	30	(17)	33	(20)	23	(12)	28	(16)	35	(21)
Moderately worse	3	(2)	9	(5)	14	(8)	3	(2)	6	(3)	9	(5)
Much worse	0	(0)	2	(1)	11	(7)	0	(0)	3	(2)	8	(5)
Missing	2	(1)	1	(1)	1	(1)	2	(1)	1	(1)	1	(1)
Total	198		176		166		198		176		165	

Inspection of the model fit statistics indicates that, in general, the model fit for the eight first-order factor solution is acceptable. Although the S-B χ^2 was significantly large (427.52, $p < 0.0001$), S-B χ^2 relative to the degrees of freedom was below the critical ratio cut-off (1.73), the SRMR and the approximation fit indices, CFI, TLI and RMSEA, were within the recommended criterion, indicating acceptable model fit (Table 4.7).

Standardised parameter estimates revealed high factor loading estimates (> 0.70) for the majority of items with their designated factor, over 80% of items had loading values above 0.80. The standardised and unstandardised parameter estimates, R-square values and the standard errors are summarised in Table 4.8.

Table 4.6. Correlations between first-order factors.

Factor	1	2	3	4	5	6	7	8
(1) Intrusiveness	1							
(2) Sense of Control	0.88	1						
(3) Cognitive	0.74	0.79	1					
(4) Sleep	0.61	0.62	0.59	1				
(5) Auditory	0.48	0.43	0.53	0.16	1			
(6) Relaxation	0.63	0.71	0.68	0.66	0.23	1		
(7) QoL	0.62	0.70	0.80	0.49	0.65	0.61	1	
(8) Emotional	0.65	0.81	0.72	0.55	0.28	0.67	0.73	1

Loading estimates for Items 11, 14 and 17 were exceptionally high (>0.95), suggesting that these items explain most of the variance within the designated subscale, and therefore potentially indicating some overlap in item content. Only one item had a factor loading below the optimal value (0.70), Item 4 from the Sense of Control factor had a factor loading estimate of 0.61, but this is still within the critical criteria (>0.40). The squared factor loadings mirrored these findings (see R^2 in Table 4.8). For the most part, the designated factors accounted for the majority of variance in the items (>75%), but the Sense of control factor which only accounted for 37% of the variance in Item 4 and 58% in Item 6.

A quick examination of the modification indices (MI) and standardised parameter change (Stdx EPC) revealed the presence of several potential sources of misfit to the model parameters (>10). The most notable were the potential cross-loading of Item 3 (Intrusiveness) on the Sense of Control factor (MI: 15.67; Stdx EPC: 0.77) and the error covariance (uniqueness) between item 19 "*How much has your tinnitus interfered with your enjoyment of social activities?*" and item 20 "*How*

Table 4.7. Summary of the model fit.

	Models	Modified	S-B χ2 (*df*)	χ2/*df*	p-value	TLI	CFI	SRMR	RMSEA (95% CI)
	First-order	None	427.52 (247)	1.73	<0.001	0.96	0.97	0.04	0.056 (0.05 – 0.06)
Second-order models	Original TFI-25	None	577.50 (267)	2.16	<0.001	0.94	0.94	0.07	0.070 (0.06 – 0.08)
	Re-specified TFI-25	Item error covariance*	542.01 (264)	2.02	<0.001	0.94	0.95	0.07	0.067 (0.06 – 0.08)
	TFI-22	None	388.26 (202)	1.92	<0.001	0.95	0.96	0.05	0.062 (0.05 – 0.07)
	Re-specified TFI-22	Item error covariance**	360.45 (200)	1.80	<0.001	0.96	0.97	0.05	0.058 (0.05 – 0.07)

S-B χ^2 = Satorra & Bentler adjusted Chi-square; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; SRMR = Standardised Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval

much has your tinnitus interfered with your enjoyment of life?" on the QoL subscale (MI: 21.43; stdx EPC: 0.45).

Inspection of these items indicated that the large error variance might be attributable to the similarity of the question wording. No adjustments were made at this level, but these potential modifications were kept in mind in the following second-order analysis. The first-order model showed acceptable fit for the data, even though there were some potential weak correlations with the Auditory factor. Therefore the second-order structure solution was examined.

4.4.4.2 Second-order structure model of the TFI

The second-order eight-factor model (TFI-25) (Figure 2.1) was subjected to CFA. In contrast to the first-order model, the fit indices were all borderline, indicating that the fit of the data to the TFI-25 model was less than optimal (Table 4.7).

Table 4.8. Parameter estimates, R-squared values and Standard Error for the first-order TFI model.

		First-order model			
First-order factors	**Items**	**β**	**B**	**SE**	**R^2**
Intrusiveness	INTR 1	0.72	1.00		0.52
	INTR2	0.79	0.88	0.06	0.63
	INTR 3	0.86	1.31	0.01	0.73
Sense of Control	SOC 4	**0.62**	1.00		**0.37**
	SOC 5	0.87	1.18	0.11	0.76
	SOC 6	0.76	1.07	0.11	0.58
Cognitive	COG 7	0.93	1.00		0.87
	COG 8	0.94	1.04	0.03	0.87
	COG 9	0.92	0.94	0.04	0.84
Sleep	SLP 10	0.92	1.00		0.85
	SLP 11	0.97	1.05	0.03	0.95
	SLP 12	0.91	1.01	0.04	0.83
Auditory	AUD 13	0.91	1.00		0.83
	AUD 14	0.99	1.08	0.03	0.97
	AUD 15	0.94	1.10	0.04	0.88
Relaxation	REL 16	0.92	1.00		0.85
	REL 17	0.97	1.05	0.03	0.93
	REL 18	0.87	0.94	0.04	0.76
Quality of life	QOL 19	0.86	1.00		0.74
	QOL 20	0.87	0.99	0.05	0.76
	QOL 21	0.89	1.02	0.05	0.80
	QOL 22	0.83	0.96	0.05	0.70
Emotional	EMO 23	0.91	1.00		0.85
	EMO 24	0.95	0.93	0.03	0.90
	EMO 25	0.82	0.94	0.05	0.67

The values presented in bold have poor associations with their designated factor, all below the recommended cut-off < 0.40. β = Standardised parameter estimate; B = Unstandardised parameter estimate; SE = Standard Error; R_2 = R-squared.

The S-B $\chi 2$ was still significantly large ($\chi 2$: 577.5; $p < 0.001$) but now the S-B $\chi 2$ relative to the degrees of freedom was marginally larger (2.2) than the critical

ratio cut-off (≤ 2.0) indicating problems with data fit. Consistent with this, the RMSEA score (and 95% CI) was less than optimal (0.07) and since the SRMR was only just within reasonable fit criteria (≤ 0.07), the RMSEA estimate could not be considered an indication of reasonable fit (SRMR should be below 0.06 for RMSEA to be reasonable) so once again poor fit is indicated (Table 4.7). However, given that the SRMR was acceptable (≤0.07) and both the TLI and CFI estimates indicated acceptable model fit, the model may improve with slight modifications (Schreiber et al., 2006). To identify the potential source of the "less than optimal" model fit, factor loading estimates and modification indices were examined. The identified parameters were re-specified accordingly, if they improved the model fit and if they were conceptually justified.

Standardised parameter estimates for the second-order model reflected those seen in the first-order model analysis (Table 4.9). Again, over 80% of items had loading values above 0.80, the same three items had exceptionally high loading estimates and only Item 4 had a loading estimate below 0.7 (still above the critical cut-off). In terms of the second-order factor accounting for the variance in the first-order factors, two first-order factors, one in particular, appear to have a weaker relationship to the second-order than the others (Table 4.9). Both the Sleep (SLP) and Auditory (AUD) factors had factor loadings below the optimal value, although the Sleep factor is only marginally below (0.68). Perhaps unsurprisingly, given the correlations observed, the loading estimate for the Auditory factor was lower (0.50) than any of the other factors, indicating a less than optimal fit with the second-order construct.

Table 4.9. Parameter estimates, R-squared values and Standard Error for the proposed TFI-25 Model and the TFI-22 Model.

First order factor	Observed variable	Original TFI-25				TFI-22			
		β	**B**	**SE**	**R^2**	**β**	**B**	**SE**	**R^2**
Intrusiveness	INTR 1	0.71	1.00		0.51	0.71	1.00		0.51
	INTR 2	0.77	0.87	0.07	0.59	0.77	0.87	0.07	0.59
	INTR 3	0.88	1.36	0.11	0.77	0.88	1.36	0.11	0.77
Sense of Control	SOC 4	**0.62**	1.00		**0.39**	0.62	1.00		**0.39**
	SOC 5	0.88	1.17	0.11	0.77	0.88	1.17	0.11	0.77
	SOC 6	0.75	1.04	0.10	0.56	0.75	1.04	0.10	0.56
Cognitive	COG 7	0.93	1.00		0.87	0.93	1.00		0.87
	COG 8	0.93	1.04	0.03	0.87	0.93	1.04	0.03	0.87
	COG 9	0.92	0.94	0.03	0.84	0.92	0.94	0.03	0.84
Sleep	SLP 10	0.92	1.00		0.85	0.92	1.00		0.85
	SLP 11	0.98	1.05	0.03	0.95	0.98	1.05	0.03	0.95
	SLP 12	0.91	1.01	0.04	0.83	0.91	1.01	0.04	0.83
Auditory	AUD 13	0.91	1.00		0.83				
	AUD 14	0.99	1.08	0.03	0.97				
	AUD 15	0.94	1.10	0.03	0.88				
Relaxation	REL 16	0.92	1.00		0.85	0.92	1.00		0.85
	REL 17	0.97	1.04	0.03	0.93	0.97	1.04	0.03	0.93
	REL 18	0.88	0.94	0.03	0.77	0.88	0.94	0.03	0.77
QoL	QOL 19	0.87	1.00		0.75	0.87	1.00		0.75
	QOL 20	0.90	1.01	0.04	0.80	0.90	1.01	0.04	0.80
	QOL 21	0.89	1.00	0.04	0.79	0.89	1.00	0.04	0.79
	QOL 22	0.80	0.91	0.05	0.65	0.80	0.91	0.05	0.65
Emotional	EMO 23	0.92	1.00		0.85	0.95	1.00		0.85
	EMO 24	0.95	0.93	0.03	0.90	0.82	0.93	0.03	0.90
	EMO 25	0.82	0.94	0.05	0.67	0.92	0.94	0.05	0.67
Second order factor	Factor								
Functional impact of tinnitus	INTR	0.85	1.58	0.13	0.72	0.85	1.57	0.13	0.72
	SOC	0.92	1.64	0.17	0.85	0.93	1.64	0.17	0.86
	COG	0.89	2.32	0.12	0.79	0.88	2.29	0.12	0.77
	SLP	**0.68**	2.06	0.16	**0.46**	**0.69**	2.09	0.16	**0.47**
	AUD	**0.50**	1.41	0.17	**0.25**				
	REL	0.77	2.06	0.14	0.60	0.78	2.09	0.14	0.62
	QOL	0.83	2.40	0.14	0.69	0.82	2.35	0.14	0.68
	EMO	0.83	2.50	0.14	0.69	0.84	2.53	0.14	0.71

The values presented in bold have poor associations with their designated factor, all below the recommended cut-off < 0.40. β = Standardised parameter estimate; B = Unstandardised parameter estimate; SE = Standard Error; R_2 = R-squared.

Inspection of the squared factor loadings revealed that, although the second-order factor accounted for more than 60% of the variance in six of the factors, it accounted for less of the variance (46%) in the Sleep factor, and only 25% in the Auditory factor. The Auditory factor association to the second-order factor and the other seven factors was very weak and as a consequence it makes considerably less contribution to the second-order construct. This calls into question the usefulness of maintaining the Auditory factor within the global score. The evidence suggests that it measures an entirely different construct than the rest of the factors and could be diluting the overall score.

Inspection of the modification indices highlighted a high degree of mis-specified parameter estimates in the dataset, the majority of which appear to be associated with the Auditory factor. For example, error covariance (MIs >10) was observed between the Auditory factor and the Cognition, Sleep, Relaxation, QoL and Emotional factors (MI range: 10.1 – 37.7). In fact, the largest MI indicated substantially large error/uniqueness covariance between the Auditory and QoL factors (MI: 37.7; Stdx EPC: 0.55). From this, it could be assumed that the model is not accounting for a relationship between these two factors. However it seems unlikely given the observed error covariance with other factors listed above. The most common explanation for error covariance is similarity in wording or in the concepts being measured, or misinterpreting item responses. The first two again seem unlikely considering that the two factors are conceptually measuring very different aspects of tinnitus and that the similarity in wording of the questions is minimal. The former could be a potential source of error, both factors included items that could be misinterpreted or cause confusion. For example, the QoL factor included the only item with a slightly different response format than the others,

which could have caused confusion in responses, whilst the Auditory items were misinterpreted by some participants assuming that the questions related to their hearing alone rather than their hearing in relation to tinnitus with comment such as "I do not have problems with my hearing" written beside the responses. However, given that this error was not apparent at item level in this analysis or the first-factor analysis, and that error covariance is observed with the other factors, the most likely explanation is that the error covariance associated with the Auditory factor are due to the addition of the second-order factor and the weak association between these factors and the other factors. The large MIs may indicate the amount of unique variance associated with the Auditory factor and the fact that this factor may not provide any additional information than the QoL factor. Therefore, although freely estimating this error covariance would improve the model fit (the χ^2 would decrease by 37), conceptually it does not make sense to adjust. Furthermore, this provides additional rationale for removing the Auditory factor within the second-order structure. However, to be conservative and for completeness the other potential mis-specifications in the model should be assessed before re-specifying the model without the Auditory factor, to ensure that error covariance observed for the Auditory factor is not a product of or being inflated by any other mis-specification within the model.

The error covariance observed in the first-factor model between item 19 and item 20 was also apparent within this model. In addition to this, error variance was identified between item 23 "*How anxious or worried has your tinnitus made you feel?*" and item 24 "*How bothered or upset have you been because of your tinnitus?*" (MI: 11.29; Stdx EPC: 1.01) on the QoL factor, and between item 1 "*What percentage of your time awake were you consciously aware of your tinnitus?*" and

item 2 "*How strong or loud was your tinnitus?*" (MI: 16.46; STdx EPC: 0.38) on the Intrusiveness factor.

Inspection of these items indicated that the large error variance for the QoL items might be attributable to the similarity of the question wording. The wording for Sense of Control items on the other hand is not similar and the error cannot be attributed to this, but it could be attributed to the difference in response structure format for item 2 compared to the other items in the factor, for instance, item 1 and item 3 have a 0 to 100% response scale, whilst item 2 has a 0 to 10 response scale. Therefore, the model was re-specified to include these parameter estimates (freely estimated) and the model fit was reassessed. For clarity, this model will be referred as "re-specified TFI-25 model" (Figure 4.5).

Although there was some evidence of cross-loading present in the data, the largest again being the cross-loading between item 3 and Sense of Control (MI: 10.53; Stdx EPC: 0.45), this was no longer apparent following the model re-specification (re-specified TFI-25 model) for the error covariance. Interestingly, the only evidence of MIs (>10) remaining indicated that QoL items were cross-loading with other factors. In particular, item 20 appears to load on four other factors (Intrusiveness, Sense of Control, Relaxation and Emotional). Although, theoretically it is clear that a question about "*enjoyment of life"* could be linked to other concepts measured by other factors, it is impractical to cross-load an item with more than one factor. Furthermore, closer inspection of the EPC (unstandardized) estimates (>0.6) indicated that any change to the model would be minute and that the absolute loading value (Stdx EPC) for the cross-loading would be less than 0.25. Therefore, the model was not re-specified with any cross-loading (Figure 4.5).

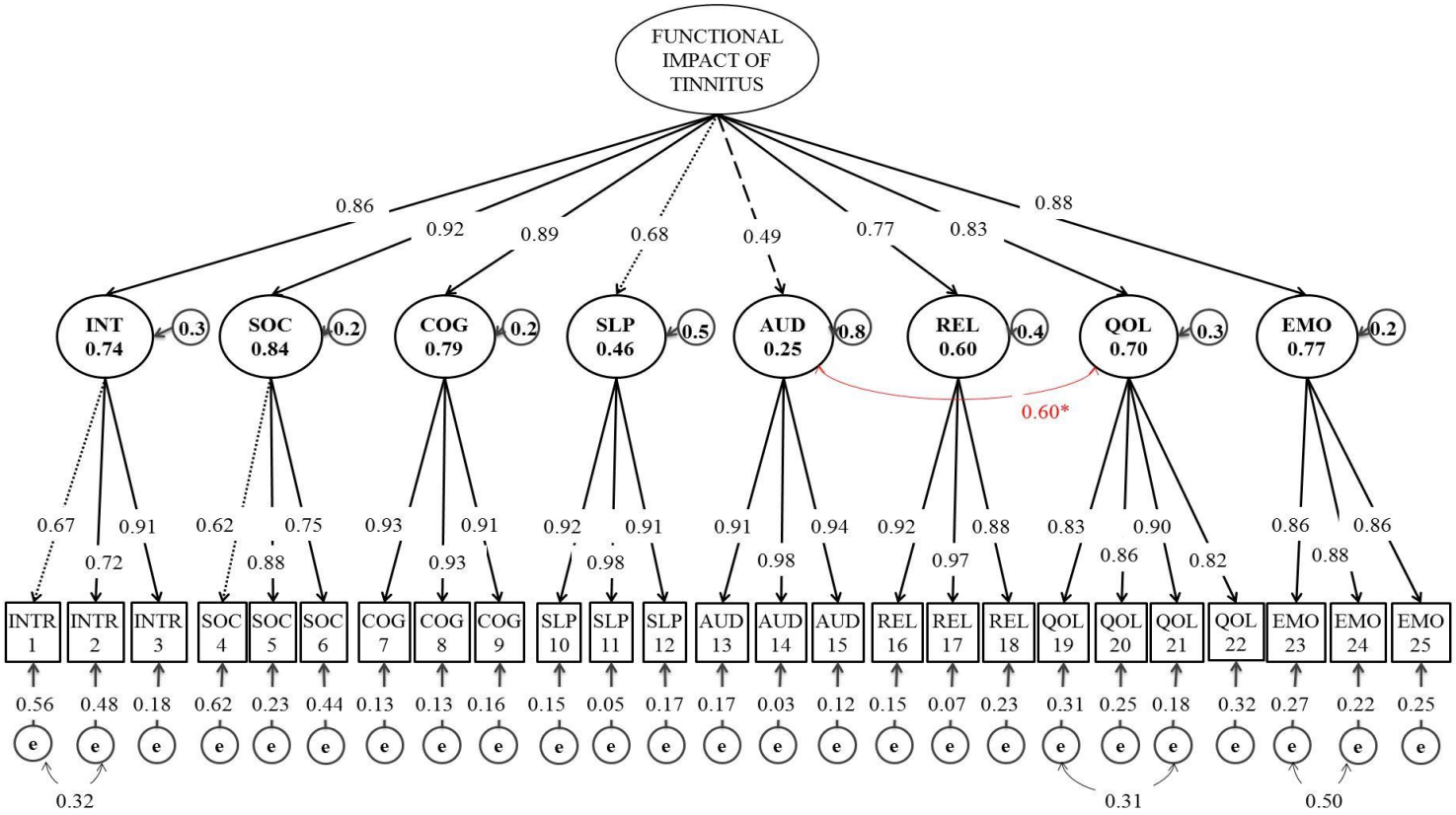

Figure 4.5. Illustrative diagram of the re-specified TFI-25 model including standardised parameter estimates and R-squared values.
Standardised parameter estimates indicate the strength of the association between the observed variables, first-order factors and the second-order factor. Red bidirectional arrow ( ) = large error covariance between factors. Solid black unidirectional arrow ( ) = a very strong association (> 0.70). Dotted unidirectional arrows ( ) = below desirable range but still acceptable (<0.70 >0.40). Dash line unidirectional arrows ( ) = poor associations. Bidirectional curved arrows ( ) = the association between the error variance (e).

4.4.4.3 Model fit for re-specified TFI-25 model

Adjusting for the error covariance between items only marginally improve model fit, but it still indicated a less than optimal to poor fit of the data to the model (Table 4.7). Furthermore, the large MIs associated with the Auditory factor were clearly still apparent and larger than previously observed (MI range: 10.7 to 44.4). Therefore, the next step was to investigate whether the Auditory factor was the source of the poor model fit. The model was re-specified without the Auditory factor. This new re-specified model is referred to as the TFI-22 model.

4.4.4.4 Model fit for TFI-22 model (Auditory factor removed)

Initially the TFI-22 model was specified using the original TFI-25 model specification, i.e. the error variance is again assumed to be uncorrelated since the model structure had changed. The model fit dramatically improved following the removal of the Auditory factor, all fit statistics indicated a reasonable fit of the data to model. The SRMR and the approximation fit indices, CFI and TLI were all within desirable criteria (Table 4.7), and although S-B χ^2 remained significant ($p< 0.001$), the χ^2/df ratio was now 1.92 so within the critical cut-off of < 2.0. Whilst the RMSEA only improved slightly (to 0.06), the 95% CI were now within the desired range, and given that the SRMR estimate was below 0.06, the RMSEA was taken to indicate reasonable model fit. Standardised parameter estimates and squared factor loadings were comparable to the original TFI-25 model (Table 4.9). Although, model fit was improved to an acceptable level, for completeness the modification indices were examined. Error covariance was once again observed, although the error observed between item 19 and item 20 in the TFI-25 model was no longer evident. The TFI-22 model was therefore re-specified to freely estimate the expected error correlations between item 1 and item 2, and item 23 and item 24 (Figure 4.6).

4.4.4.5 Model fit for re-specified TFI-22 model

Following this final re-specification, although S-B χ^2 still remained significant ($p< 0.001$), all other model fit statistics were slightly improved as expected and indicate an acceptable model fit. The re-specification of error covariance parameters did marginally reduced the factor loading estimate for those items associated with the error (although still above the recommended criteria), suggesting that the items loading estimates were previously inflated with unique variance. The standardised parameter estimates and R-squared values for the final TFI-22 model are given in Figure 4.6.

4.4.5. Validity of the TFI

Pearson's correlation coefficients between the TFI global and subscale scores and the THI global score from all four time intervals are displayed in Table 4.10. The results were as predicted. TFI global scores consistently showed strong positive correlations with the THI global scores ($r > 0.80$) across all four time intervals. Therefore, the TFI demonstrates acceptable convergent validity indicating that it measures a tinnitus construct that is similar to that measured by other multi-item tinnitus questionnaires. For the TFI subscales, weak ($r = 0.41$) to strong ($r = 0.86$) positive correlations were observed with the THI global scores (Table 4.10). The THI showed the strongest correlations with the Emotional and QoL subscales, as predicted, closely followed by the Cognition subscale and the weakest correlation with Auditory subscale, which estimate was notably smaller than any of the other subscales, possibly reflecting the fact that the THI does not fully address auditory problems. Therefore the TFI and THI both measure similar properties of tinnitus.

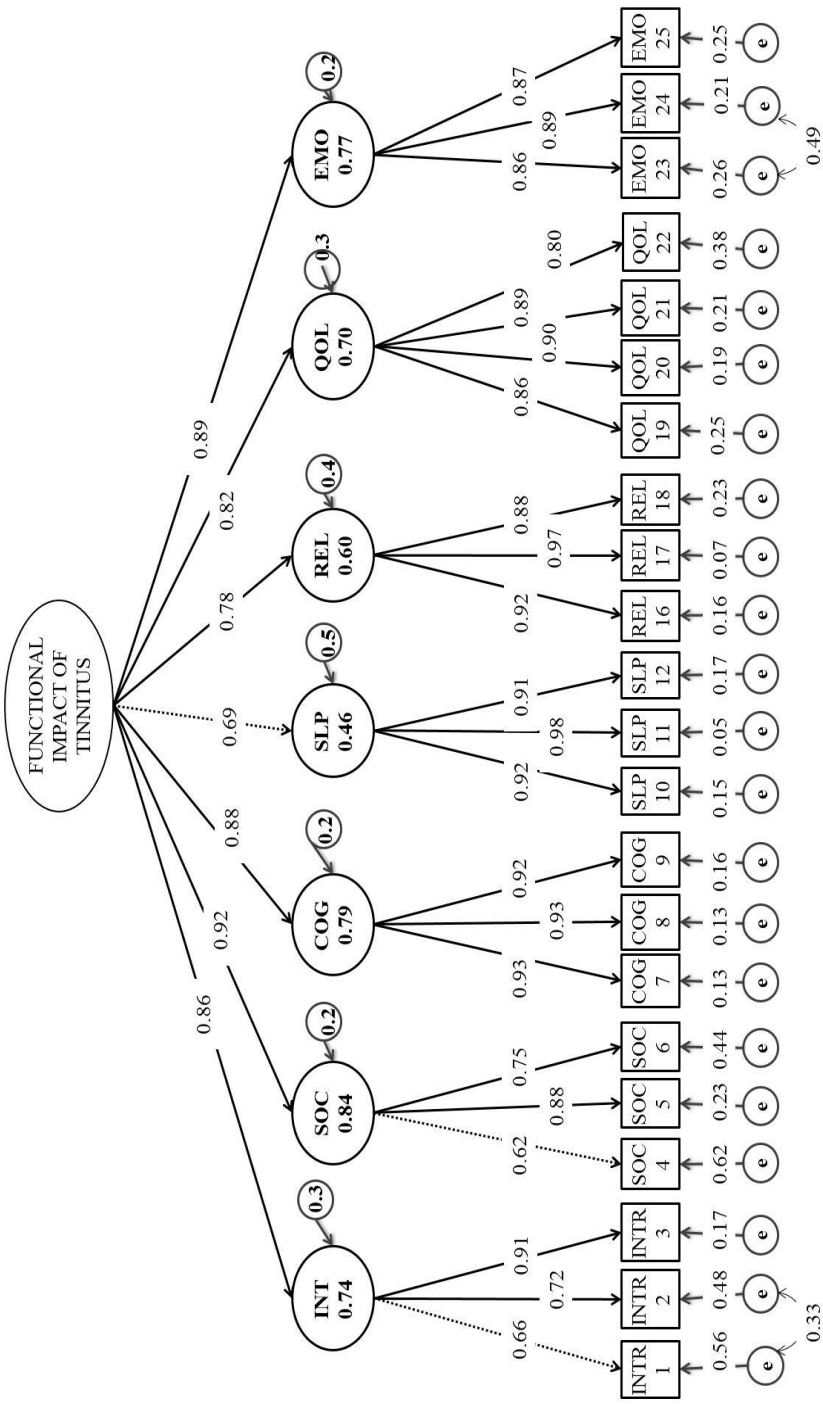

Figure 4.6. Illustrative diagram of the re-specified TFI-22 model including standardised parameter estimates and R-squared values.
Standardised parameter estimates indicate the strength of the association between the observed variables, first-order factors and the second-order factor. Solid black unidirectional arrow (⟶) = a very strong association (>0.70). Dotted unidirectional arrows (⇢) = below desirable range but still acceptable (<0.70 >0.40). Bidirectional curved arrows (↷) = the association between the error variance (e).

Table 4.10. Pearson's correlation coefficients between the TFI global and subscale scores and the THI global score for four time points

	Baseline		3 months		6 months		9 months	
Scale	*n*	THI	*n*	THI	*n*	THI	*n*	THI
Tinnitus Functional Index	*255*	0.85	*195*	0.83	*175*	0.86	*165*	0.85
Intrusiveness	*251*	0.62	*190*	0.65	*168*	0.70	*157*	0.78
Sense of Control	*251*	0.67	*195*	0.67	*173*	0.69	*164*	0.72
Cognitive	*255*	0.74	*192*	0.74	*175*	0.77	*165*	0.75
Sleep	*253*	0.61	*195*	0.66	*175*	0.66	*164*	0.69
Auditory	*254*	0.41	*193*	0.49	*175*	0.60	*165*	0.57
Relaxation	*254*	0.67	*194*	0.65	*174*	0.74	*163*	0.75
QoL	*255*	0.76	*195*	0.77	*175*	0.82	*165*	0.80
Emotional	*255*	0.79	*194*	0.86	*175*	0.84	*165*	0.86

4.4.6. Reliability of the TFI as a measure of tinnitus severity

4.4.6.1 Internal consistency

Alpha estimates for the global TFI and THI scores were all extremely high ($\alpha > 0.95$), exceeding the recommended criteria ($\alpha \leq 0.90$; Table 4.11). Although, the number of items (n=25) could have inflated these estimates, the 95% CI were also very high with a narrow range suggesting that there is overlap in content. Even though the TFI subscales would be expected to have lower alpha estimates because of the number of items alone, the estimates and CIs are similar to those observed for the global TFI. They were extremely high and in most cases were above the recommended criteria. The Intrusiveness and Sense of Control subscales were the only ones consistently within the recommended criteria, except for the 9 month data in which all of the estimates have inflated possibly due to the smaller sample size.

Table 4.11. Cronbach's alpha estimates (95% CI) for the THI global and TFI global and subscale scores over the four time points.

	QP1		QP2		QP3		QP4	
Scale	n (missing)	Cronbach's alpha (95% CI)	n (missing)	Cronbach's alpha (95% CI)	n (missing)	Cronbach's alpha (95% CI)	n (missing)	Cronbach's alpha (95% CI)
Tinnitus Functional Index[a]	255(0)	**0.96** (0.95 – 0.97)	196(0)	**0.97** (0.97 – 0.98)	175(0)	**0.98** (0.97 – 0.98)	165(0)	**0.98** (0.98 – 0.99)
Intrusiveness	251(4)	0.83 (0.79 – 0.86)	191(5)	0.89 (0.86 – 0.91)	163(12)	0.89 (0.85 – 0.91)	157(8)	0.92 (0.89 – 0.94)
Sense of Control	251(4)	0.79 (0.74 – 0.83)	196(0)	0.88 (0.85 – 0.91)	173(2)	0.90 (0.87 – 0.92)	164(1)	0.92 (0.90 – 0.94)
Cognition	255(0)	0.95 (0.94 – 0.96)	193(3)	**0.96** (0.95 – 0.97)	175(0)	**0.96** (0.94 – 0.97)	165(0)	**0.98** (0.97 – 0.98)
Sleep	253(2)	0.95 (0.94 – 0.96)	196(0)	**0.96** (0.95 – 0.97)	175(0)	**0.97** (0.96 – 0.97)	164(1)	**0.97** (0.96 – 0.98)
Auditory	254(1)	**0.96** (0.95 – 0.97)	194(2)	**0.97** (0.96 – 0.98)	175(0)	**0.97** (0.96 – 0.98)	165(0)	**0.98** (0.98 – 0.99)
Relaxation	254(1)	0.95 (0.93 – 0.96)	195(1)	**0.96** (0.94 – 0.97)	173(1)	**0.96** (0.95 – 0.97)	163(2)	**0.97** (0.96 – 0.98)
QoL	255(0)	0.92 (0.91 – 0.94)	196(0)	0.94 (0.93 – 0.95)	175(0)	0.95 (0.93 – 0.96)	165(0)	**0.96** (0.95 – 0.97)
Emotional	255(0)	0.92 (0.91 – 0.94)	195(1)	0.93 (0.92 – 0.95)	175(0)	0.95 (0.94 – 0.96)	165(0)	**0.96** (0.95 – 0.97)
Tinnitus Handicap Inventory[b]	255(0)	0.94 (0.93 – 0.95)	195(1)	0.94 (0.92 – 0.95)	175(0)	0.94 (0.93 – 0.96)	165(0)	0.95 (0.93 – 0.96)

Bold = exceeding recommended criteria (> 0.95)

Inspection of the inter-item correlations corroborates these findings. Inter-item correlations at baseline ranged 0.14 to 0.93 (Table 4.12), with the highest inter-item correlations between the items clearly identifying their designated subscales. Most notable were the extremely low correlations (r ~ 0.1) between the Auditory (13 – 15) and Sleep (10 - 12) subscale items, suggesting that these subscales are unrelated in content. The rest of the correlations between the items are generally low to moderate (within criteria), indicating the expected variability and overlap in content. Inter-item correlations for the follow-up data appeared susceptible to a reduction in the variability in the sample. As sample size decreased all correlation coefficients increased. These findings suggest that there is concerning overlap in content within the subscales and in turn the global TFI.

4.4.6.2 Test-retest reliability

Test-retest reliability and agreement estimates between the three time intervals are summarised in Table 4.13 for the "baseline comparison" group and Table 4.14 for the "3 month comparison" group. In both groups, the $ICC_{agreement}$ for the TFI global score was 0.87 (95%CI: 0.80 – 0.93), indicating excellent reliability of the TFI to distinguish people with tinnitus from each other. All subscale scores showed similarly high reliability with ICCs ranging 0.69 to 0.86 for both groups. However, the 95% CIs did indicate larger variability and lower reliability than accounted for by the ICC estimate. For example, the ICC 95%CI for the Sleep subscale (0.57 – 0.80) had a wide range with the lower value indicating that in a random sample the reliability could be markedly lower, below recommended guidelines for high reliability. The lower bound would indicate moderate evidence of reliability.

Table 4.12. Inter-item correlations for all 25 TFI items from baseline scores.

	QP1	Intrusiveness			Sense of control			Cognition			Sleep			Auditory			Relaxation			QoL				Emotional		
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
INT	Q1	1																								
	Q2	0.64	1																							
	Q3	0.62	0.65	1																						
SOC	Q4	0.38	0.36	0.43	1																					
	Q5	0.49	0.61	**0.68**	0.54	1																				
	Q6	0.46	0.56	0.61	0.49	0.65	1																			
COG	Q7	0.51	0.57	**0.66**	0.50	**0.68**	0.56	1																		
	Q8	0.39	0.52	0.60	0.45	0.64	0.50	**0.87**	1																	
	Q9	0.40	0.50	0.60	0.43	0.62	0.52	**0.84**	**0.87**	1																
SLP	Q10	0.29	0.42	0.55	0.32	0.49	0.49	0.56	0.50	0.52	1															
	Q11	0.33	0.45	0.55	0.33	0.50	0.48	0.57	0.49	0.52	**0.90**	1														
	Q12	0.35	0.45	0.54	0.31	0.53	0.43	0.55	0.47	0.51	**0.83**	**0.89**	1													
AUD	Q13	0.42	0.42	0.38	0.28	0.38	0.37	0.47	0.48	0.43	0.16	**0.15**	0.16	1												
	Q14	0.38	0.40	0.36	0.24	0.37	0.32	0.49	0.50	0.45	**0.14**	**0.14**	0.17	**0.90**	1											
	Q15	0.36	0.35	0.32	0.26	0.34	0.28	0.45	0.47	0.44	**0.15**	**0.15**	0.17	**0.86**	**0.93**	1										
REL	Q16	0.31	0.45	0.55	0.40	0.57	0.54	0.61	0.54	0.57	0.63	0.61	0.56	0.24	0.19	0.19	1									
	Q17	0.33	0.46	0.58	0.46	0.59	0.52	0.65	0.60	0.61	0.65	0.62	0.57	0.24	0.21	0.23	**0.88**	1								
	Q18	0.36	0.42	0.50	0.43	0.49	0.46	0.52	0.46	0.48	0.52	0.50	0.44	0.26	0.20	0.23	**0.82**	**0.84**	1							
QOL	Q19	0.34	0.43	0.47	0.37	0.52	0.39	0.60	0.63	0.60	0.31	0.34	0.36	0.54	0.56	0.59	0.41	0.47	0.38	1						
	Q20	0.35	0.45	0.58	0.42	0.63	0.47	0.65	**0.66**	**0.66**	0.44	0.48	0.48	0.40	0.44	0.43	0.53	0.62	0.50	**0.80**	1					
	Q21	0.36	0.40	0.50	0.35	0.54	0.45	0.63	0.64	0.63	0.36	0.41	0.39	0.54	0.58	0.58	0.48	0.54	0.45	**0.77**	**0.77**	1				
	Q22	0.31	0.36	0.44	0.37	0.51	0.37	**0.67**	**0.68**	0.65	0.41	0.42	0.42	0.57	0.62	0.61	0.42	0.48	0.34	0.67	**0.67**	**0.76**	1			
EMO	Q23	0.35	0.39	0.54	0.43	0.61	0.52	0.60	0.58	0.57	0.48	0.49	0.49	0.23	0.25	0.25	0.55	0.60	0.51	0.53	**0.66**	0.61	0.55	1		
	Q24	0.36	0.42	0.59	0.47	**0.69**	0.57	0.64	0.62	0.61	0.47	0.47	0.47	0.25	0.24	0.22	0.57	0.62	0.50	0.50	**0.68**	0.57	0.56	**0.87**	1	
	Q25	0.35	0.47	0.63	0.38	0.63	0.52	0.59	0.59	0.60	0.51	0.53	0.53	0.27	0.27	0.21	0.54	0.58	0.45	0.46	**0.67**	0.56	0.56	**0.75**	**0.77**	1

Black bold = high inter-item correlations. Red bold = poor inter-relatedness. Blue = within criteria (0.15 – 0.55)

Table 4.13. Reliability of Tinnitus Functional Index (TFI) scores from baseline comparisons: Standard Error of Measurement (SEM), Intra-class correlations (ICC), Smallest Detectable Change (SDC), Limits of Agreement (LoA) between three administrations.

Baseline comparison		Mean (±SD)			Difference			Reliability	SEM		Agreement				
Scale	n	Baseline (T0)	Retest (T1)	Retest (T2)	Mean diff	SE	SD diff	ICC (95%CI)	Con	Agree	SDC	LoA	LoA Lower limit (95% CI)	LoA Upper limit (95% CI)	%
Tinnitus Functional Index	55	50.8 (±25.1)	45.9 (±22.8)	44.9 (±23.1)	-5.4	1.0	7.2	0.87 (0.80 - 0.93)	5.1	7.4	20.6	14.2	-19.64 (-23.2 to -16.1)	8.8 (5.2 to 12.3)	76
Intrusiveness	47	64.0 (±24.3)	55.6 (±23.2)	54.7 (±23.3)	-9.8	1.9	13.2	0.79 (0.63 - 0.88)	9.3	10.9	30.1	25.8	-35.6 (-42.3 to -28.9)	19.7 (12.3 to 27.1)	89
Sense of Control	48	61.9 (±24.3)	56.5 (±24.2)	55.0 (±24.0)	-6.2	1.6	10.7	0.79 (0.69 - 0.87)	7.6	11.0	30.6	21.0	-27.2 (32.6 to -21.8)	14.8 (9.4 to 20.2)	79
Cognitive	50	40.9 (±28.4)	39.3 (±27.2)	38.0 (±26.0)	-2.2	1.9	13.5	0.86 (0.79 - 0.91)	9.6	10.1	27.9	26.5	-28.7 (-35.5 to -22.0)	24.3 (17.6 to 31.0)	72
Sleep	48	52.6 (±32.0)	46.4 (±29.2)	47.1 (±29.4)	-4.8	1.9	13.3	0.69 (0.57 - 0.80)	9.4	14.1	39.0	25.9	-30.8 (-37.2 to -24.3)	21.1 (14.6 to 27.6)	73
Auditory	50	47.7 (±30.1)	47.5 (±28.8)	44.4 (±28.2)	-1.7	2.3	16.2	0.83 (0.74 - 0.89)	11.5	12.1	33.5	31.8	-33.5 (-41.4 to -25.6)	30.1 (22.2 to 38.0)	88
Relaxation	50	62.0 (±29.3)	55.9 (±26.4)	53.5 (±27.4)	-7.4	2.0	14.4	0.75 (0.63 - 0.84)	10.2	13.6	37.8	28.3	-35.7 (-42.7 to -28.7)	20.8 (13.8 to 27.8)	72
QoL	49	38.6 (±32.4)	34.7 (±28.9)	33.3 (±29.0)	-4.6	1.6	11.3	0.79 (0.70 - 0.87)	8.0	11.8	32.6	22.2	-26.7 (-32.2 to -21.2)	17.6 (12.1 to 23.1)	76
Emotional	50	42.8 (±31.5)	35.7 (±29.7)	38.4 (±30.2)	-5.7	2.1	14.6	0.82 (0.75 - 0.88)	10.3	12.3	34.2	28.6	-34.4 (-41.5 to -27.2)	22.9 (15.7 to 30.0)	84

SE = Standard Error. Mean diff = mean difference between the three administrations scores. SD diff = Standard deviation of the difference. ICC = Intra-class correlations. SEM con= Standard Error of Measurement for consistency. SEM agree= Standard Error of Measurement for agreement. SDC = Smallest Detectable Change. LoA = Limits of Agreement. CI = Confidence Interval. % = percentage of agreement.

Table 4.14. Reliability of Tinnitus Functional Index (TFI) scores from 3 month comparisons: Standard Error of Measurement (SEM), Intra-class correlations (ICC), Smallest Detectable Change (SDC), Limits of Agreement (LoA) between three administrations.

3 month comparison		Mean (±SD)			Difference			Reliability	SEM		Agreement				
Scale	n	Baseline (T0)	Retest (T1)	Retest (T2)	Mean diff	SE	SD diff	ICC (95%CI)	Con	Agree	SDC	LoA	LoA Lower limit (95% CI)	LoA Upper limit (95% CI)	%
Tinnitus Functional Index	55	49.3 (±25.4)	44.2 (±23.3)	42.7 (±23.3)	-3.1	1.4	10.3	0.87 (0.80 - 0.93)	7.3	8.0	22.2	20.2	-23.2 (-28.1 to -18.4)	17.1 (12.3 to 21.9)	93
Intrusiveness	52	63.0 (±23.8)	54.0 (±22.8)	52.3 (±23.8)	-5.3	3.2	22.7	0.80 (0.63 - 0.89)	16.1	10.7	29.6	44.5	-49.7 (-60.7 to -38.8)	39.2 (28.3 to 50.2)	100
Sense of Control	55	61.7 (±24.8)	57.0 (±24.3)	53.9 (±23.9)	-3.3	1.2	9.2	0.71 (0.58 - 0.81)	6.5	11.6	32.1	18.1	-21.4 (-25.7 to -17.1)	14.7 (10.4 to 19.1)	66
Cognitive	55	40.5 (±28.5)	36.9 (±27.5)	35.9 (±26.1)	-2.3	1.9	14.1	0.86 (0.79 - 0.91)	10.0	10.3	28.4	27.7	-30.0 (-36.6 to -23.4)	25.4 (18.8 to 32.0)	93
Sleep	54	48.4 (±34.5)	44.9 (±30.5)	43.4 (±31.2)	-2.0	2.6	18.9	0.77 (0.67 - 0.85)	13.4	14.0	41.2	37.0	-39.1 (-48.0 to -30.1)	35.0 (26.0 to 43.9)	86
Auditory	56	44.9 (±30.4)	434.7 (±28.3)	42.1 (±27.6)	-1.5	1.9	14.7	0.82 (0.74 - 0.88)	10.4	12.1	33.5	28.8	-30.2 (-37.1 to -23.4)	27.2 (20.4 to 34.1)	90
Relaxation	56	59.3 (±30.0)	54.1 (±26.6)	50.7 (±28.4)	-4.3	2.2	16.5	0.78 (0.68 - 0.86)	11.7	12.7	36.7	32.4	-36.7 (-44.4 to -28.9)	28.1 (20.4 to 35.8)	86
QoL	55	36.7 (±31.9)	33.1 (±29.2)	31.6 (±28.9)	-2.5	1.7	12.3	0.83 (0.75 - 0.89)	8.7	11.1	30.8	24.0	-26.6 (-32.3 to -20.8)	21.5 (15.7 to 27.3)	80
Emotional	53	39.6 (±31.1)	34.5 (±30.7)	35.2 (±30.2)	-2.2	3.5	25.8	0.88 (0.81 - 0.92)	18.3	10.8	29.9	50.6	-52.8 (-65.2 to -40.5)	48.3 (36.0 to 60.7)	100

SE = Standard Error. Mean diff = mean difference between the three administrations scores. SD diff = Standard deviation of the difference. ICC = Intra-class correlations. SEM con= Standard Error of Measurement for consistency. SEM agree= Standard Error of Measurement for agreement. SDC = Smallest Detectable Change. LoA = Limits of Agreement. CI = Confidence Interval. % = percentage of agreement.

4.4.6.3 Test-retest agreement

For the "baseline comparison" group, the LoA and SDC estimates for the TFI global were comparable with one another. The LoA score was 14.2 (± Mean diff of -5.4) and the SDC score was 20.6.

Both indicate that a reasonably large change score is required to detect the true change that occurs in the scores either with worsening or improvements of tinnitus (Table 4.13). Slightly larger estimates were observed for the "3 month comparison" data, with both the LoA (20.2±-3.1) and SDC (22.2). At 3 months, the findings suggest that a change in TFI global of 23 would be required to detect a true change in scores (Table 4.14). Although slightly larger, this estimate does seem to account for more of the observed variability in the repeated measure change score than the estimates from the "baseline comparison" data. In particular, only four change scores were found to be outside the defined LoA (for one participant both change scores between the time intervals were outside). Therefore 93% agreement was observed, only just below the recommended value of 95% agreement (Figure 4.7). In the "baseline comparison" data, however, 12 change scores were found outside the defined LoA (again both difference scores for one participant were outside), and therefore only 76% of change scores were within the limits. This is considerably lower than recommended (Figure 4.8). In this case, the LoA and SDC estimates from the "3 month comparison" data are taken as the more accurate representation of the variability seen in the data for participants that are self-defined "unchanged" and therefore a change in score of 23 is considered an indication of a true change in scores.

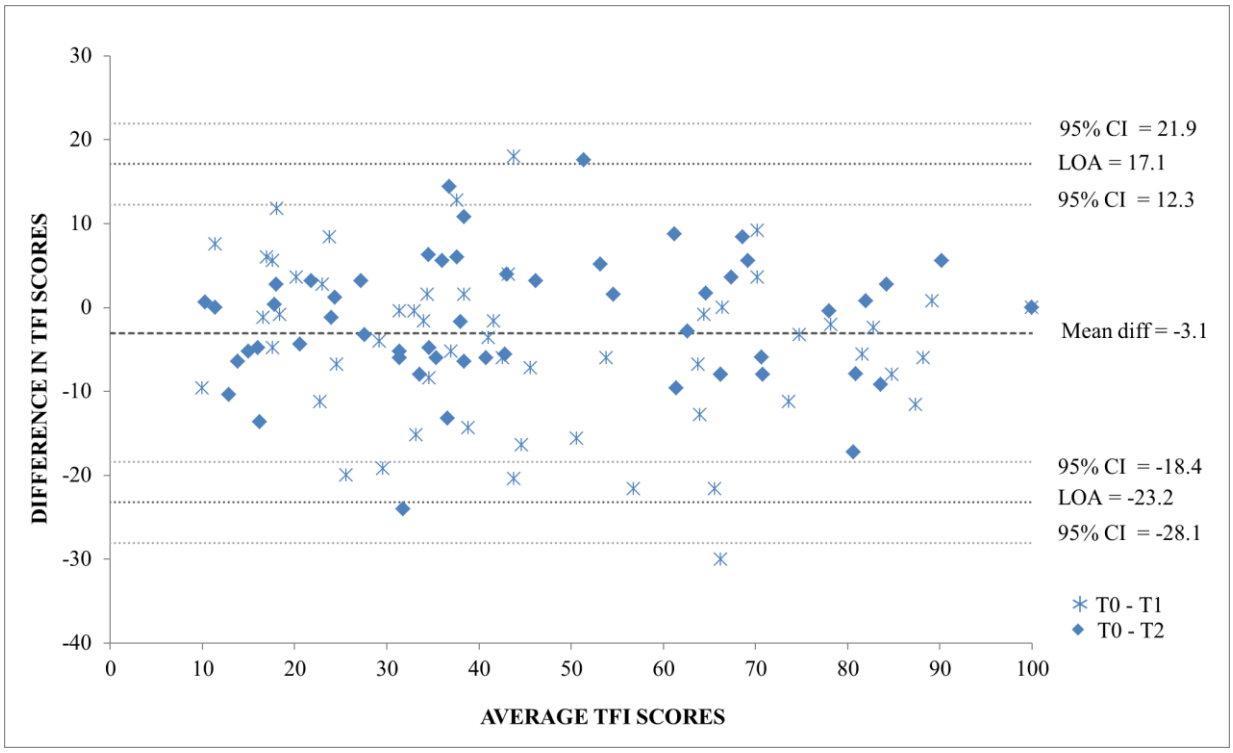

Figure 4.7. Bland-Altman plot of test-retest agreement for repeated measures of the TFI global scores for self-defined "stable" participants from 3 month comparison data.

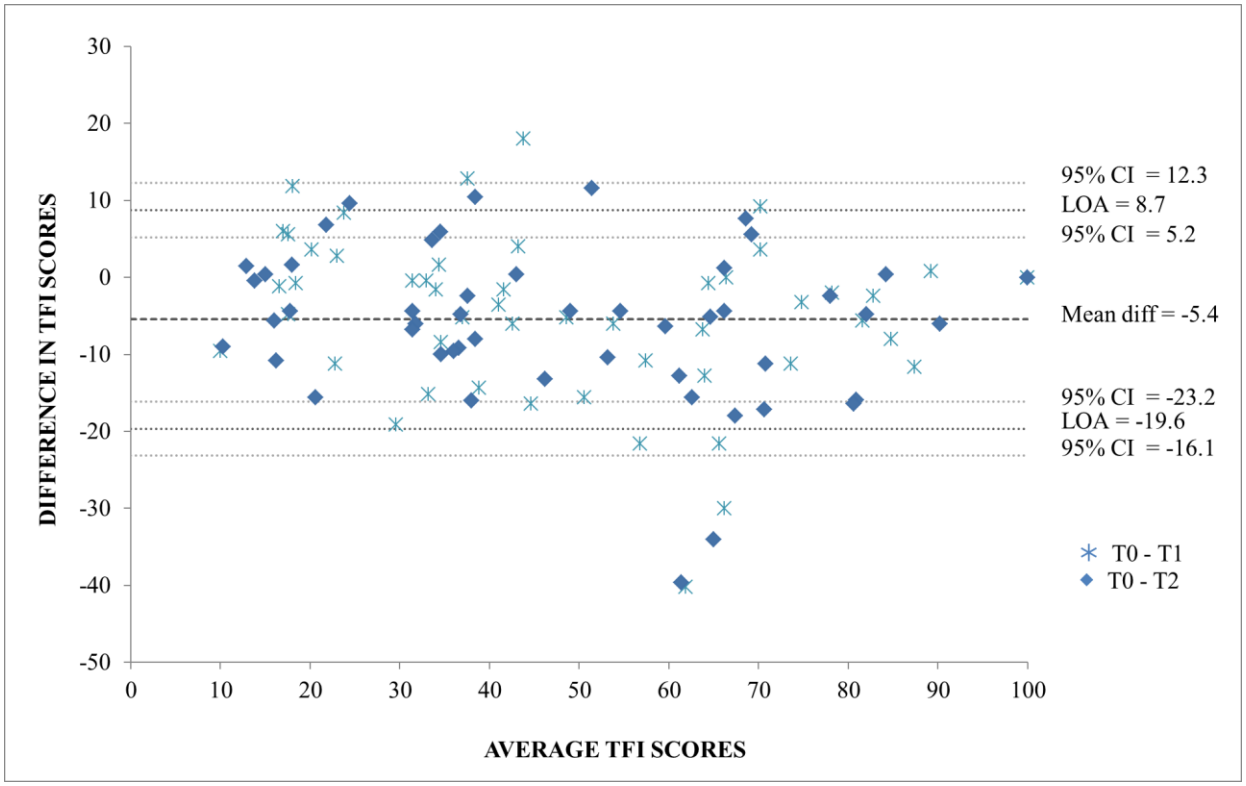

Figure 4.8. Bland-Altman plot of test-retest agreement for repeated measures of the TFI global scores for self-defined "stable" participants from baseline comparison data.

According to Terwee et al. (2007), this SDC score is appropriate for individual assessment, the change score should be reduced by a factor of √n to account for the reduced measurement error in group assessment. Therefore in this case the SDC for group assessment would only be 3, which unexpectedly low and given that the SEM estimates are both more than 7 points, these might provide a better indication of measurement error for group assessment.

In terms of the TFI subscales, the limits of agreement and smallest detectable change estimates were in general considerably larger than the estimates for the global score and there were inconsistencies between the estimates for some of the subscales in both groups (Tables 4.13 – 4.14). The LoA for the "baseline comparison" data ranged from 27 to 36 points, with the largest variability observed in the Intrusiveness and Relaxation subscales. The SDC scores were largely comparable with the LoA estimates in five of the subscales, but for the Sleep and QoL subscales there was a notable difference in the estimates, with the SDC identifying a larger estimate of true change. For example, the LoA for the Sleep subscale indicate that a change of at least 30 points would be required to detect a "true change", whilst the SDC score suggests that a change of 39 points would be required. A possible reason for these differences is that the SDC is calculated using the SEM_{agree} and therefore takes into consideration a higher proportion of the variance in the data which can lead to higher estimates of change in some cases. This difference is usually reflected in the two estimates for SEM, the SEM_{con} value (estimated using the SD_{diff}) is expected to be slightly lower than the value for the SEM_{agree}

The "3 month comparison" data presented in Table 4.14 again show that the LoA and SDC estimates were slightly larger at 9 months than at baseline. In this case, the LoA ranged from 21 (Sense of control subscale) to 52 points for the

Emotional subscale. This large estimate was unexpected and is not comparable to the SDC. In fact, the LoA estimates for the Intrusiveness (50) and Emotional subscales (52) were considerably larger than the SDC, which would suggest that only a change in score of 30 is required to detect real change (similar to the estimates in the baseline data). Other than any unexpected variability in the change scores that was not identified by the SEM_{agree}, which was calculated using the total scores for each time interval, there is no clear reason why these estimates are inconsistent or so large. Although this inconsistency was not apparent for the Emotional subscale in the "baseline comparison" data and the SDC estimate is very similar, because this inconsistency cannot be explained for these subscales, we cannot make any strong conclusions about the level of agreement.

For the most part, the SDC scores were reasonably consistent with the LoA estimates. Ignoring the problematic subscales, the estimates in the "3 month comparison" data do once again account for slightly more of the observed variability in the repeated measures, with slightly higher percentages of agreement, than the estimates from the "baseline comparison" data (Tables 4.13 – 4.14). However, 95% agreement was not observed in any of the subscales, only the Cognition subscale agreement was above 90%, and therefore these subscales appear to be more susceptible to the variability over long periods of time than expected. Perhaps shorter timeframes would produce more consistent results (examined in Chapter 5).

4.4.7. Responsiveness of the TFI to detect changes

4.4.7.1 The ability of the TFI items to detect changes in responses

Response frequency distributions for each item on the TFI were examined for floor and ceiling effects (Figure 4.9 and Table 4.15). Eleven out of 25 items failed to meet the *a priori* definition of non-significant floor or ceiling effects (i.e. observed in no

more than 15% of respondents on the 11-point scale). More specifically, ceiling effects were observed in five items from the Intrusiveness (item 1), Sense of Control (items 4, 6), Sleep (item10) and Relaxation (item18) subscales, with responses of '10' being observed for between 17% and 29% of the population. Of these, only Item 4 consistently showed ceiling effect across time intervals. This suggests that this item is not sensitive to change.

Extreme floor effects were observed in six items from three subscales. Two subscales had multiple items with floor effects. The lowest response '0' was observed for between 18 and 26% of participants in two items from the Auditory subscale (items 14, 15), three items from the QoL subscale (items 19 – 22) and one item from Emotional subscale (item 25), with the largest floor effects (> 24%) observed in the QoL items. Furthermore, these floor effects were clearly apparent over all of the datasets and were consistently high. Consequently, these subscales are limited in their ability to detect improvements of tinnitus and changes in scores which in turn reduces the chances of the TFI being responsive to treatment-related changes.

4.4.7.2 The ability of the TFI to detect changes in scores

The sample sizes in two of the seven global rating of change categories, the "much improved" and "much worse" categories, were not sufficient to make meaningful comparisons (Table 4.5). Therefore the responses in these categories were amalgamated with those in the "moderate improved" and moderate worse" categories so that five global ratings of change groups remained. The mean changes for the global TFI within the five ratings of change groups from 3, 6 and 9 months compared to baseline ("baseline comparison" data) and compared to 3 months ago ("3 month comparison" data) are presented in Table 4.16.

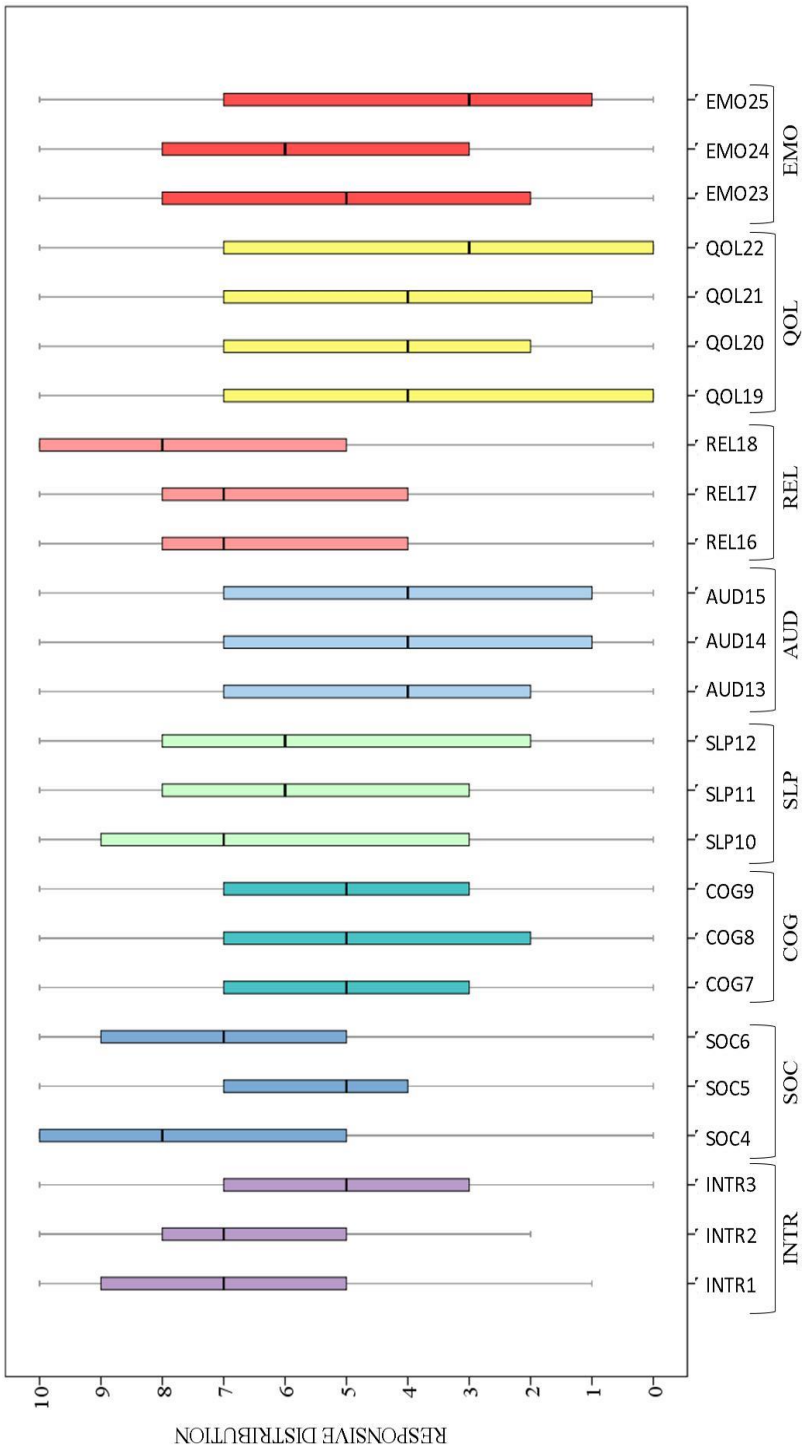

Figure 4.9. Response frequency distributions for each TFI item within their subscales allowing for examination of floor and ceiling effects. Ceiling effects are evident from the position of the upper quartile and medium on the upper end of the scale, i.e. on response options 9 and 10. The floor effects are evident by the position of the first quartile and medium on the lower end of the scale, i.e. on response options 0 and 1. Colours represent the items associated with each subscale.

Table 4.15. Percentage of responses in each response category option for the TFI items.

N255 items	Percentage of responses for items on the TFI											% Missing	Mean ±SD
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		
INT1	0.0	2.8	4.0	7.1	7.5	8.3	5.5	15.8	16.6	9.9	**22.5**	0.8	6.94 ± 2.61
INT2	0.0	0.0	1.6	5.9	9.4	10.6	11.8	19.7	20.5	9.4	11.0	0.4	6.80±2.07
INT3	2.0	12.6	9.1	12.6	9.5	13.0	9.5	9.1	8.3	5.5	8.7	0.8	4.96±2.87
SOC4	2.4	3.2	2.8	6.8	4.0	10.8	10.4	8.8	13.5	8.4	**29.1**	1.6	6.97±2.85
SOC5	1.6	3.5	6.3	10.2	8.2	20.4	13.3	12.2	12.9	5.5	5.9	0.0	5.58±2.41
SOC6	0.8	2.4	2.4	5.9	5.5	13.8	9.8	15.7	15.0	9.8	**18.9**	0.4	6.82±2.48
COG7	6.7	6.3	8.2	11.8	9.8	9.8	9.0	17.3	11.8	3.5	5.9	0.0	5.06±2.81
COG8	12.5	7.5	7.8	12.9	7.1	11.8	9.4	12.2	12.9	2.7	3.1	0.0	4.50±2.89
COG9	6.7	8.6	9.8	15.3	8.2	12.5	9.4	12.9	10.2	3.1	3.1	0.0	4.58±2.70
SLP10	9.4	5.5	5.9	6.7	6.7	7.1	4.7	12.2	15.3	9.0	**17.6**	0.0	5.93±3.32
SLP11	10.2	6.7	7.5	7.5	5.5	9.4	4.3	14.2	13.8	7.1	13.8	0.4	5.50±3.30
SLP12	13.0	5.5	9.8	6.3	5.9	9.1	4.3	14.2	11.4	5.9	14.6	0.4	5.28±3.39
AUD13	14.9	8.6	8.6	9.8	9.8	7.8	9.0	11.0	11.0	3.5	5.9	0.0	4.43±3.11
AUD14	**20.4**	9.8	9.0	7.1	9.4	7.5	10.6	10.6	7.5	5.1	3.1	0.0	3.98±3.11
AUD15	**18.1**	11.0	6.3	8.7	7.9	7.9	7.1	10.6	7.9	8.7	5.9	0.4	4.37±3.32
REL16	4.7	3.1	7.1	8.3	3.9	8.7	9.1	14.2	17.3	9.1	14.6	0.4	6.20±2.93
REL17	4.3	3.5	9.4	7.5	4.3	8.6	10.2	13.7	16.9	8.6	12.9	0.0	6.04±2.93
REL18	3.1	3.1	5.1	4.7	2.7	8.2	7.5	9.4	14.5	13.7	**27.8**	0.0	7.08±2.93
QOL19	**25.1**	8.8	8.0	7.6	4.0	11.6	6.8	9.6	8.4	4.4	6.0	1.6	3.95±3.33
QOL20	14.9	9.8	9.0	10.6	7.8	9.0	7.1	8.2	9.0	5.9	8.6	0.0	4.47±3.27
QOL21	**23.9**	8.2	10.2	7.8	7.1	11.8	3.9	8.6	8.6	5.5	4.3	0.0	3.85±3.23
QOL22	**25.6**	11.0	9.4	5.9	7.5	7.5	7.5	8.3	7.5	4.3	5.5	0.4	3.72±3.30
EMO23	9.4	9.0	10.2	11.4	3.9	7.5	7.1	11.0	11.8	6.7	12.2	0.0	5.11±3.30
EMO24	3.5	6.7	10.2	8.6	7.1	12.5	6.7	12.5	9.4	9.0	13.7	0.0	5.65±3.03
EMO25	**24.7**	7.8	9.4	8.6	4.7	9.4	5.5	6.7	9.4	4.7	9.0	0.0	4.06±3.47

% missing data and mean ±SD (Standard Deviation) reported for each item. Bold = exceeding criteria (ratings of either 0 points or 10 point being observed in <15% of respondents).

Table 4.16. Descriptive statistics for TFI global scores classified into global rating of changes categories for 3, 6 and 9 months.

TFI		3 months				6 months				9 months			
	Perception of change	n	Mean (SD)	Range	Diff	n	Mean (SD)	Range	Diff	n	Mean (SD)	Range	Diff
Compared to baseline	Much to moderately improved	30	-22.1 (20.3)	-63.6 8.40	-18	34	-26.0 (17.2)	-56.8 - 4.0	-18.9	35	-25.3 (19.1)	-61.6 - 16.4	-16.2
	Slightly improved	39	-12.8 (10.2)	-40.8 - 6.40	-8.7	33	-12.7 (11.8)	-36.0 - 15.2	-5.6	24	-12.0 (16.7)	-52.1 - 11.6	-2.9
	No change	101	-4.1 (12.0)	-46.4 - 24.8		67	-7.1 (13.5)	-55.2 - 20.4		48	-9.1 (12.7)	-34.5 - 18.4	
	Slightly worse	23	2.5 (12.1)	-18.8 - 22.8	6.6	30	1.4 (14.9)	-25.6 - 31.7	8.5	33	-0.7 (16.1)	-28.8 - 38.0	8.4
	Moderately to much worse	3	-2.5 (11.4)	-13.5 - 9.17	1.6	11	7.2 (14.2)	-17.2 - 28.0	14.3	25	7.6 (16.5)	-28.4 - 48.5	16.7
Compared to 3 months ago	Much to moderately improved	–	–	–	–	28	-6.1 (11.3)	-26.8 - 17.2	-4.3	22	-5.3 (15.2)	-23.7 - 51.2	-3.8
	Slightly improved	–	–	–	–	28	-4.2 (15.0)	-52.8 - 22.0	-2.4	26	2.2 (14.6)	-25.2 - 42.4	3.7
	No change	–	–	–	–	82	-1.8 (10.7)	-49.6 - 35.0		64	-1.5 (9.5)	-24.8 - 34.4	
	Slightly worse	–	–	–	–	28	6.7 (12.1)	-14.0 - 32.4	8.5	35	5.5 (12.5)	-28.8 - 38.8	7.0
	Moderately to much worse	–	–	–	–	9	12.7 (9.9)	-2.8 - 26.0	14.5	17	7.5 (12.3)	-7.2 - 37.2	9.0

SD = Standard deviation. Diff = the difference between mean scores in the no change groups and the improved and worsened groups (much-to-moderately improved/worse and slightly improved/worse).

In general, the pattern in the mean change scores between the different ratings of change groups was as expected across the nine months for both datasets. Although, for "3 month comparison" data, there was very little difference in the mean scores between the global ratings of change groups. The mean change scores decreased with lower self-reported improvements and increased with greater self-reported deterioration (worsening). For example, the mean change scores for 'much-to-moderately improved' group were lower than that observed in the 'slightly improved' group, which again were lower than the 'no change' and 'worse' rating groups. There was one exception to this pattern for the "baseline comparison" data, mean change scores observed in the 'moderately-to-much worse' rating group at 3 months was lower than expected, but there were insufficient responses within this category at 3 months and as a consequence the mean change within this category should be ignored.

The difference in the change scores were as predicted, with larger differences for the 'much-to-moderately' improved and 'worsened' groups from 'no change' than the differences for the 'slightly improved' and 'worsened' comparison to 'no change'. These differences did however slightly vary over time for the "3 month comparison" data. For example, the change scores at 9 months were smaller than those observed at 3 months for the 'improved' groups (Table 4.16).

In terms of the TFI subscales, mean change scores from the "baseline comparison" data showed the expected pattern between the ratings of change groups (Table 4.17), with lower mean change scores in the improved group than the no change across all time intervals.

Table 4.17. Mean (SD) and mean difference for TFI subscale scores in 'improved', 'no change' and 'worsened' categories for baseline comparison data at 3, 6 and 9 months.

		3 months			6 months			9 months		
		n	Mean (SD)	Diff	n	Mean (SD)	Diff	n	Mean (SD)	Diff
THI	Improved	69	-12.3 (15.0)	-9.1	67	-15.2 (16.5)	-9.5	59	-16.0 (16.8)	-6.7
	No change	101	-3.2 (12.0)		67	-5.7 (13.6)		48	-9.3 (16.2)	
	Worsened	26	-0.5 (12.1)	2.7	41	0.7 (15.3)	6.4	58	1.3 (15.9)	10.6
TFI	Improved	69	-16.8 (15.9)	-12.7	67	-19.4 (16.1)	-12.3	59	-19.9 (19.2)	-10.8
	No change	101	-4.1 (12.0)		67	-7.1 (13.5)		48	-9.1 (12.7)	
	Worsened	26	1.9 (11.8)	6	41	2.9 (14.7)	10	58	2.9 (16.7)	12
INT	Improved	69	-20.1 (21.7)	-13.1	67	-25.7 (22.4)	-13.8	59	-27.9 (20.4)	-9.7
	No change	101	-7.0 (22.4)		67	-11.9 (18.1)		48	-18.2 (22.6)	
	Worsened	26	-5.0 (22.8)	2.0	41	3.7 (21.3)	15.5	58	-2.5 (23.2)	15.7
SOC	Improved	69	-20.7 (25.1)	-16.6	67	-25.9 (22.2)	-16.1	59	-23.1 (25.5)	-13.0
	No change	101	-4.2 (18.3)		67	-9.8 (23.7)		48	-10.1 (20.2)	
	Worsened	26	3.7 (22.8)	7.9	41	0.7 (23.2)	10.5	58	-0.6 (22.0)	9.6
COG	Improved	69	-12.2 (21.8)	-7.4	67	-15.5 (20.4)	-10.7	59	-19.9 (26.4)	-13.2
	No change	101	-4.8 (21.6)		67	-4.8 (19.5)		48	-6.7 (15.4)	
	Worsened	26	2.8 (20.2)	7.6	41	2.8 (20.0)	7.7	58	3.6 (24.5)	10.3
SLP	Improved	69	-17.3 (29.3)	-12.5	67	-22.3 (30.1)	-15.6	59	-24.7 (28.8)	-12.4
	No change	101	-4.8 (22.1)		67	-6.7 (25.7)		48	-12.4 (26.1)	
	Worsened	26	0.1 (18.0)	4.9	41	3.2 (18.7)	9.8	58	2.5 (24.6)	14.9
AUD	Improved	69	-10.1 (21.3)	-13.1	67	-8.8 (24.7)	-5.4	59	-4.1 (28.5)	-1.6
	No change	101	2.9 (23.7)		67	-3.4 (18.8)		48	-2.5 (17.5)	
	Worsened	26	2.2 (25.4)	-0.8	41	12.2 (23.4)	15.6	58	14.1 (26.9)	16.6
REL	Improved	69	-22.8 (25.2)	-15.7	67	-24.4 (28.9)	-14.5	59	-26.2 (31.0)	-13.1
	No change	101	-7.1 (23.5)		67	-9.9 (21.7)		48	-13.1 (24.8)	
	Worsened	26	2.4 (22.7)	9.5	41	-1.5 (24.2)	8.4	58	-0.2 (24.5)	12.8
QOL	Improved	69	-15.5 (19.3)	-12.0	67	-14.3 (21.6)	-8.2	59	-16.7 (23.7)	-12.3
	No change	101	-3.5 (19.3)		67	-6.1 (20.6)		48	-4.4 (17.6)	
	Worsened	26	6.0 (21.3)	9.5	41	4.9 (24.0)	11.0	58	6.0 (21.2)	10.4
EMO	Improved	69	-17.1 (22.1)	-12.1	67	-22.6 (24.9)	-17.3	59	-19.9 (23.9)	-10.2
	No change	101	-5.1 (17.6)		67	-5.3 (18.8)		48	-9.7 (22.5)	
	Worsened	26	1.7 (20.1)	6.7	41	0.1 (24.3)	5.4	58	-1.3 (23.2)	8.4

SD = Standard deviation. Diff = the difference between mean scores in the 'no change' group and the improved and worsened groups.

The magnitude of change (i.e. the difference) between ‘improved’ and ‘no change’ was also reasonably consistent over time for all the subscales, indicating that a change in scores from 10 to 15 points indicated improvements. The magnitude of change between ‘worse’ and ‘no change’ groups was again reasonably consistent over time for most of the subscales, with the possible exception of the Intrusiveness and Auditory subscales. Interestingly, in the Auditory subscale, the mean change scores for participants reporting that their tinnitus has become worse are larger at 6 and 9 months than at 3 months (and compared to any other subscale) and the magnitude of change reflects this. Therefore, from this low-level analysis, we can observe that the changes in the TFI global and subscale scores are reflecting the changes in the global rating categories, especially for the baseline comparison.

Spearman’s correlations coefficients comparing the TFI global and subscales mean change scores with the five (and three) anchor ratings of change scores (baseline comparison and 3 month comparison) at 3, 6 and 9 months are reported in Table 4.18.

As predicted, for the “baseline comparison” data, the correlations between the TFI global scores and the five anchor ratings of change for 3, 6 and 9 month were moderate (Spearman’s rho = 0.5). On the other hand, the correlations for the subscales ranged from moderate to weak, suggesting that the TFI subscales scores may not be reflecting the change ratings as much as previously assumed. Unexpectedly, the “3 month comparison” data indicated moderate to weak relationships between the TFI global and the five (and three) anchor ratings of change scores at 6 and 9 months.

Table 4.18. Correlations between the TFI and the global rating of change

		Baseline comparison						3 month comparison			
		T0-T1		T0-T2		T0-T3		T1-T2		T2-T3	
N° of global change categories		*5*	*3*	*5*	3	*5*	3	*5*	3	*5*	3
Total	TFI	-0.46	-0.45	-0.54	-0.51	-0.52	-0.48	-0.36	-0.34	-0.33	-0.31
Subscales	*INT*	-0.33	-0.33	-0.50	-0.48	-0.54	-0.49	-0.30	-0.32	-0.12	-0.10
	SOC	-0.37	-0.36	-0.49	-0.47	-0.43	-0.39	-0.27	-0.26	-0.20	-0.18
	COG	-0.24	-0.22	-0.34	-0.32	-0.37	-0.37	-0.26	-0.25	-0.19	-0.17
	SLP	-0.31	-0.29	-0.42	-0.39	-0.46	-0.42	-0.28	-0.28	-0.26	-0.23
	AUD	-0.22	-0.22	-0.38	-0.33	-0.36	-0.31	-0.28	-0.25	-0.19	-0.17
	REL	-0.35	-0.33	-0.36	-0.34	-0.40	-0.38	-0.22	-0.22	-0.18	-0.15
	QOL	-0.33	-0.33	-0.33	-0.29	-0.44	-0.40	-0.22	-0.21	-0.34	-0.32
	EMO	-0.31	-0.30	-0.39	-0.38	-0.35	-0.31	-0.30	-0.28	-0.22	-0.23

Therefore, this anchor rating of change using the 3 month reference may not be reliably categorising the TFI global scores which could adversely affect the estimates of change. This was a consideration in the following analysis.

ROC analysis was initially conducted comparing participants reporting slight improvement (n = 39) on the global rating of change with those reporting no change at 3 months (n = 101). The ROC curve generated for this comparison is presented in Figure 4.10. The AUC exceeded the recommended criteria (AUD = 0.7) with 95% CIs (0.68 – 0.82) that indicated a reasonably good level of accuracy at identifying improvement based on small changes. However, the sample size for "slight improvement" group was small in comparison to the "no change" group, therefore to increase the power in the comparison, the global rating categories were reduced to 'improved', 'no change' and 'worsened' (see Tables 4.17).

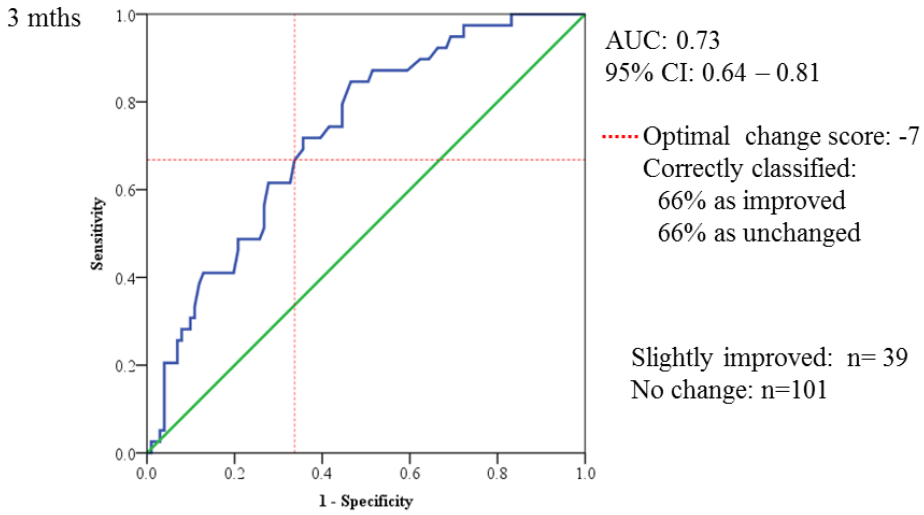

Figure 4.10. Receiver operating characteristic (ROC) curve for identifying changes on the TFI global that signify slight improvements based on the responses to the global rating change question at 3 months.
Green line indicates 50% probability of correctly classifying improvement.

The AUC for the comparison between the 'improved' and 'no change' group at 3 months was slightly higher at 0.75 (95% CIs = 0.68 – 0.82) indicating that the TFI global scores have reasonably good level of accuracy when identifying changes in scores (Figure 4.11). For the "baseline comparison" data, the AUC for the global TFI at 6 months indicated the same, whilst at 9 months the accuracy had fallen below the recommended criteria (AUC = 0.67) with 95% CIs (0.57 – 0.77) (Figure 4.11). This indicates that in a random sample there is potentially a 43% probability that the TFI will be unable to identify and discriminate people who have improved from those who did not.

For the "3 month comparison" data, the AUC for the global TFI at 6 months is alarmingly lower than that observed above and the recommended criteria (0.58), indicating poor accuracy in detecting improvement (Figure 4.12). It is important to remember, this does not necessarily mean that the global TFI score is unable to identify improvement at 3 months. It simply indicates that the global rating of change becomes susceptible to recall bias.

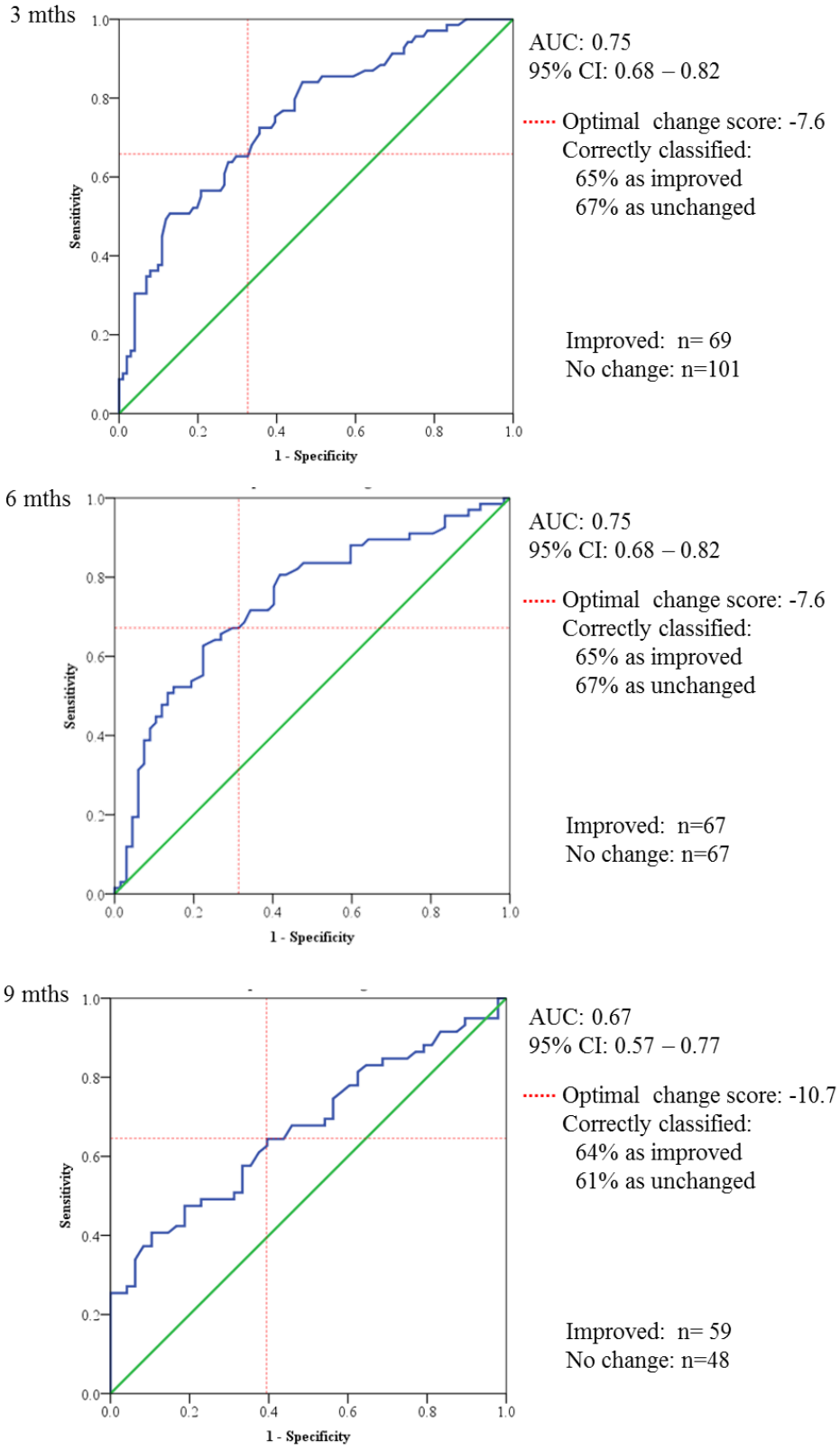

Figure 4.11. Receiver operating characteristic (ROC) curve for identifying changes on the TFI global that signify improvements based on the responses to the global rating change question at 3, 6 and 9 months compared to baseline.
Green line indicates 50% probability of correctly classifying improvement.

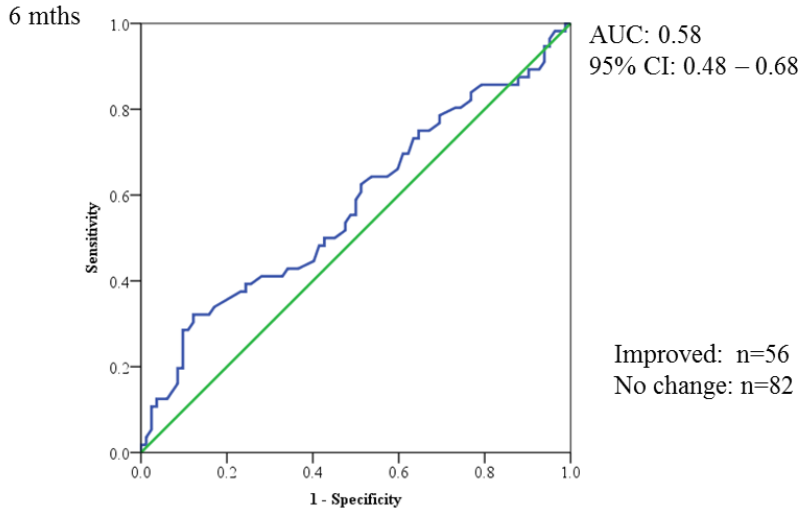

Figure 4.12. Receiver operating characteristic (ROC) curve for identifying changes on the TFI global that signify improvements based on the responses to the global rating change question at 3, 6 and 9 months compared to 3 months ago.
Green line indicates 50% probability of correctly classifying improvement.

In fact, given that the most likely clinical practice is that baseline data will be used for comparisons with data measured at any other time; 3 month data were not used in any other analyses.

The AUC values at 3 months for the Intrusiveness, Sense of Control, Relaxation and QoL subscales were just below the 0.7 criteria, although the 95% CIs suggest a 40% chance that the subscale would be unable to identify improvement. The Cognition, Sleep, Auditory and Emotional subscales were all below the recommended criteria indicating reasonably poor accuracy in detecting improvements (Table 4.19). The AUC values continue to fall over 6 and 9 months, suggesting that the subscales are not as reliable over time as the global score, which is consistent with the findings observed for agreement (section 4.4.6.3).

To identify whether the global TFI is able to detect changes in scores that are associated with self-reported worsening of tinnitus, ROC analysis was conducted comparing participants reporting worsening of their tinnitus on the global rating of change with those reporting no change.

Table 4.19. Characteristics of the Receiver Operating Characteristic analysis and the optimum cut-off point for TFI subscales

	T0 – T1 (3 months)				T0 – T2 (6 months)				T0 – T3 (9 months)			
	Improved (69) vs Unchanged (101)				Improved (67) vs Unchanged (67)				Improved (59) vs Unchanged (48)			
Scale	AUC (95%CI)	Optimal value	Sens	Spec	AUC (95%CI)	Optimal value	Sens	Spec	AUC (95%CI)	Optimal value	Sens	Spec
Intrusiveness	0.69 (0.60 – 0.77)	-11.66	63%	69%	0.68 (0.59 – 0.77)	-13.34	60%	58%	0.63 (0.52 – 0.74)	-18.34	59%	56%
Sense of Control	0.69 (0.61 – 0.77)	-8.34	66%	63%	0.73 (0.64 – 0.82)	-13.34	67%	66%	0.64 (0.54 – 0.75)	-13.34	63%	63%
Cognitive	0.61 (0.52 – 0.69)	-6.66	57%	56%	0.66 (0.56 – 0.75)	-8.34	60%	61%	0.65 (0.55 – 0.75)	-8.34	62%	60%
Sleep	0.64 (0.56 – 0.73)	-8.34	57%	57%	0.67 (0.57 – 0.76)	-11.66	64%	67%	0.63 (0.53 – 0.74)	-16.65	64%	65%
Auditory	0.64 (0.56 – 0.73)	-1.66	53%	59%	0.56 (0.46 – 066)	-1.66	54%	54%	0.54 (0.43 – 0.65)	-1.66	54%	56%
Relaxation	0.69 (0.60 – 0.77)	-13.33	61%	69%	0.64 (0.55 – 0.74)	-16.67	61%	61%	0.64 (0.54 – 0.75)	-15	61%	60%
Quality of life	0.69 (0.61 – 0.77)	-6.25	65%	65%	0.62 (0.53 – 0.72)	-6.25	60%	58%	0.66 (0.56 – 0.77)	-6.25	61%	63%
Emotional	0.66 (0.57- 0.74)	-8.34	62%	68%	0.73 (0.64 – 0.82)	-8.34	65%	70%	0.63 (0.52 – 0.73)	-11.67	55%	65%

AUC = Area Under the Curve. CI = Confidence Interval. Sens = Sensitivity. Spec = Specificity.

However, despite collapsing the ratings of change categories, very few patients reported a worsening of their tinnitus at 3 months, therefore ROC analysis was not conducted on this data. ROC analysis conducted on the 6 months global TFI change scores indicated an AUC value of 0.66, with the 95% CIs once again (0.56 – 0.77) indicating that the possibility of poor discrimination between 'no change' and 'worsening; (Figure 4.13). However, the AUC value (0.70) at 9 months indicates good accuracy of the global TFI to discriminate participants whose tinnitus has become worse from those did not change (Figure 4.13).The difference between the different AUC estimates could be attributed to the increased number of participants reporting their tinnitus had worsened at 9 months compared to 6 and 3 months, adding more stability to the comparisons. Therefore, the global TFI potentially has the ability to detect worsening in scores, but because of the small sample size we cannot confidently conclude the accuracy of this ability.

4.4.7.3 Estimating effect size for improved, no change and worse responses

The ES were calculated for the global TFI and subscales within the three perceived rating of change categories for 3, 6 and 9 months (Figure 4.14). For the improved groups, ES were as predicted (medium to large positive ES) across the TFI global and the majority of subscales scores for all time intervals. Large ES were observed for the TFI global scores in the 'improved' group, ranging from 1.1 to 1.2, whilst ES for THI were somewhat smaller, ranging from 0.9 to 1. This finding confirms that the TFI is slightly more responsive to changes in scores than the THI (see Meikle et al, 2012). The only subscale with a small ES for the 'improved' group was the Auditory subscale scores, and the ES notably decreased by 9 months.

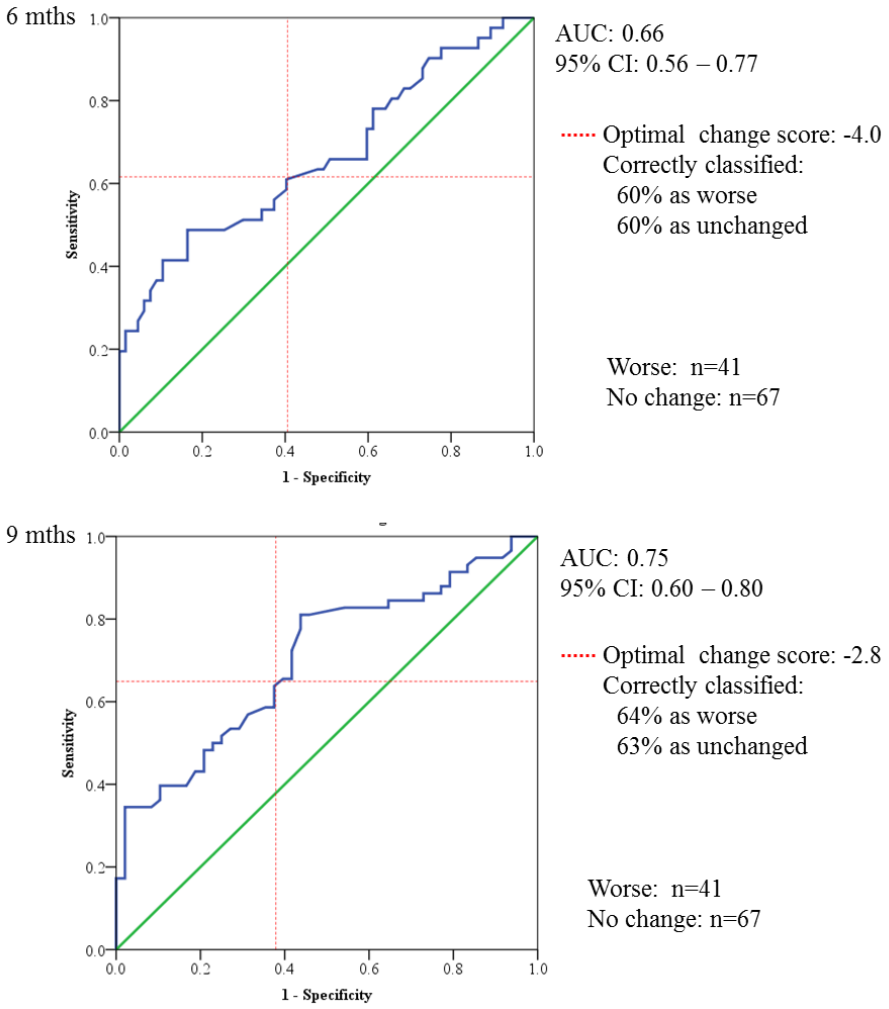

Figure 4.13. Receiver operating characteristic (ROC) curve for identifying changes on the TFI global that signify worsening based on the responses to the global rating change question at 6 and 9 months compared to baseline.
Green line indicates 50% probability of correctly classifying improvement.

The ES for the 'worsened' groups were somewhat smaller than predicted but were in general in the right direction (negative), and in this case the largest effect was observed for the Auditory subscale. The ES for the 'no change' groups were considerably larger than expected. In some cases, they were considerably larger than zero indicating that for participants that report stable tinnitus there is still reasonably large variability in TFI and THI scores, which should be considered when conducting clinical trials with control groups that do not receive the treatment.

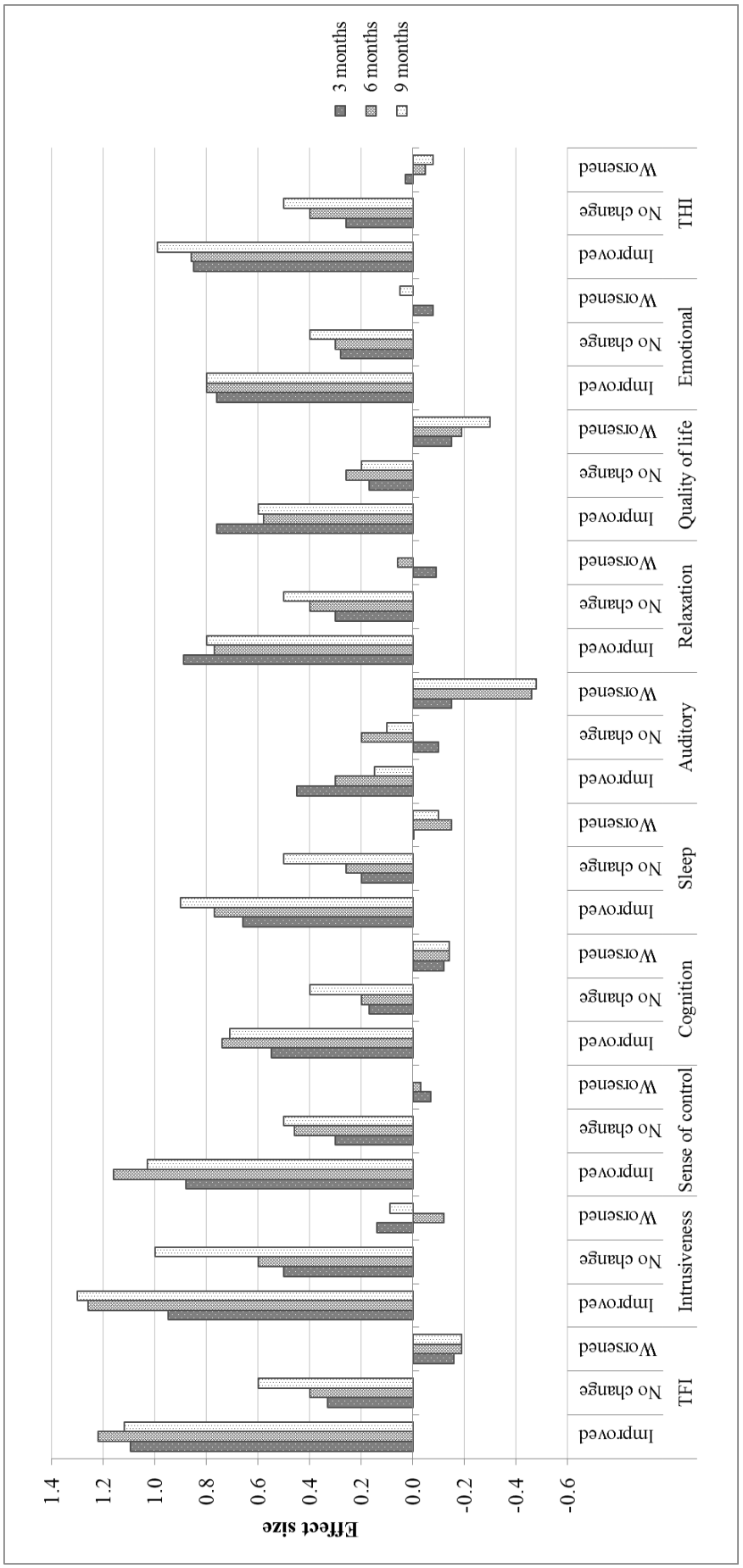

Figure 4.14. Effect sizes for TFI global and subscales and the THI corresponding to improved, no change and worsened groups at 3, 6 and 9 months.

4.4.8. Interpretability of the TFI scores

4.4.8.1 Grading system

The distribution and descriptive statistics for the TFI global baseline scores (n= 255) corresponding to the quartile analysis, the THI grading system and the ratings of perceived problem are presented in Figures 4.15 – 4.18.

Quartile analysis divided the data into four categories ranging from 0 – 36 (mean: 25.6), 37 – 52 (mean: 44.5), 53 – 67 (mean: 60.8), 68 – 100 (mean: 81.5) (Figure 4.15). A one-way ANOVA indicated that these categories were significantly different from each other ($F(3, 251) = 783.10$, $p > 0.001$). The distribution of scores based on the THI grading system were significantly different between categories ($F(3, 251) = 138.83$, $p > 0.001$) as were the perceived problem ratings ($F(3, 249) = 89.34$, $p > 0.001$). Although, both indicate broader ranges within each category (Figures 4.16 – 4.17), the mean scores within each category were similar across the different approaches. The highest percentage of responses in each of the categories indicates similar ranges to that identified in the quartile analysis.

For the ROC analysis, the ranges in scores within the perceived problem categories were used, although some adjustments were made to clearly define the categories. In particular, if there were any large conflicts in classification of the score between their perceived problem rating (i.e. identifying a very big problem) and THI grading (mild problem) then the score classification was adjusted based on the TFI score (Figure 4.18).

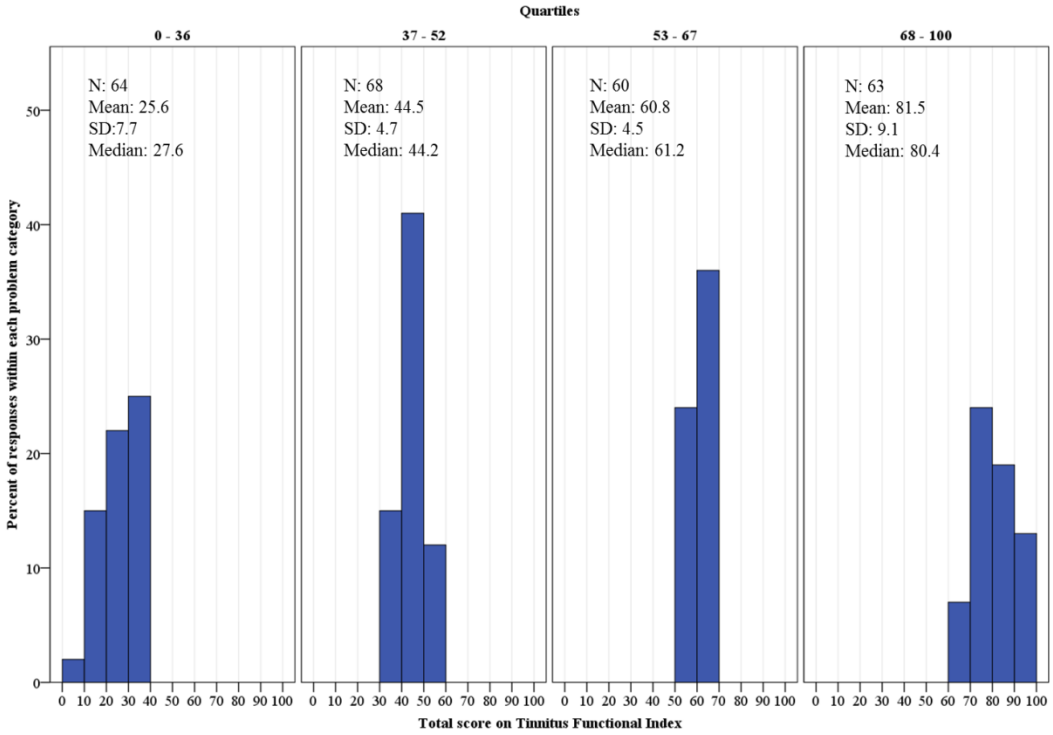

Figure 4.15. Distribution of the TFI global scores separated into quartiles

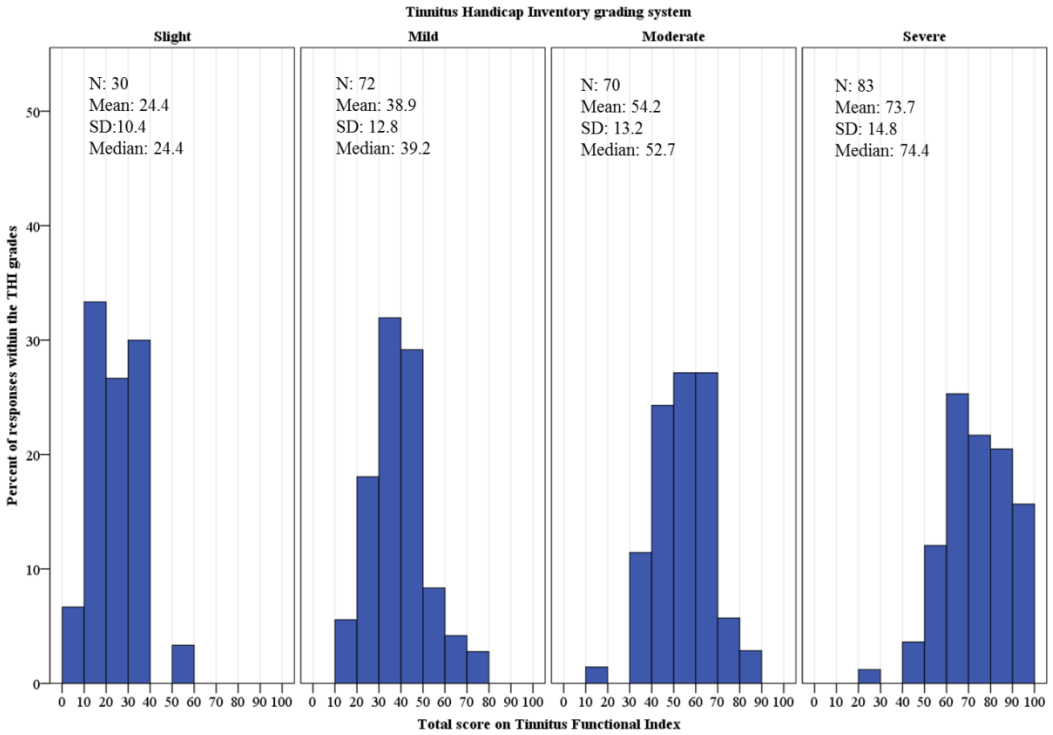

Figure 4.16. Distribution of the TFI global scores corresponding to the THI grades of tinnitus severity

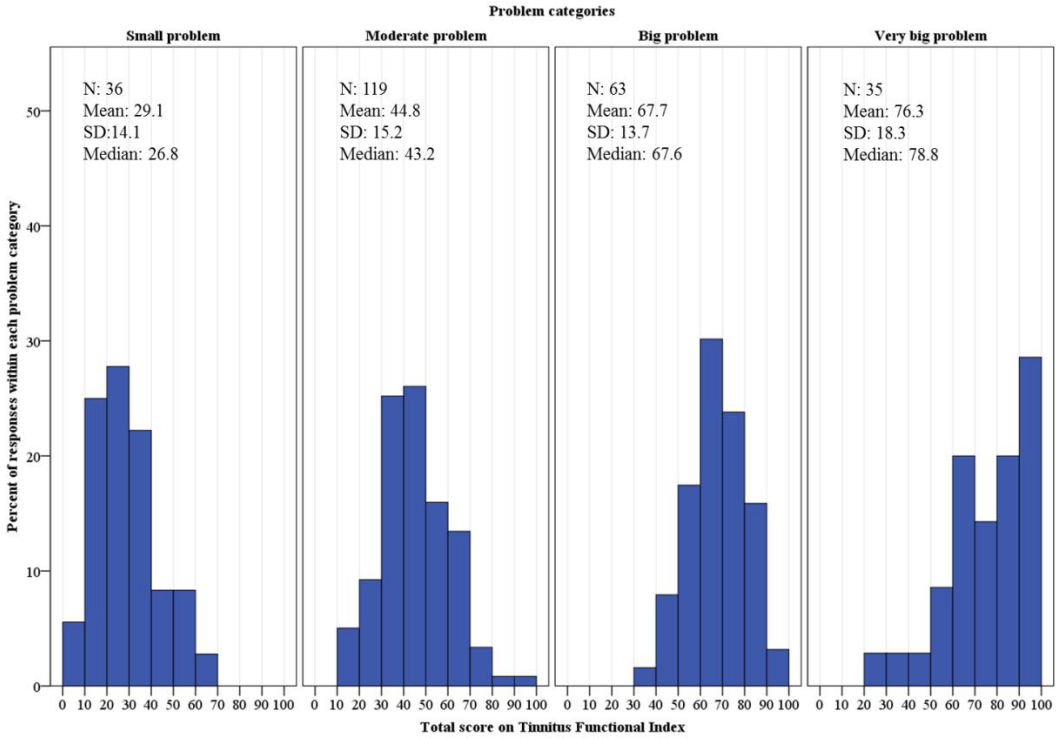

Figure 4.17. Distribution of the TFI global scores corresponding to the problem rating categories

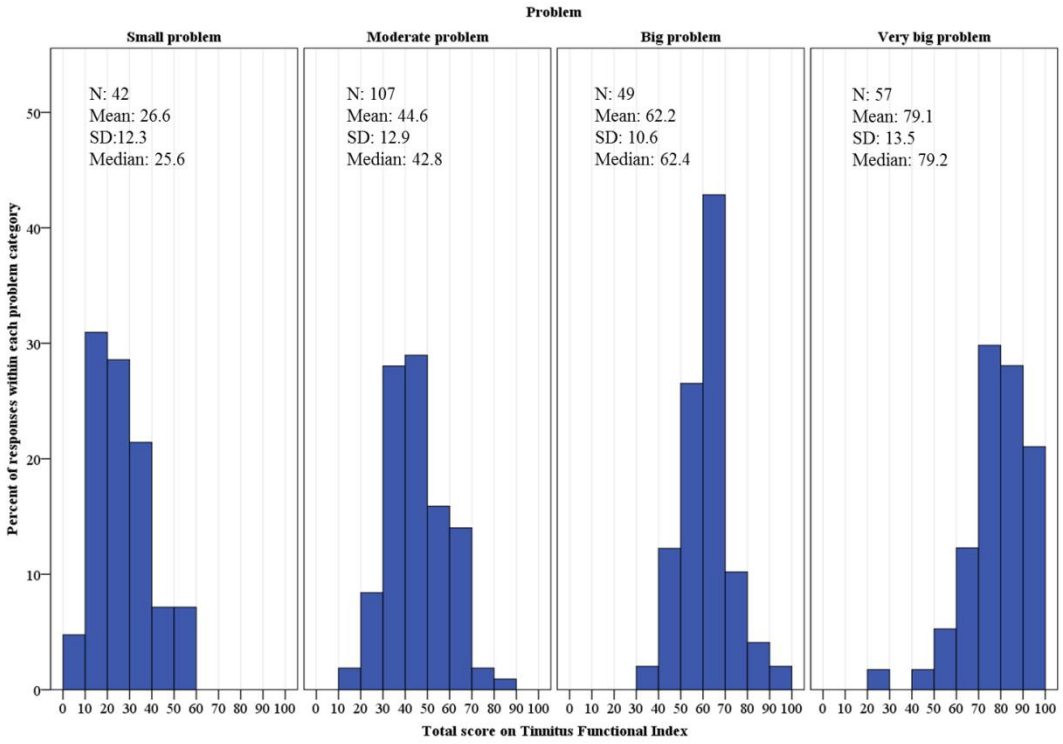

Figure 4.18. Distribution of the TFI global scores in the final categories for ROC analysis

ROC analysis was conducted comparing each problem category with the adjoining lower problem category, for example comparing participants reporting moderate problems (n = 107) with those reporting small problems (n = 42) (see Figure 4.19). The AUC in all three comparisons (AUC ≤0.85) exceeded the recommended criteria (AUC<0.7) indicating excellent ability to discriminate participants reporting different levels of perceived problems. The sensitivity and specificity rates were plotted for multiple possible cut-off points for each analysis. As a diagnostic tool, priority was place on identifying participants with the higher level of problem with their tinnitus.

Examination of the ROC curve and the estimate cut-off values for detecting participants with moderate problems from those with small problems (Table 4.20), indicated that a cut-off value of 28 approximates the optimal cut-off value that was sensitive to identifying moderate problems (94%) from small problems (60%). Therefore, global TFI scores below 28 indicate small problems with tinnitus. The estimate cut-off values for detecting big problems from moderate problems (Table 4.21) and the corresponding ROC curve indicate that a cut-off value of 47 points is optimal for discriminating participants who have big problems from those with moderate (Figure 4.19). Moderate problems with tinnitus are therefore identified by global TFI scores in the range of 28 and 46. To identifying participants reporting very big problems from those reporting big problems an optimal cut-off value of 65 points was identified as correctly classifying 93% of participants as having very big problems and 60% as having big problems (Table 4.22; Figure 4.19). The interpreted grading system is given in Table 4.23.

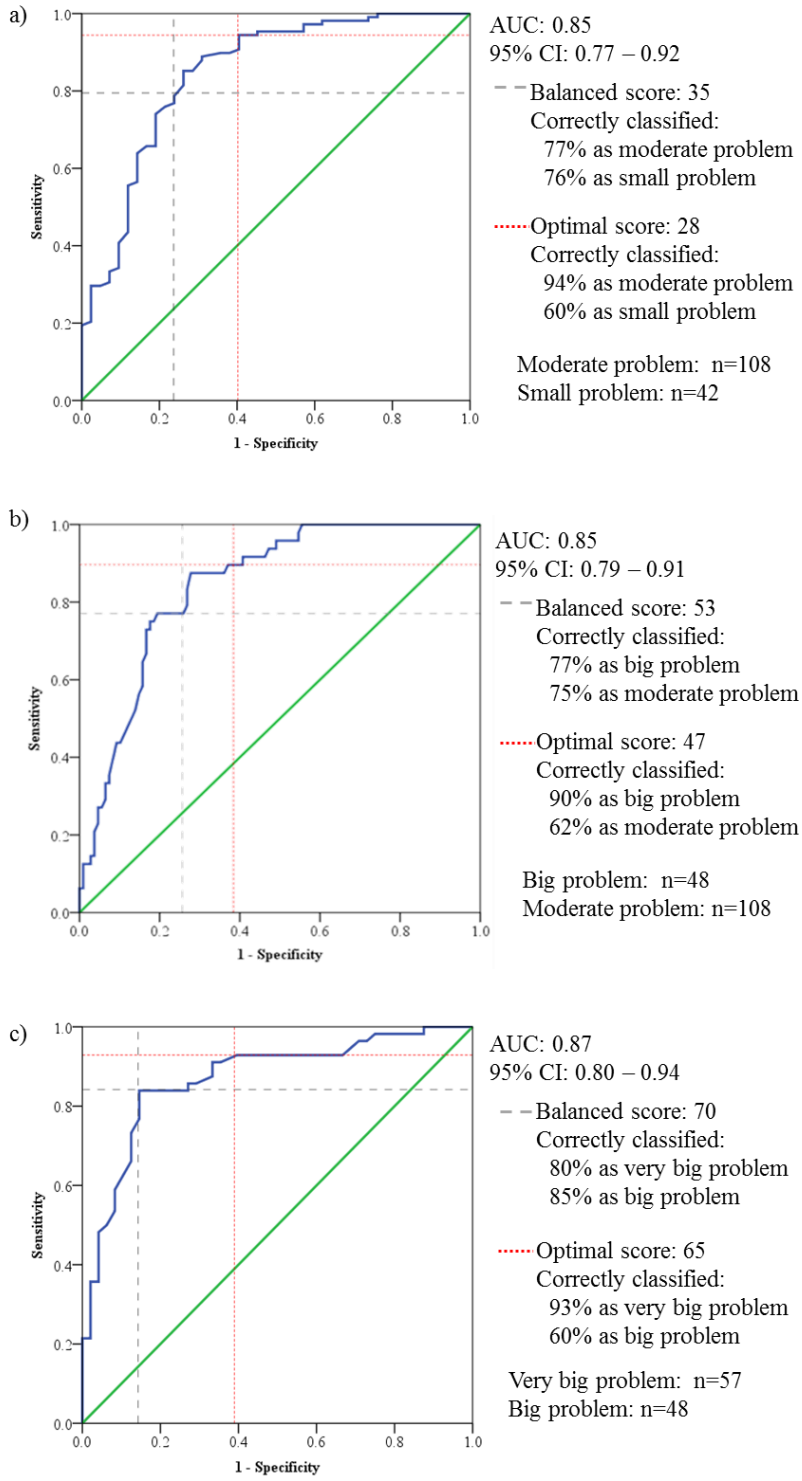

Figure 4.19. Receiver operating characteristic (ROC) curves for identifying optimal cut-off values for different levels in tinnitus severity using the global TFI.
(a) Moderate problem vs Small problems. (b) Big problem vs Moderate problem. (c) Very big problem vs Big problem Green line indicates 50% probability of correctly classifying improvement.

Table 4.20. Optimal grading, cut-off score, sensitivity and specificity rates for identifying small problems with tinnitus using the global TFI.

Small Problem				
Optimal grading	Cut off score	Sensitivity	Specificity	1-Specificity
	7	1.00	0.00	1
	10	1.00	0.05	0.95
	12	1.00	0.07	0.93
	13	1.00	0.12	0.88
	14	1.00	0.17	0.83
	15	1.00	0.21	0.79
	16	1.00	0.24	0.76
	17	0.99	0.24	0.76
	18	0.98	0.26	0.74
	19	0.98	0.33	0.67
	20	0.98	0.38	0.62
	21	0.97	0.38	0.62
	22	0.97	0.43	0.57
	23	0.95	0.43	0.57
	24	0.95	0.45	0.55
	25	0.95	0.48	0.52
	26	0.95	0.52	0.48
	27	0.94	0.55	0.45
	28	**0.94**	**0.60**	0.41
	29	0.90	0.64	0.36
	30	0.89	0.69	0.31
	31	0.85	0.71	0.29
	32	0.84	0.74	0.26
	33	0.82	0.74	0.26
	34	0.78	0.76	0.24
	35	**0.77**	**0.76**	0.24
	36	0.74	0.81	0.19
	37	0.73	0.81	0.19
	38	0.72	0.81	0.19
	39	0.66	0.81	0.19
	40	0.57	0.86	0.14
	41	0.56	0.88	0.12
	43	0.53	0.88	0.12
	45	0.44	0.88	0.12
	46	0.44	0.88	0.12
	47	0.41	0.91	0.12
	48	0.37	0.91	0.10
	49	0.35	0.91	0.10
	50	0.32	0.93	0.10
	51	0.30	0.95	0.07
	52	0.28	0.98	0.05
	53	0.26	0.98	0.02
	55	0.20	0.98	0.02
	56	0.19	1.00	0.02
	57	0.19	1.00	0.02
	58	0.17	1.00	0.00
	60	0.16	1.00	0.00
	61	0.15	1.00	0.00
	62	0.13	1.00	0.00
	64	0.10	1.00	0.00
	65	0.08	1.00	0.00
	67	0.06	1.00	0.00
	68	0.05	1.00	0.00
	70	0.04	1.00	0.00
	72	0.03	1.00	0.00

Table 4.21. Optimal grading, cut-off score, sensitivity and specificity rates for identifying moderate problems with tinnitus using the global TFI.

Moderate problem				
Optimal grading	Cut off score	Sensitivity	Specificity	1-Specificity
	29	1.00	0.00	0.90
	30	1.00	0.11	0.89
	31	1.00	0.15	0.85
	32	1.00	0.16	0.84
	33	1.00	0.19	0.81
	34	1.00	0.22	0.78
	35	1.00	0.24	0.76
	36	1.00	0.26	0.74
	37	1.00	0.27	0.73
	38	1.00	0.32	0.68
	39	1.00	0.35	0.65
	40	1.00	0.43	0.57
	41	1.00	0.45	0.55
	42	0.96	0.46	0.54
	43	0.96	0.51	0.49
	44	0.94	0.53	0.47
	45	0.92	0.55	0.45
	46	0.92	0.60	0.40
	47	**0.90**	**0.62**	0.38
	48	0.88	0.65	0.36
	49	0.88	0.66	0.34
	50	0.88	0.70	0.30
	51	0.88	0.71	0.29
	52	0.83	0.74	0.26
	53	**0.77**	**0.75**	0.25
	54	0.77	0.79	0.22
	55	0.77	0.80	0.20
	56	0.77	0.81	0.19
	57	0.73	0.83	0.17
	58	0.71	0.84	0.16
	59	0.67	0.84	0.16
	60	0.63	0.85	0.15
	61	0.56	0.86	0.14
	62	0.50	0.88	0.12
	63	0.48	0.89	0.11
	64	0.44	0.91	0.09
	65	0.33	0.94	0.07
	66	0.33	0.94	0.06
	67	0.27	0.94	0.06
	68	0.21	0.96	0.04
	70	0.15	0.96	0.04
	71	0.15	0.97	0.03
	73	0.13	0.97	0.03
	74	0.13	0.98	0.02
	76	0.13	0.99	0.01
	77	0.10	0.99	0.01
	78	0.08	0.99	0.01
	80	0.06	0.99	0.01
	81	0.06	1.00	0.00
	83	0.04	1.00	0.00
	88	0.02	1.00	0.00

Table 4.22. Optimal grading, cut-off score, sensitivity and specificity rates for identifying big problems with tinnitus using the global TFI.

Big problem				
Optimal grading	Cut off score	Sensitivity	Specificity	1-Specificity
	48	1.00	0.00	0.90
	49	1.00	0.13	0.88
	51	0.98	0.13	0.88
	52	0.98	0.17	0.83
	53	0.98	0.21	0.79
	54	0.98	0.23	0.77
	57	0.98	0.25	0.75
	58	0.96	0.29	0.71
	59	0.93	0.35	0.65
	60	0.93	0.42	0.58
	61	0.93	0.48	0.52
	62	0.93	0.50	0.50
	63	0.93	0.54	0.46
	64	0.93	0.56	0.44
	65	**0.93**	**0.60**	0.40
	66	0.86	0.71	0.29
	67	0.84	0.73	0.27
	68	0.84	0.79	0.21
	69	0.82	0.85	0.15
	70	**0.80**	**0.85**	0.15
	71	0.77	0.85	0.15
	73	0.73	0.88	0.13
	74	0.71	0.88	0.13
	75	0.68	0.88	0.13
	76	0.66	0.88	0.13
	77	0.59	0.92	0.08
	78	0.55	0.92	0.08
	79	0.54	0.92	0.08
	80	0.50	0.94	0.06
	81	0.48	0.96	0.04
	82	0.43	0.96	0.04
	83	0.41	0.96	0.04
	84	0.38	0.96	0.04
	85	0.36	0.96	0.04
	86	0.36	0.98	0.02
	87	0.34	0.98	0.02
	88	0.32	0.98	0.02
	89	0.25	0.98	0.02
	90	0.21	0.98	0.02
	91	0.21	1.00	0.00
	92	0.18	1.00	0.00
	94	0.13	1.00	0.00
	95	0.11	1.00	0.00
	99	0.02	1.00	0.00
	100	0.00	1.00	0.00

Table 4.23. Grading system for the TFI global

Grades	Range	Nº of participants (%) in each category	Mean (±SD)
Small problem	7 - 28	39 (15)	20.9 (±6.3)
Moderate problem	29 - 53	73 (29)	38.9 (±5.2)
Big problem	54 - 65	71 (28)	57.1 (±5.6)
Very big problem	66 - 100	72 (28)	79.6 (±9.8)

4.4.8.2 Interpreting changes in scores

A minimal important change score was identified using the mean change scores for the different levels of perceived change (ratings of change) reported in Table 4.16. Figure 4.20 displays the changes in scores for global TFI across the five self-perceived levels of change. There is a distinct pattern with a gradual increase in scores from 'much-to-moderately improved' to 'moderately-to-much worse' for all the time intervals, although the magnitude of change differs slightly across the times intervals. At 3 months, the magnitude of change between the 'no change' and 'slightly improved' groups indicated that the minimum change in scores that should be meaningful for patients was -8.7, which was comparable to the change observed by Meikle et al. (2012) for these same groups (-9.1). This magnitude of change for these groups did slowly decreases over time. For example, at 9 months the minimal important change was only -2.9, suggesting that perhaps smaller changes become more important at later time points. To take a conservative approach the minimum change identified at 3 months is considered the minimal important change (MIC) for slight improvements. Interestingly, the degree of change between 'much-to-moderately improved' and 'no change' was reasonably consistent over time, suggesting that a change in scores of -18 would definitely indicate meaningful change for improvements.

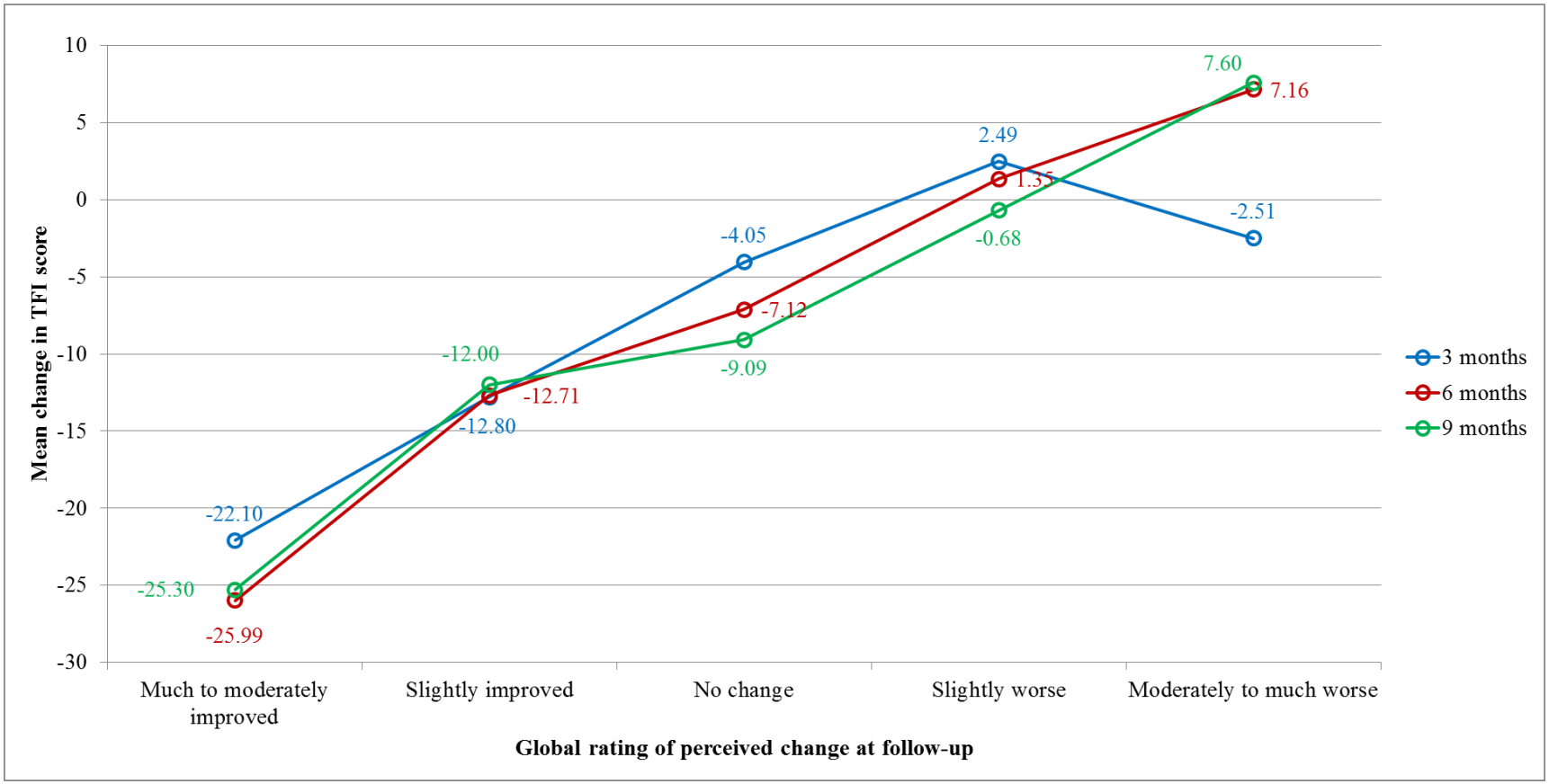

Figure 4.20. TFI global scores at 3, 6 and 9 months corresponding to Global rating of perceived change groups.

Therefore, even though this value is considerably higher than the MIC identified at 3 month for detecting slight improvements (-8.7), for comparative purposes, it was also considered when integrating the anchor and distributed based methods.

To assess whether the MIC estimates were dependent on the magnitude of the baseline values, the distribution of the TFI global scores were examined again within the three ratings of perceived change but stratified by baseline grading group (Figure 4.21). The degree of change between the 'improved' and 'no change' categories did differ depending on the baseline value across all three time intervals. Participants with higher baseline scores are more likely to report larger changes in scores to register an improvement than those with smaller problems at baseline. The difference in scores between improved and no change reflects this pattern, with MIC estimates ranging from -5.5 (small problem) to -13.9 (big problem) at 3 months (Table 4.24). The MIC estimates for the 'big' to 'very big problem' baseline scores were slightly larger than the MIC identified for slight improvement. Unfortunately, due to sample size restrictions within each baseline grading and corresponding perceived change group, no further assessments could be conducted.

Examination of the ROC curve analysis for detecting improvement (presented in section 4.4.7.2), indicated that for slight improvement at 3 months the optimal cut-off value was -7.0 points on the global TFI, correctly classifying 65% of participants as improved and 67% as unchanged (Figure 4.10). This optimal score was reasonable consistent when the ratings of perceived change categories were collapsed to just identify improvement (-7.6) and was the same at 6 months, indicating that a decrease in scores of -8 would indicate improvements (Figure 4.11).

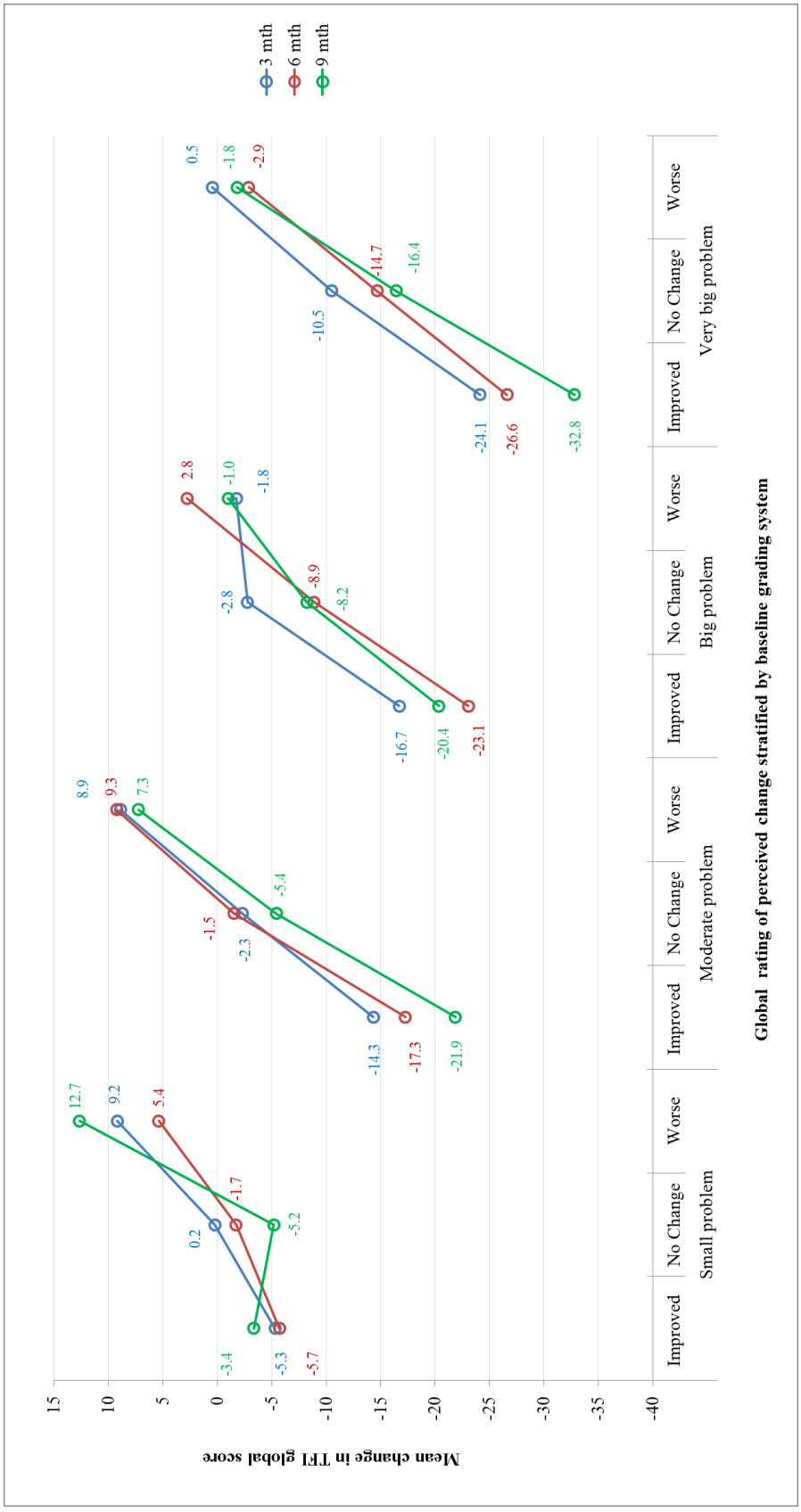

Figure 4.21. TFI global scores at 3, 6 and 9 according to global rating of perceived change groups classified by baseline grading system..

Table 4.24. Mean change scores (SD) and mean difference for TFI global scores classified by baseline grading system and global rating of perceived change.

Tinnitus Functional Index		T0 – T1 (3 months)			T0 – T2 (6 months)			T0 – T3 (9 months)		
Grade	Perception of change	n	Mean (SD)	Diff imp	n	Mean (SD)	Diff imp	n	Mean (SD)	Diff imp
Small problem	Improved	8	-5.3 (9.6)		11	-5.7 (9.6)		11	-3.3 (8.1)	
	No change	22	0.2 (9.1)	-5.5	13	-1.7 (7.1)	-4.0	10	-5.2 (8.9)	1.8
	worse	3	9.2 (14.3)		5	5.4 (18.9)		8	12.7 (21.9)	
Moderate problem	Improved	20	-14.3 (10.3)		18	-17.3 (12.8)		15	-21.9 (12.0)	
	No change	30	-2.3 (10.2)	-12.0	20	-1.5 (7.9)	-15.7	15	-5.4 (12.0)	-16.5
	worse	4	8.9 (14.1)		11	9.3 (12.5)		16	7.3 (17.0)	
Big problem	Improved	21	-16.7 (15.9)		23	-23.1 (16.1)		22	-20.4 (20.2)	
	No change	22	-2.8 (11.8)	-13.9	13	-8.9 (10.2)	-14.2	9	-8.2 (15.0)	-12.1
	worse	10	-1.8 (9.8)		11	2.8 (14.6)		14	-1.0 (16.1)	
Very big problem	Improved	20	-24.1 (19.6)		15	-26.6 (18.9)		11	-32.8 (22.8)	
	No change	27	-10.5 (13.8)	-13.6	21	-14.7 (18.4)	-11.9	14	-16.4 (12.1)	-16.4
	worse	9	0.5 (11.9)		14	-2.9 (14.2)		20	-1.8 (12.7)	

SD = Standard deviation. Diff imp = the difference between mean change scores in the no change and improved groups.

In contrast to the MIC decrease observed above, the optimal cut-off value for improvement at 9 months was -10.7, suggesting that a larger value is required to identify improvements at 9 months than previously indicated (Figure 4.11). However, the ROC curve (AUC) for 9 months was slightly below the recommended criteria suggesting that the global TFI may possibly be unable to detect the improvements as accurately at this time. Therefore the optimal value (-7.6) from 3 months (and 6 months) is provided as the evidence of detecting improvement using ROC analysis.

Finally, the distribution of the global TFI scores for the improved and no change participants at 3 months were plotted in the visual anchor-based MIC distribution plots (Figures 4.22 and 4.23). Plotted with the distributions in Figure 4.22 are the MIC estimate for slightly improvement and the larger estimate for much-to moderately improvement, the ROC optimal value, the SEM estimates and the SDC/LoA estimate (section 4.4.6.3).

It is clearly apparent that for the SEM estimates and the ROC optimal value are largely comparable, and therefore the ROC optimal value cannot be untangled from the measurement error. The MIC estimate (-8.7) on the other hand is slightly above these estimates and therefore can be considered representing change above measurement error. However, inspection of the two distributions suggests that the proportion of participants in the 'no change' group would still be identified after this point; there is still reasonably high variability in the data beyond the MIC which may inflate the possible change scores. The larger MIC estimate for 'much-to-moderately improved' group (-18) is clearly associated with smaller variability and a large peak in participants identifying improvement.

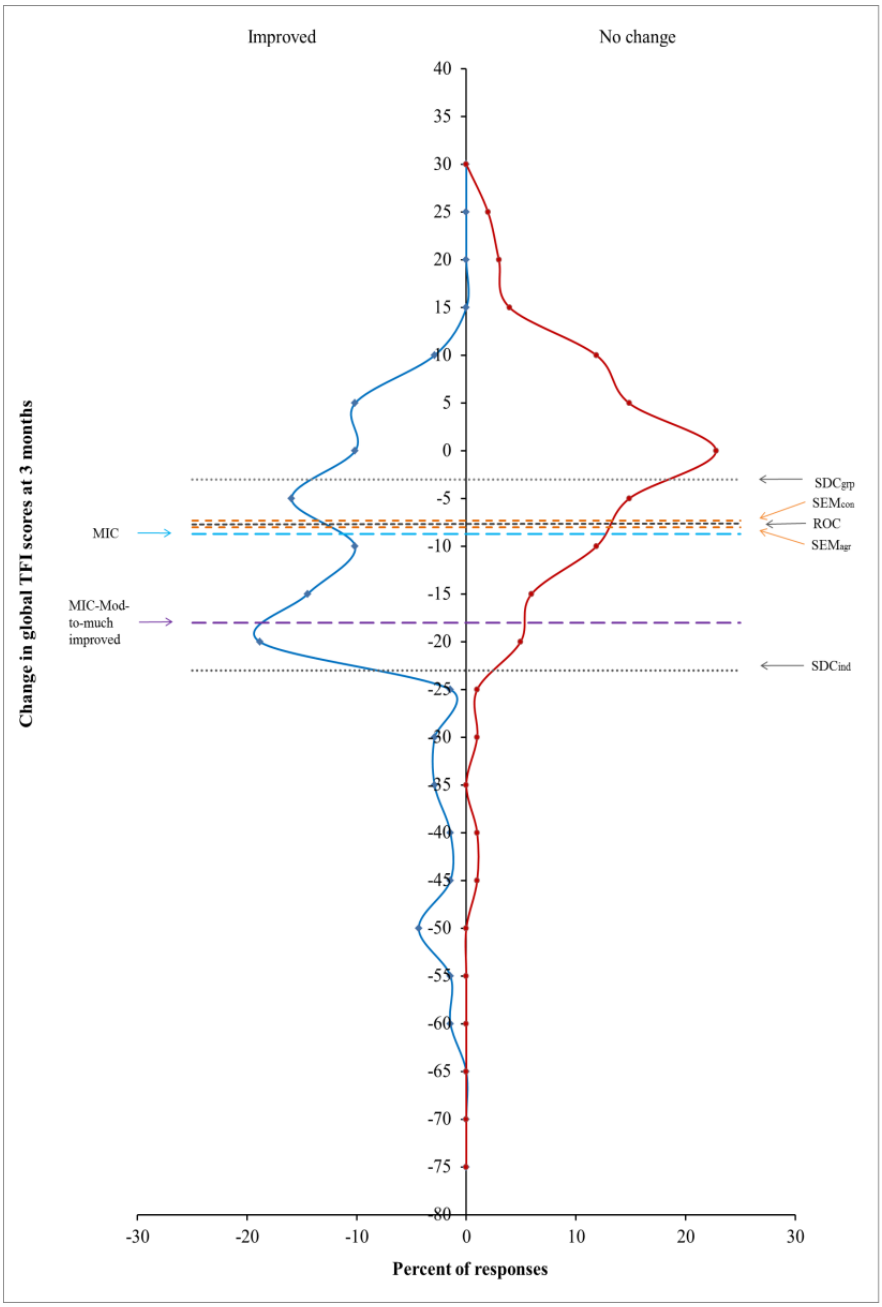

Figure 4.22. Distributions (expressed in percents) of the changes in scores on the global TFI for tinnitus patients who reported improvements in tinnitus and those who reported no change in tinnitus at 3 months.

Horizontal lines indicate Standard Error of Measurement estimates for agreement (SEM_{agr}) and consistency (SEM_{con}) (dotted orange), optimal value from Receiver Operating Characteristic (ROC) analysis (dashed black) and Smallest Detectable Change estimates for groups (SDC_{gp}) and individual (SDC_{ind}) assessment (dotted grey). The MIC = -8.7 (▬ ▬) is above measurement error but there is too much variability in the data to be confident that it is a reliable change. If MIC = -18 (▬ ▬) then there are much fewer 'no change' responses so the change is more reliable.

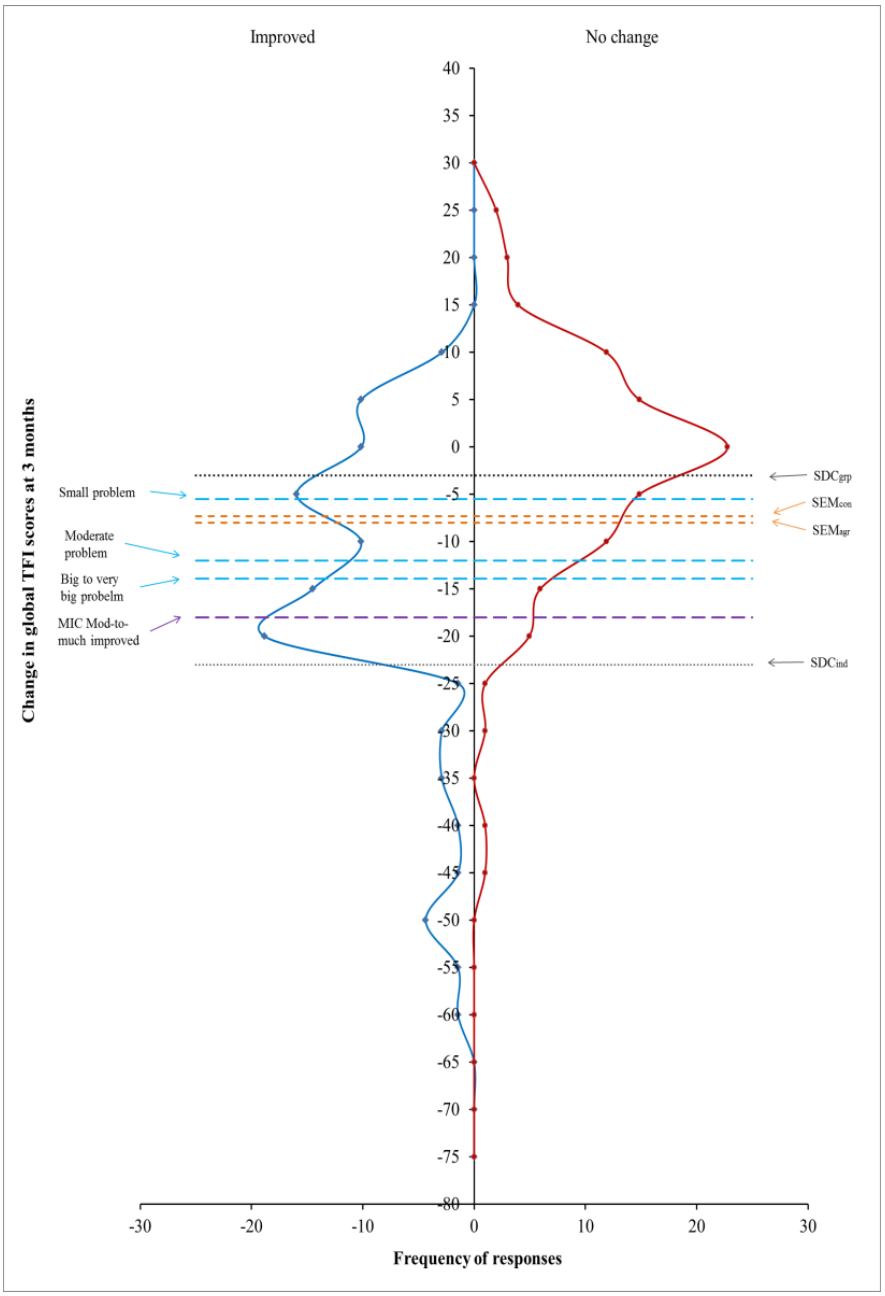

Figure 4.23. Distributions (expressed in percents) of the changes in scores on the global TFI for tinnitus patients who reported improvements in tinnitus and those who reported no change in tinnitus at 3 months with baseline minimal important change (MIC) estimates.
The MIC (-18) is the most reliable change score as it is associated with the fewest responses in 'no change' relative to improved and it is higher than any other MIC estimates associated with baseline problem TFI scores. MIC = Minimal Important Change. SEM_{con} = Standard Error of Measurement for consistency. SEM_{agr} = Standard Error of Measurement for agreement. SDC_{ind} = Smallest Detectable Change for individual assessment. SDC_{gp}= Smallest Detectable Change for group assessment.

Therefore, from this it could be assumed that an MIC of -18 would more clearly identify the true improvement above variability and error. However, although the SDC for group assessment is below the SEM, the SDC for individual assessment

would suggest that an estimate of more than 23 points is required to be above the variability. Examination of the plot showed that at this point, the proportion of participants identified as having improved has considerably reduced whilst the variability is only marginally reduced, the proportion of participants identifying no change is now less than 2% rather than 5%. Therefore, a change score of -23 would be desirable, but given that SDC is not reflecting patients' perceived improvement, the change score of -18 is chosen to represent the minimal important change as it overcomes the majority of the variability, and exceeds the measurement error. This MIC score would also account for the effects of high baseline values such as those in the big (46 – 65) to very big problem categories (65 – 100) (Figure 4.23).

4.5. SUMMARY

The psychometric evaluation performed here provides the first independent evidence, in a UK clinical population, of how reliably the TFI measures different aspects of tinnitus impact, and how well it distinguishes between individuals and detects changes in scores.

Validity

The TFI has good convergent validity being comparable to the construct of tinnitus severity measured by the THI. Classical testing using CFA did not confirm the eight factor structure proposed by Meikle et al. (2012). The Auditory factor was the source of the poor fit as it was consistently unrelated to the other factors and the model dramatically improved when a seven-factor model was re-specified without the Auditory factor. From Chapter 4, my recommendation for a UK clinical population would be for an alternative TFI structure with only seven factors. The Auditory subscale should be thought of and used as a stand-alone subscale, if at all.

Reliability

The global TFI and subscale scores all had reasonably high test-retest reliability, but there was poor agreement between the TFI subscales scores for a "stable" population over the three time intervals (T0, T1 (3 mths) and T2 (6 mths). My recommendation would be for clinicians and researchers to remain mindful of that there is not absolute agreement and so natural variability within and between patients should be expected.

Responsiveness

Substantial floor effects on items were noted. At 3 and 6 months follow-up, the size of the change in global TFI score corresponded well to individual patients' perceived global improvement. But at 9 months, participants experiencing improvements were harder to discriminate from those who remained unchanged. My conclusion is that the TFI is somewhat limited in its responsiveness to detecting improvements beyond 6 months. My recommendation would be that the subscales should be used with caution to identify treatment-related change.

Interpretability

This is the first report to integrate approach using both anchor-based and distribution-based methods and thereby identify a minimal important change score that accounts for both patient perceived benefit and measurement error. My findings indicate a minimal important change score of -18 points. My recommendation would be *not* to use the 13-point difference proposed by Meikle et al. (2012) as a minimal important score for a UK clinical population.

CHAPTER 5. UK VALIDATION OF THE TFI IN A LARGE RESEARCH POPULATION

5.1. INTRODUCTION

International standards proposed by Landgrebe et al. (2012) state that for tinnitus trials at least one validated questionnaire should be clearly defined as a primary outcome measure. For tinnitus research, participants represent a mixed general/experimental population including those who had previously attended clinical appointments for their tinnitus, and those who had never sought medical help for tinnitus. Therefore, the questionnaire needs to be appropriate for this population. What was clearly apparent from the critical evaluation of the current tinnitus questionnaires (Chapter 3) is that, although the majority are used in research, only the TRQ was developed and validated with this population in mind. The reliability and validity of the TFI (25-item prototype 2) is therefore only understood relative to the USA clinical population in which the properties were identified in. It cannot be assumed that the questionnaire will show the same properties when administered to a research volunteer population in the UK.

5.2. AIM AND HYPOTHESIS

The aim of this study was to assess (a) the reliability of the proposed eight-factor TFI structure, verifying item identification with each factor and the underlying construct using CFA, and (b) the ability of the TFI to reliably measure tinnitus severity, distinguishing between individual differences in tinnitus-related symptoms, and responsively measure treatment-related changes in tinnitus.

5.3. METHODS

This was a retrospective analysis of data collected during a two-centre clinical trial (RESET2, clinicaltrials.gov ID:NCT01541969) conducted at the NIHR Nottingham Hearing BRU and the University College London Ear Institute (Hoare et al., 2013).

5.3.1. Approvals

Data were collected in accordance with the permissions granted by the Nottingham 1 NHS Research Ethics Committee and the Sponsor (Nottingham University Hospitals NHS Trust) as part of the protocol described in Hoare et al. (2013).

5.3.2. Participants and Procedure

The eligibility assessment for the trial provided data for the psychometric validation analysis. Assessment included a percentage rating of tinnitus annoyance (PR-A), a Visual Analogue Scale of tinnitus loudness (VAS-L), the TFI, THI, THQ, Beck's Anxiety Inventory (BAI; (Beck & Steer, 1990)) and Beck's Depression Inventory (BDI-II; (Beck et al., 1996)), and the World Health Organisation Quality of Life Bref (WHOQOL-BREF; (The WHOQOL group, 1998)). In total, 294 people with tinnitus (212 male, 82 female) completed most or all of the eligibility assessments. The average age of the participants was 52.8 years (range: 18 to 82). The average duration of tinnitus was 9.0 years (range: 4 months to 50 years). None of the participants were receiving any clinical interventions for their tinnitus at the time of assessment. However, participants were motivated to seek a specific treatment by volunteering for this clinical trial.

Ninety-five eligible participants completed the TFI a second time before beginning the trial, providing data for reliability assessments. The mean time interval between the test and retest was 41 days. Only 44 participants completed the TFI

within the recommended time interval (7-21 days) and so analysis was conducted on this subset only (mean interval: 15 days, SD = 7).

5.3.3. Missing data

Due to some participants not fulfilling the eligibility assessment criteria (visit 1) or being withdrawn at certain points of the eligibility assessment, not all 294 people complete all assessments. Complete case analyse were conducted in which only complete questionnaire datasets were analysed after listwise deletion.

For analysis of the factor structure, internal consistency and item response distributions (floor and ceiling effects) only complete TFI item data was used (i.e. completed all 25 items). The effective sample was 283 participants after list-wise deletion. In 9 of the 11 removed cases, the TFI was not completed at all, whilst in 2 cases only one item was missing (defined as MCAR). For the remaining analysis of construct validity, overall scores were calculated for all of the assessments listed below. Forty-seven participants did not complete all assessments and therefore after listwise deletion the effective sample was 247 participants.

5.3.4. Measures

Descriptions of TFI and THI administrations of procedures were provided in 4.3.4.

5.3.4.1 Percentage Rating Annoyance (PR-A)

As part of the Tinnitus Case History Questionnaire, participants were asked to rate the percentage of time awake they were annoyed by their tinnitus (0-100%).

5.3.4.2 Visual Analogue Scale of loudness (VAS-L)

As part of the 'Tinnitus Tester' computerised test (Roberts et al., 2006, 2008), participants rated the loudness of their tinnitus on a Borg CR100 scale (Borg & Borg,

2001). Participants mark the loudness of their tinnitus at any point along the numerical scale, with word descriptors utilised as an anchor points at 0 for "extremely weak," 30 for "moderate," 50 for "strong," 70 for "very strong," and 100 for "extremely strong" tinnitus loudness.

5.3.4.3 Tinnitus Handicap Questionnaire (THQ)

For each of the 27 items, participants indicate their agreement with each item, by assigning a number between zero (strongly disagree) and 100 (strongly agree). The mean global score reflects the sum of all responses, averaged to give a global score out of 100. Higher scores indicate higher levels of tinnitus handicap. Two of the factors identified by Kuk et al. (1990) are considered reliable to be calculated as separate independent subscales assessing the physical, emotional and social effects and hearing and communication ability.

5.3.4.4 Beck's Depression Inventory – II (BDI-II)

The BDI-II provides a measure of depressive symptomatology, in particular mood and physical effects (Beck et al., 1996; Dozois et al., 1998; Segal et al., 2008). Participants select statements characterising how they have felt over the previous two weeks, and each of the 21 items is rated on a categorical scale (0–3 points). Responses are summed to form a global score out of 63, with higher scores indicating higher levels of depressive symptomatology.

5.3.4.5 Beck's Anxiety Inventory (BAI)

The BAI is a reliable measure of the clinical anxiety which lists 21 common symptoms associated with clinical anxiety (Beck & Steer, 1990; Steer et al., 1993). Participants rate how much they were bothered by each symptom over the previous

week on a categorical scale (0-3 points). Responses are summed to give a global score out of 63 (higher scores indicate greater anxiety).

5.3.4.6 World Health Organisation Quality of Life-BREF (WHOQOL-BREF)

The WHOQOL-BREF is a 26-item questionnaire which provides a broad reliable measurement of perceived quality of life embedded in a cultural, social and environmental context (The WHOQOL group, 1998; Skevington et al., 2004). Although, the WHOQOL-BREF produces four domain scores (physical health, psychological, social relationships and environment), only one of those items measures overall quality of life and general health ("How would you rate your quality of life?"). Only this item is used in this study. This item has 5 response options being (1) "very poor"; (2) "poor"; (3) "neither poor nor good"; (4) "good"; and (5) "very good". The score is transformed onto a 100 point scale, using the WHOQOL-BREF conversion method (The WHOQOL Group, 1998).

5.3.5. Data screening

TFI data were screened for outliers, linearity and multicollinearity. There was no evidence of univariate outliers in the boxplots and histograms, but Mahalanobis distance statistic indicated that there were nine multivariate outliers (Mahalanobis d-squared: 90.72 to 59.15, $p \leq 0.0001$). Although, kurtosis and skewness did not exceed the recommended cut-off points (for kurtosis = 2.00; skewness = 7.00; (Curran et al., 1996)), Mardia's normalised coefficient estimate was 37, considerably larger that the recommended value of < 5 , indicating non-normality in the distribution of the data that requires control.

In terms of the data for all questionnaires (global and subscales scores) being used in the analysis for construct validity, none violated the assumptions relating to

multicollinearity and linearity; analysis of tolerance indices and Variance Inflation Factor (VIF) all met the cut-off points of > 0.10 and <10, respectively (Menard, 2002; Myers, 2000).

5.3.6. Analysis plan

The methods and criteria applied to the questionnaire data were described in detail in Chapter 2. Listed below are the specific methods used and specifications that apply to this dataset in particular. CFA was performed in Mplus 7 (Muthén & Muthén, 2012), whilst the reliability, validity and item distributions analyses were performed in SPSS (v.21.0) and Microsoft excel.

5.3.6.1 Confirmation of the eight-factor structure of the TFI

CFA followed the same analysis plan and fit statistics as those described in Chapter 4. Since there was non-normality in the data the same adjusted estimations methods were applied (maximum likelihood parameter with adjusted Satorra-Bentler scaled Chi-square).

5.3.6.2 Validity of the TFI

To evaluate convergent validity, the global TFI scores were compared to THQ and THI global scores, and VAS-L and PR-A in the same population using Pearson's correlations. The TFI was predicted to have high convergent validity with both questionnaires (correlation > 0.60). Based on reports by Adamchic et al. (2012a, 2012b) and Hiller and Goebel (2006) which compared the TFI with other multi-item tinnitus questionnaires and single item scales, loudness and annoyance single measures only moderately correlate with multi-item measures of tinnitus distress. Therefore the TFI was predicted to show weak convergent validity (correlation < 0.6) with VAS-L and PR-A.

To evaluate discriminant validity, TFI global scores were compared with scores on general health questionnaires (BAI, BDI-II, WHOQOL-BREF) completed by the same participants. Since general health and QoL questionnaires measure general constructs of health, not the tinnitus-specific construct measured by the TFI, it was predicted that acceptable discriminant validity would be indicated by weak to moderate correlations (< 0.6).

As a secondary analysis, the validity of the individual TFI subscales constructs were assessed in relation to the strength of the relationship between them and other questionnaires and their subscales. In order to assess the strength of each individual subscale without the influence of the other subscales, partial correlations and multiple linear regression analyses were performed. It was expected that given the constructs measured by the subscales, the emotional subscale of the TFI would moderately correlate with the BDI-II and BAI scores, and that the QoL subscale of the TFI would moderately correlate with WHOQOL-BREF scores. Based on previous evaluations of the THI and THQ, it is expected the global scores for these questionnaires would correlate with the emotional subscale of the TFI (Kuk et al., 1990; Newman et al., 1996; Baguley et al., 2000; Kennedy et al., 2004).

5.3.6.3 Reliability of the TFI

Reliability was assessed using three methods; Internal consistency, test-retest reliability, and test-retest agreement. Cronbach's alpha and inter-item correlations were calculated for the items in all the questionnaires as per recommendations but only the TFI global and subscales are examined closely. $ICC_{agreement}$, SEM_{con} (SD_{diff}), limits of agreement and smallest detectable change were calculated for the TFI global and subscale scores from two administrations (test-retest).

5.3.6.4 Responsiveness

Response frequency distributions for TFI item data were examined to detect the presence of floor and ceiling effects. The SEM and SDC scores were calculated using test-retest data (method described in 2.2.2.4).

Due to the retrospective nature of the study, interpretability using a global rating of change was not examined. SDC score provides the initial indication of an important change.

5.4. RESULTS

5.4.1. Inspection of the distribution of scores

Descriptive statistics for all questionnaire measures, including the TFI subscales are shown in Appendix A Table 2 (Fackrell et al., 2016). Scores on tinnitus severity questionnaires were moderate (~ 40/100 in each case). For depression and anxiety, mean scores were low, although the range was broad. Cumulative frequency distributions for global TFI, THI and THQ are given in Appendix A Figure 2 (Fackrell et al., 2016). THI global scores were slightly positively skewed towards the lower end of the scales, i.e. 70% of participants scored below 50. THQ global scores had very few higher value scores with all participants scoring less than 70. Compared with these two questionnaires, the TFI global scores appear to be more evenly distributed across the scale, and cover a broad range of scores.

5.4.2. Confirmation of the eight-factor structure of the TFI

5.4.2.1 First-order model analysis

To assess the first-order factor correlations and model fit without the second-order factor, the initial first-order eight-factor model was subjected to CFA. Correlation between the first-order factors ranged from very weak (r = 0.11) to extremely strong

(r = 0.85), but most were strong, with 85% having r = 0.60 (Appendix A Table 2; (Fackrell et al., 2016)). The Auditory factor showed unacceptably weak correlations with all the other factors, from an extremely weak correlation with Sleep (r = 0.11) to a moderate correlation with QoL (r = 0.43). In this case, the Auditory factor appears to be completely unrelated to the Sense of Control, Sleep, Relaxation, and Emotional factors. The Emotional factor showed strong correlations with Sense of Control and QoL factors, indicating potential error variance between the items content or the way on which the items are worded. This was taken into consideration when examining the second-order structure.

Inspection of the model fit statistics indicated that the model fit for the eight first-order factor solution was acceptable (Table 5.1). The S-B χ^2 was significantly large (473.39, p <0.0001), but the S-B χ^2 relative to the degrees of freedom was below the critical ratio cut-off (1.91), the SRMR and the approximation fit indices, CFI, TLI and RMSEA, were all within the recommended criterion, indicating acceptable model fit (Table 5.1).

Standardised parameter estimates for the model revealed high factor loading estimates (> 0.70) for all the items with their designated factor, over 75% of items had loading values above 0.80 (Table 5.2). Item 1 from the Intrusiveness had a factor loading estimate only slightly below the optimal value at 0.68. The loading estimate for Item 4 (Sense of Control) is somewhat lower, although still above the critical criteria (>0.4) indicating that it is potentially contributing less to the designated factor than the other items in the factor. Item 11 from the Sleep factor and Item 14 from the Auditory factor had exceptionally large loading estimates, indicating that these items explained the most variance within the designated subscale and therefore potentially contribute more to the subscale score.

Table 5.1. Summary of the model fit.

	Models	Modified	S-B $\chi 2$ (*df*)	$\chi 2/df$	p-value	TLI	CFI	SRMR	RMSEA (95% CI)
	First-order	None	473.39 (247)	1.91	<0.001	0.95	0.96	0.04	0.057 (0.05 – 0.07)
Second-order	Original TFI-25	None	578.95 (267)	2.17	<0.001	0.94	0.95	0.06	0.064 (0.06 - 0.07)
	Re-specified TFI-25	Item error covariance*	498.48 (264)	1.89	<0.001	0.95	0.96	0.05	0.056 (0.05 - 0.06)

Model based on first-order eight-factor structure, proposed second-order eight-factor structure and re-specified TFI-25 model for the final factor structure with modifications. S-B χ^2 = Satorra & Bentler adjusted Chi-square; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; SRMR = Standardised Root Mean Square Residual; RMSEA = Root Mean Square Error of Approximation.

The squared factor loadings mirrored these findings and highlighted some other likely limitations in the model (see R^2 in Table 5.2). The Sense of Control factor accounted for only 33% of the variance in item 4 and 57% in Item 6. Two items in the Intrusiveness factor and Item 22 in QoL factor had lower squared factor loadings than the estimates for the other items in the TFI. Only 46% and 51% of the variance in Item 1 and Item 2 was accounted for by the Intrusiveness factor and only 59% of the variance in Item 22 was accounted for by the QoL factor. Potentially these items might not be as appropriate for the research population, but this would need to be further assessed with second-factor model and floor and ceiling effects.

An examination of the modification indices (MI) and standardised parameter change (Stdx EPC) revealed the presence of more than 10 potential sources of misfit to the model parameters (MI >10). In particular, error covariance (uniqueness) was identified between item 16 "*How much has your tinnitus interfered with your quiet resting activities?*" and item 18 "*How much has your tinnitus interfered with your*

Table 5.2. Parameter estimates, R-squared values and Standard Error for first-order model structure.

		First-order model			
First-order factors	**Items**	**β**	**B**	**SE**	**R^2**
Intrusiveness	INTR 1	0.68	1.00		**0.45**
	INTR 2	0.72	0.77	0.08	**0.51**
	INTR 3	0.77	1.13	0.11	0.60
Sense of Control	SOC 4	0.57	1.00		**0.33**
	SOC 5	0.88	1.12	0.10	0.77
	SOC 6	0.75	1.10	0.11	0.57
Cognitive	COG 7	0.94	1.00		0.88
	COG 8	0.93	0.96	0.03	0.88
	COG 9	0.91	0.90	0.03	0.82
Sleep	SLP 10	0.88	1.00		0.78
	SLP 11	0.97	1.12	0.04	0.95
	SLP 12	0.91	1.04	0.04	0.82
Auditory	AUD 13	0.92	1.00		0.85
	AUD 14	0.98	1.10	0.03	0.96
	AUD 15	0.89	1.09	0.03	0.79
Relaxation	REL 16	0.93	1.00		0.87
	REL 17	0.94	0.98	0.02	0.88
	REL 18	0.82	0.92	0.04	0.67
QoL	QOL 19	0.83	1.00		0.69
	QOL 20	0.91	1.14	0.05	0.82
	QOL 21	0.85	0.96	0.06	0.73
	QOL 22	0.76	0.92	0.06	**0.59**
Emotional	EMO 23	0.89	1.00		0.81
	EMO 24	0.90	1.06	0.04	0.81
	EMO 25	0.83	0.86	0.04	0.68

The values presented in bold have poor associations with their designated factor, all below the recommended cut-off < 0.40. β = Standardised parameter estimate; B = Unstandardised parameter estimate; SE = Standard Error; R2 = R-squared.

ability to enjoy 'peace and quiet'?" (MI: 35.62; EPC: 1.45) on the Relaxation subscale, and between item 19 "*How much has your tinnitus interfered with your enjoyment of social activities?*" and item 21 "*How much has your tinnitus interfered with your relationships with family, friends and other people?*" (MI: 25.72; EPC:

1.05) on the QoL subscale. This might be attributable to the similarity of the question wording.

Although Item 22 strongly loaded onto the QoL, the content of item is also related to that measured by the Cognitive factor (MI: 25.93; EPC: 1.22). Item 22 asks "*How often did your tinnitus cause you to have difficulty performing your work or other tasks, such as home maintenance, school work, or caring for children or others?*". In this context, the focus is on assessing "difficulties in performing work or tasks" which could be attributed to cognitive processes. Therefore there is logic to this cross-loading. No adjustments were made at this level, but these potential modifications were kept in mind in the following second-order analysis. The first-order model showed acceptable fit for the data, and even though the weak correlations between the Auditory factor and other factors might suggest removing it from the second-order, all factors were maintained to evaluate the fit of the data to second-order model as it was intended (Figure 2.1). For the same reason, the error covariance observed above were not initially estimated in the model.

5.4.2.2 Second-order eight-factor structure of the original TFI-25 model

The model fit was borderline, and was only slightly worse than the fit for the first-order model (Table 5.1). Unsurprisingly, S-B χ^2 was still significantly large (χ^2: 578.95; $p < 0.001$) and the S-B χ^2 relative to the degrees of freedom (df = 267) was now marginally higher (2.1) than the critical ratio cut-off (≤ 2.0).

The SRMR and the approximation fit indices were acceptable albeit less than optimal, suggesting fit could improve with modifications (Schreiber et al., 2006). Factor loading estimates and modification indices were examined to provide a more detailed picture of the model fit and the potential source of the "less than optimal"

fit. Identified parameters were only re-specified if they were conceptually justified and improved model fit.

Standardised parameter estimates for the model replicated those reported for the first-order factor model. The same two items had the lowest factor loading estimates and Items 11 and 14 were again exceptionally large. The factor loading estimates for six of the eight factors with the second-order factor were all above the optimal criteria (0.7) indicating that they all measure the same underlying construct, namely the functional impact on tinnitus. The Auditory and Sleep factors were below the criteria. The Sleep factor loading estimate was marginally below, suggesting that it contributed slightly less to the second-order construct than the other first-order factors. What was most concerning, was that the Auditory factor loading estimate was only 0.31 indicating a very weak relationship to the second-order factor. The squared factor loadings once again mirrored these findings and both the Sleep and Auditory factor estimates were noticeably lower than the others (see R^2 in Appendix A Table 4; (Fackrell et al., 2016)). These estimates indicate that the second-order factor only accounted for 9% of the variance in the Auditory factor. The rest of the squared factor loadings for the factors and items ranged from 0.33 to 0.95. As in Chapter 4, the conclusion is that the Auditory factor is potentially measuring a different construct.

Examination of the modification indices indicated the presence of the same error covariance and cross-loading identified above. Although, specifying the cross-loading of Item 22 might marginally lower the loading estimates for this item, it is of interest to examine the strength in both relationships as this might indicate that item 22 is associated with the wrong subscale or that it is redundant due to poor definition of the construct in which it measures. Therefore, the error covariance and cross-

loading were freely estimated in the re-specified model (Appendix A Table 4; (Fackrell et al., 2016)).

5.4.2.3 Model fit for re-specified TFI-25 model

The Model fit improved and was similar to the fit of the first-order model. SRMR improved and the approximation fit indices were all within desirable limits (Table 5.1), although S-B χ^2 remained < 0.001, the χ^2/df ratio was now 1.89 so within the critical cut-off of < 2.0. Re-specification of the parameters identified as error covariance marginally reduced the factor loading estimate for those items associated with the error, suggesting that the items loading estimates were previously inflated with unique variance. Although factor loading estimates were expected to marginally fall due to the cross-loading, re-specification of the parameters to adjust for cross-loading item 22 substantially reduced the loading estimates for this item on both factors to 0.4 and 0.43 (Appendix A Table 4; (Fackrell et al., 2016)). This was unexpected, but it does suggest that the item is similarly associated with the two factors. The standardised parameter estimates and R-square values for the final model are given in Appendix A Figure 3 (Fackrell et al., 2016).

5.4.3. Validity

Pearson's correlation coefficients between the global scores on all measures (TFI, THI, THQ, VAS-L, PR-A, BDI-II, BAI and global WHOQOL-BREF) are displayed in Appendix A Table 6 (Fackrell et al., 2016).

For convergent validity, results were as predicted. TFI global scores showed strong positive correlations with the THI and THQ global scores (r = 0.82 in both cases) and moderate positive correlations with the VAS-L (r = 0.46) and PR-A (r = 0.58). Therefore, the TFI demonstrated acceptable convergent validity indicating that

it measures a tinnitus construct that is similar to that measured by other multi-item questionnaires.

For most of the TFI subscales, moderate to strong positive pairwise correlations were observed with THI and the THQ global scores (Appendix A Table 7; (Fackrell et al., 2016)). However, when the influence of the remaining subscales were held constant, partial correlations indicated that only the Emotional subscale remained meaningful with a moderate to weak correlation with both the THI and THQ (THI, *pr* = 0.31 and THQ, *pr* = 0.29, respectively), whilst the Auditory subscale maintained a moderate correlation with the THQ (THQ *pr* = 0.41). To confirm the strength of these associations, a series of multiple linear regression analyses were conducted. These beta values (β) mirrored the same pattern as shown by the partial correlations indicating that the TFI is measuring similar properties of tinnitus as in the THI and THQ and of auditory difficulties as in the THQ.

Furthering this, correlations between TFI subscales and the two major subscales of the THQ were initially strong to moderate (Appendix A Table 8; (Fackrell et al., 2016)). However, partial correlation coefficients demonstrated that only the TFI Emotional and Sleep subscales remained meaningfully associated with THQ subscale 1, with a moderate correlation (*pr* = 0.36 and *pr* = 0.31 respectively) and only the TFI Auditory subscale remained strongly associated with THQ subscale 2 (*pr* = 0.71). Acceptable convergent validity was therefore only shown by the TFI Auditory subscale and the THQ hearing and communication subscale.

For discriminant validity, results were also as predicted. Moderate correlations were observed between the TFI global scores and BDI-II (*r* = 0.57), BAI (*r* = 0.39), and WHOQOL-BREF global item scores (*r* = 0.48). Therefore, the TFI

measures construct(s) are distinct from those measured by more general health domains.

Partial correlations between individual TFI subscales and general health, with the remaining subscales held constant, yielded a distinct pattern of results. As predicted, the TFI Emotional subscale meaningfully correlated with all three general health questionnaires (Appendix A Table 7; (Fackrell et al., 2016)). Unexpectedly, the QoL subscale only weakly correlated with WHOQOL-BREF (r = -0.13). The only other notable correlation was the weak correlation between the BDI-II and the TFI Cognitive subscale (r = 0.25). β estimates mirrored findings from the partial correlation analyses, although they were marginally higher. For example, the Cognitive subscale showed a moderate association with the BDI-II. Perhaps the Cognitive subscale is somewhat sensitivity to aspects of cognitive difficulty associated with generalised depression. Overall, these results suggest an acceptable degree of discriminant validity. The BDI-II and BAI are associated with the emotional subscale as expected, whilst unexpectedly the WHOQOL-BREF showed little association with the QoL subscale.

5.4.4. Reliability

5.4.4.1 Internal consistency

Inter-item correlations ranged 0.06 to 0.90 (Table 5.3). Most notably, the Auditory subscale items exhibited the lowest correlations with the other subscales items, in particular extremely low correlations (r > 0.15) with the Sense of Control item 4, Sleep subscale items 10, 11 and 12, and the Emotional subscale item 23 were observed. This potentially indicates that the Auditory and Sleep subscales are unrelated in content.

Table 5.3. Inter-item correlations between all 25 items within the designated subscales.

	QP1	Intrusiveness			Sense of control			Cognition			Sleep			Auditory			Relaxation			Quality of life				Emotional		
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
INT	Q1	1																								
	Q2	0.55	1																							
	Q3	0.56	0.48																							
SOC	Q4	0.26	0.32	0.35	1																					
	Q5	0.38	0.53	0.60	0.53	1																				
	Q6	0.50	0.59	0.51	0.41	**0.66**	1																			
COG	Q7	0.39	0.45	0.49	0.42	**0.69**	0.50	1																		
	Q8	0.34	0.40	0.48	0.43	**0.66**	0.46	**0.89**	1																	
	Q9	0.35	0.46	0.48	0.48	**0.70**	0.50	**0.84**	**0.84**	1																
SLP	Q10	0.25	0.34	0.39	0.28	0.52	0.37	0.50	0.50	0.52	1															
	Q11	0.27	0.34	0.40	0.25	0.50	0.37	0.50	0.49	0.50	**0.86**	1														
	Q12	0.34	0.35	0.44	0.27	0.50	0.39	0.49	0.49	0.48	**0.79**	**0.89**	1													
AUD	Q13	0.25	0.26	0.21	**0.13**	0.18	0.22	0.26	0.30	0.26	**0.09**	**0.07**	**0.10**	1												
	Q14	0.25	0.23	0.22	**0.14**	0.18	0.18	0.29	0.34	0.27	**0.11**	**0.11**	**0.14**	**0.90**	1											
	Q15	0.25	0.25	0.18	**0.09**	0.17	0.18	0.28	0.30	0.25	**0.06**	**0.07**	**0.11**	**0.82**	**0.87**	1										
REL	Q16	0.34	0.46	0.48	0.44	0.64	0.58	0.61	0.58	0.62	0.52	0.50	0.51	0.23	0.21	0.19	1									
	Q17	0.35	0.47	0.53	0.47	**0.67**	0.60	**0.69**	0.65	**0.71**	0.59	0.59	0.60	0.22	0.22	0.20	**0.87**	1								
	Q18	0.26	0.45	0.41	0.43	0.57	0.53	0.50	0.46	0.52	0.44	0.42	0.43	0.21	0.19	0.17	**0.81**	**0.74**	1							
QoL	Q19	0.34	0.33	0.47	0.25	0.52	0.40	0.54	0.56	0.55	0.31	0.32	0.34	0.39	0.39	0.35	0.50	0.56	0.37	1						
	Q20	0.31	0.42	0.56	0.39	**0.66**	0.46	0.64	**0.66**	**0.67**	0.38	0.40	0.40	0.30	0.30	0.28	0.60	**0.67**	0.51	**0.76**	1					
	Q21	0.31	0.28	0.51	0.28	0.53	0.37	0.59	0.62	0.58	0.36	0.36	0.39	0.37	0.39	0.33	0.45	0.55	0.37	**0.78**	**0.75**	1				
	Q22	0.29	0.35	0.43	0.34	0.58	0.38	**0.67**	**0.68**	**0.66**	0.38	0.42	0.41	0.27	0.32	0.31	0.47	0.53	0.41	0.59	**0.68**	0.64	1			
EMO	Q23	0.29	0.43	0.52	0.39	**0.68**	0.51	0.65	0.63	0.65	0.44	0.46	0.49	**0.14**	0.17	**0.13**	0.56	0.65	0.48	0.57	**0.73**	**0.66**	**0.67**	1		
	Q24	0.34	0.49	0.54	0.45	**0.70**	0.60	**0.66**	0.62	**0.71**	0.46	0.46	0.50	0.16	0.16	0.17	0.58	**0.67**	0.51	0.56	**0.72**	0.60	0.60	**0.81**	1	
	Q25	0.24	0.39	0.48	0.39	0.65	0.46	0.61	0.60	**0.66**	0.43	0.41	0.42	0.20	0.20	0.19	0.50	0.60	0.45	0.51	**0.69**	0.61	0.55	**0.74**	**0.74**	1

Blank bold = high inter-item correlations. Red bold = poor inter-relatedness. Blue = within criteria (0.15 – 0.55)

The highest inter-item correlations were between items on the designated subscales and notably between the QoL subscale and Emotional and Cognition subscales. Item 22 is highly correlated with all the items on the Cognition subscale, reflecting the cross-loading identified above. Item 20 is highly related to all three items in the Emotional subscale, potentially indicating an overlap in construct being measured by these items. Otherwise the remaining items generally showed acceptable low to moderate correlations with one another, indicating expected variability in item content.

Alpha estimates for the global TFI scores were high and within criterion (α = 0.80, Appendix A Table 1). Alpha estimates for the TFI subscales were also extremely high, just within the criterion (< 0.60 - > 0.95). Only the Intrusiveness subscale was unacceptably low (0.58) and this further indicates poor fitting items within this dataset, which was also reflected in the lower factor loadings in the TFI structure.

5.4.4.2 Test-retest reliability and agreement

$ICC_{agreement}$ for the TFI global score was 0.91, indicating excellent test-retest reliability. All subscale scores showed similarly high ICCs, ranging 0.81 to 0.95, and the 95% CIs indicated that the estimates were reasonably reliable (Table 5.4). Only the Sense of Control had wide CIs indicated larger unaccounted for variability that could reduce the reliability in a random sample (ICC 95% CI 0.65 – 0.90).

5.4.4.3 Test-retest agreement

The SDC and LoA values for the global and each of the subscale scores were largely comparable (Table 5.4). The TFI global scores had an SDC score of 22.4, whereas the LoA score was 22.1 (±-0.3).

Table 5.4. Reproducibility of Tinnitus Functional Index (TFI) scores: Intra-class correlations (ICC) and limits of agreement between two administrations.

	Mean (±SD)		Difference			Reliability	SEM		Agreement				
Scale	Baseline (T0)	Retest (T1)	Mean diff	SE	SD diff	ICC (95%CI)	Con	Agree	SDC	LoA	LoA Lower limit (95% CI)	LoA Upper limit (95% CI)	%
Tinnitus Functional Index	45.3 (±20.1)	45.6(±19.4)	-0.3	1.7	11.5	0.91 (0.84 - 0.95)	8.1	8.1	22.4	22.1	-22.4 (-28.4 to -16.4)	21.8 (15.7 to 27.8)	93
Intrusiveness	57.1 (±19.1)	58.8 (±21.3)	-1.7	1.6	10.7	0.92 (0.82 - 0.96)	7.6	7.7	21.3	21.1	-22.7 (-28.4 to -17.1)	19.4 (13.8 to 25.0)	93
Sense of Control	58.1 (±22.8)	57.6 (±20.9)	0.5	2.7	17.7	0.81 (0.65 - 0.90)	12.5	12.5	34.8	34.8	-34.2 (-43.5 to -24.9)	35.3 (26.0 to 44.6)	96
Cognitive	39.2 (±38.2)	41.9 (±24.3)	-2.7	2.6	16.8	0.89 (0.79 - 0.94)	11.8	12.0	33.2	32.8	-35.5 (-44.3 to -26.7)	30.2 (21.4 to 40.0)	93
Sleep	41.9 (±31.6)	41.2 (±30.1)	0.7	2.7	18.1	0.91 (0.83 - 0.95)	12.8	12.8	35.5	35.5	-34.8 (-44.4 to -25.3)	36.2 (26.7 to 45.7)	93
Auditory	33.9 (±29.7)	36.1 (±30.2)	-2.3	2.1	13.6	0.95 (0.90 - 0.97)	9.6	9.7	26.6	26.9	-28.9 (-36.0 to -21.8)	24.3 (17.2 to 31.5)	93
Relaxation	64.6 (±25.9)	62.9 (±25.3)	1.7	3.0	19.6	0.83 (0.69 - 0.91)	13.9	14.0	38.5	38.7	-36.8 (-47.1 to -26.5)	40.3 (29.9 to 50.6)	89
QoL	35.1 (±26.1)	34.0 (±24.6)	1.1	2.7	17.8	0.86 (0.75 - 0.92)	12.6	12.6	34.9	34.9	-33.7 (-43.1 to -24.4)	36.0 (26.7 to 45.4)	93
Emotional	36.0 (±28.1)	36.6 (±27.5)	-0.6	2.8	18.8	0.87 (0.77 - 0.93)	13.3	13.3	36.8	36.8	-37.4 (-47.3 to -27.6)	36.2 (26.4 to 46.1)	91

Mean diff = mean difference scores between the repeated measure; SE = Standard Error between the repeated measures; SD diff = Standard deviation of the difference; ICC = Intra-class correlations; SEM con = Standard error of measurement based on SD diff; SEM agree = Standard error of measurement accounting for variance between time, between individuals and random error; SDC = Smallest detectable change; LoA = Limits of agreement.

The SDC and LoA for the TFI subscales were generally considerably larger than the global TFI estimates, except for the Intrusiveness subscale which showed a similar level of variability as the global. The SDC and LoA scores are generally similar. These differences are because SEM_{agree} value (used to calculate the SDC) takes into consideration the variability over time, which is not accounted for in the calculation for the LoA using the SD_{diff}.

Some of the repeated measure change scores in TFI global and subscale scores were not within the identified agreement limits. For three participants, the differences between the TFI global scores were below the defined LoA and therefore only 93% agreement was observed. 95% agreement between scores was only observed for the TFI Sense of Control subscale (Table 5.4). Considering that the Relaxation subscale had the largest limits of agreement estimates, this still did not account for all the variability in the scores, with 11% of participants' difference scores outside the desired limits. The remaining subscales were all borderline, with the majority indicating 93% agreement.

5.4.5. Responsiveness

Response frequency distributions for each item on the TFI were examined for floor and ceiling effects (Appendix A Figure 5; (Fackrell et al., 2016)). Seventeen out of 25 items were rated by more than 15% of participants at the highest (10 points) or lowest (0 points) response option, failing to meet the *a priori* criteria. More precisely, for items 7, 13, 10, 9, 8, 11, 12, 15, 23, 14, 20, 19, 22, 21, and 25, respectively the lowest category of zero was endorsed by 16 to 41% of participants, indicating severe floor effects. Closer inspection shows that all the items in the QoL subscale displayed floor effects which were notably larger than the other subscales.

Only item 4 and item 18 displayed ceiling effects, with responses of 10 (maximum) being observed for 22% and 25% of the population.

SDC scores and SEM identified for the TFI global and subscale scores in the reliability section (5.4.4) provide the first stage of evidence for the TFI's responsiveness in a research population. For the TFI global score, the SDC score was above or below 22.4, similar to the limits of agreement estimate, both indicating that a change in TFI global scores of more than 22 points is required to detect the "true change" above the natural variability and measurement error. In contrast, the smallest detectable change scores for the TFI subscales indicate that much larger changes in scores are required before a "true change" is represented above error. For example, for the Relaxation subscale the scores would need to change more than 39 points to be above measurement error and indicate a "true" change.

5.5. SUMMARY

The psychometric evaluation performed here provides the first account of how reliably the TFI measures tinnitus impact and how well it distinguishes between individual differences in a research population.

Validity

The TFI has good construct validity and converged on the same construct of tinnitus as the THI and THQ. Discriminant validity findings indicated that the TFI score is clearly a different measure from those of generalised depression, anxiety, or quality of life.

CFA indicated an alternative TFI structure was required to best explain the data captured in the general tinnitus population. Scores on the Auditory subscale provided little additional information about the functional impact of tinnitus. But

internal consistency and test-retest reliability of the Auditory factor were generally high. The QoL subscale was somewhat problematic as (i) large floor effects were identified for all four items within the subscale, and (ii) the subscale failed to show any relationship to the single item facet on overall QoL and general health. Item 22 from the QoL subscale was consistently shown to be associated with the Cognition factor. From Chapter 5, my conclusion for a UK research population is that the Auditory subscale is a stand-alone scale. My recommendation would be for researchers to remain mindful that Item 22 appears to be measure elements of both QoL and cognition.

Reliability

The global TFI and subscale scores all had reasonably high test-retest reliability and agreement.

Responsiveness

There were substantial floor effects on the majority of items, particularly for items in the Cognition, Sleep and Auditory subscales. My conclusion is that the TFI is somewhat limited in its responsiveness to detecting treatment-related benefits in a research population.

Interpretability

My findings using SDC indicate that a change in TFI scores of at least 22 points is required to identify “true change”.

CHAPTER 6. RASCH ANALYSIS: AN IN-DEPTH ASSESSMENT OF ITEM RESPONSIVENESS AND THE STRUCTURE OF THE TFI

6.1. INTRODUCTION

Within hearing research, the efforts to develop and validate questionnaires have focused on traditional psychometric methods that have substantially drawn from classical test theory principles. To reiterate, classical test theory is a theory of measurement error that uses total scores as the basis for analysis. A persons "true score" is directly unobservable. Every observed score is made up of measurement error and the person actual "true" attitude or attribute on the latent construct that is being measured (Raykov & Marcoulides, 2011). Within classical test theory, ordinal total scores (defined as interval by Stevens (1946), are postulated to generate interval-level measurements. In Chapters 4 and 5, the validity (structure), reliability, responsiveness and interpretability of the TFI were examined using methods underpinned by this theory and based on best practice guidelines provided by Terwee et al. (2007). There are some limitations associated with classical test theory, such as overlooking individual response patterns in favour of the total scores, assuming that standard error applies to all scores and the assumption that ordinal total scores approximate interval-level measurements (Hays et al., 2000; Hobart et al., 2007; Hobart & Cano, 2009). The difference between 0 and 1 cannot be assumed to be the same as the difference between 3 and 4 (Stucki et al., 1996). To complement classical test theory, new or modern psychometrics methods, sometimes referred to as latent trait models, are emerging as the next logical progression in test development and validation.

Latent trait models include Rasch measurement and Item Response Theory. These are mathematical models that specify the probabilities of person responses to a number of items and the resulting measurement continuum (Wright, 1977). Similar to classical test theory, the relationship between the observed score and unobserved true measurement are evaluated, but in this case the focus is on the association between an individual's unobserved measurement (underlying level of latent trait) and the chosen item response categories (Wright, 1977; Hays et al., 2000; Hobart & Cano, 2009). The unobserved "true" measurement is indicated on an interval-level continuum rather than the combined overall score. The models incorporate the ideals of fundamental measurement from mathematics and medicine; (i) Interval-level structure and estimates that do not vary across the different characteristics of the underlying trait similar; (ii) Person ability and item difficulty are both linked to behaviour and are therefore two separate elements of measurement that should be estimated separately (Andrich, 1988; Andrich & Van Schoubroeck, 1989; Wright, 2000; Hobart & Cano, 2009). The person's response to any item is a product of their level of tinnitus severity (impact) and the difficulty of the task. Both latent models incorporate a variety of different elements (i.e. item and person parameters(Rasch, 1966a, 1966b)), each method uses slightly different parameters, to untangle measurement properties.

Rasch measurement is the simplest model with the fewest parameters, just person ability (or disability, i.e. severity levels) parameters (one per person) and item difficulty parameters (one per item). Unlike Item Response Theory, the total score is considered a sufficient statistics for estimating person ability and item difficulty (Hagquist, 2001; Hamon & Mesbah, 2002; Embertson & Reise, 2000; Bond & Fox, 2007). Regardless of the number of items with high or low scores, persons with the

same total score are assumed to have the same level of severity ability. Similarly, item difficulty estimates are not influenced by the specific person's scores. All items are assumed to equally relate to the trait. No additional parameters are introduced to the model to improve data fit (Embertson & Reise, 2000; Hays et al., 2000). In Item Response Theory models, the total score is not considered sufficient; additional parameters such as item discrimination values and guessing values are included in the model measurement to estimate person ability (Hays et al., 2000; Warne et al., 2012). This changes the structure of the model so that persons with the same total score can have different estimates of the levels of severity. For instance, the discrimination parameters (similar to item-total correlations) calculates the difference between item responses and total scores which identifies the level of discrimination at trait level associated with each item (Hays et al., 2000). Discrimination is regarded as the rates in which the expected scores change relative to the latent measure, i.e. the extent to which success on an item corresponds to success on the whole test (Hagquist, 2001). Separate person locations are estimated for items based on their level of discrimination. Items are not equally related to the latent trait being measured. Unlike Rasch measurement, these parameters can be added at later stages of the modelling. Varying the item parameters maximises the likelihood that data will fit the model (Hambleton & Jones, 1993; Embertson & Reise, 2000). Items can be characterised by one, two, or three parameters, whilst the sample of persons are characterised by a distribution and, again unlike Rasch measurement, persons are not individually parameterised (Hagquist, 2001; Bond & Fox, 2007). Item Response Theory models are in general considered data driven, whilst in Rasch measurement the model is prioritised (Cano & Hobart, 2011). The data are compared to the model which is built upon measurement requirements, not on data assumptions. If the data

does not conform to the model then the data is re-examined for reasons why. No additional parameters are included in the model.

For the purpose of validating the TFI, Rasch measurement was the most appropriate choice, providing the simplest explanation of the data and identifying any potentially problematic items, without resulting in too many changes to the questionnaire response structure. Additionally, Rasch provides individual person parameter evidence for the reliability of the questionnaire for the given sample population. A variety of fit statistics assess the extent to which the items and persons accord to model expectations and although the model does not include the additional discrimination parameters, the item discrimination is accounted for in these fit statistics. Therefore there is no need to apply a modelling method that includes additional discrimination parameters for items as the aim of this study is to validate an existing questionnaire retaining as much of the original structure as possible. The TFI was designed for all the items to *equally* contribute to the subscales or overall scores (second-order structure). Therefore if the data does not fit then this is an indication that the TFI is not working.

A number of measurement requirements underpin the Rasch model; (i) item and person parameters can be estimated separately; (ii) rating scales should be defined on an interval-level continuum; (iii) the model is given primacy (mentioned above); and, most importantly, (vi) the property of invariance. The next section will provide a brief overview of Rasch measurement models and the property of invariance.

6.1.1. Rasch measurement model

The Rasch measurement model was originally developed by the Danish mathematician Georg Rasch (1966) for dichotomous scales. In the model, the observed response patterns are compared to the expected responses using the total scores and probabilistic scaling. The formula of which is:

$$In\left(\frac{\pi_{ni}}{1-\pi_{ni}}\right) = \beta_n - \delta_i \qquad (6.1)$$

where the probability (π_{ni}) of a person (n) endorsing the item (i) (positive response given two responses are possible) is governed by the difference between the person's level of the attribute (β_n) and the level of attribute expressed by the item (δ_i). The model places both items and person parameters on a linear logarithmic (log-odds) scale (*In*) (Masters, 1982; Ostini & Nering, 2006; Pallant & Tennant, 2007; Bond & Fox, 2007).

Extensions of the dichotomous model have since been developed for polytomous outcomes with more than two response category thresholds (the intersection point between two adjacent categories). Here, the thresholds are either constrained to be equal across items using one set of threshold estimates (Rating scale model; (Andrich, 1978)) or thresholds can vary between items (Partial credit model; (Masters, 1982). The thresholds have individual difficulty estimates based on the 50/50 probability of choosing one category over another.

In the Rating scale model (equation 6.2), the separate threshold difficulty estimates (τ_k) for the probability (π_{nik}) of a person (n) endorsing any given response category on any item (i) are estimated only once for each threshold (k) across the entire item set.

$$In\left(\frac{\pi_{nik}}{1-\pi_{nik}}\right) = \beta_n - \delta_i - \tau_k \qquad (6.2)$$

For example, the threshold difficulty for the first threshold ($\tau_{1)}$ is estimate from the probability of endorsing the second category over the first category across all items. The threshold estimates ($\tau_{k)}$ alongside item difficulty estimates are treated as a separate set of estimates for the entire item set (Bond & Fox, 2007; Masters, 1982; Andrich, 1978).

In contrast, the Partial credit model provides individual threshold estimates (*k*) that are unique to each individual item (*i*).

$$In\left(\frac{\pi_{nik}}{1-\pi_{nik}}\right) = \beta_n - \delta_{ik} \qquad (6.3)$$

Where τ_k, from equation 6.2, is replaced with δ_{ik}, allowing thresholds estimates to vary in number and in distance across the item set (Bond & Fox, 2007; Masters, 1982).

Both the models probabilistic functions are similar to and based on the deterministic Guttman pattern (Guttman, 1950). The Guttman pattern determines the patterns of responses to the items based on ability level, for example if an individual were to get an easy question wrong then the Guttman model would determine that the individual would get all harder questions wrong (Table 6. 1).

In the Rasch model, given a person's ability (e.g. high severity), the probability is that the person would endorse all the items severe categories is higher than them endorsing the lower severity categories. However, unlike the Guttman pattern, the Rasch model allows for random variation in the response patterns, i.e. two individuals with the same level of ability or severity can have varying response patterns (Table 6.2).

Table 6.1. Example of the Guttman pattern.

		Easy → Hard						Total score
		1	2	3	4	5	6	
Least able	Person 1	0	0	0	0	0	0	0
↓	Person 2	1	0	0	0	0	0	1
	Person 3	1	1	0	0	0	0	2
	Person 4	1	1	1	0	0	0	3
	Person 5	1	1	1	1	0	0	4
	Person 6	1	1	1	1	1	0	5
Most able	Person 7	1	1	1	1	1	1	6

* 0 = incorrect answer; 1 = correct answer

Finally, the property of invariance is an essential requirement in Rasch measurement. Rasch proposed that the principles of fundamental measurement as seen in mathematics and medicine should be applied to measurement in behavioural and social sciences, in particular the basic criterion of invariance and unidimensionality (Andrich, 1988; Bond & Fox, 2007). The questionnaire should work the same way across every individual, irrespective of individual differences. In order to make valid comparisons, the item functioning should be the same for males as females when measuring health as would be expected when measuring height, for example. Item functioning should be invariant across different groups of individuals such as those of different gender, age, or ethnic backgrounds. All items within a questionnaire or subscale should simultaneously measure the same single underlying construct and are equally contributing to the overall score (unidimensionality; Marais & Andrich 2008).

Table 6.2. Example of the variability in Rasch pattern.

		Easy → Hard						Total score
		1	2	3	4	5	6	
Least able	Person 1	0	0	0	0	0	0	0
↓	Person 2	0	0	1	0	0	0	1
	Person 3	1	1	0	0	0	0	2
	Person 4	0	1	1	0	0	1	3
	Person 5	1	0	0	1	1	1	4
	Person 6	1	1	1	0	1	1	5
Most able	Person 7	1	1	1	1	1	1	6

* 0 = incorrect answer; 1 = correct answer. Red values would be classified as unexpected in a Guttman pattern. For example, Person 5 would be expected to answer the easier questions correctly; the Rasch model is able to predict the variations in data.

A tinnitus questionnaire with subscales that reflect the presence of different aspects of a construct, more than one trait, could violate the assumption of unidimensionality which could lead to reduced variance in person estimates and reliability (Marais & Andrich, 2008). To minimise the chances that the differences between two samples of the populations are due to measurement problems rather than health differences, no health construct other than tinnitus should affect the measurement precision.

The TFI is multidimensional (Meikle et al., 2012) and therefore violates the fundamental assumption of unidimensionality. The eight subscales however were proposed to all equally contribute to an overall score of the functional impact of tinnitus. These subscales can therefore be thought of as individual subtests that are unidimensional and the data from each can be calibrated using standard Rasch measurement model procedures. The items in these subscales can also be transformed into eight ‘testlets’ which combine to produce the overall construct of

the TFI. This can then be considered a unidimensional construct of the overall score, with all the items measuring the same underlying construct.

6.1. AIMS AND HYPOTHESIS

The aim is to use Rasch measurement to explore overall model fit, individual item and person characteristics, the validity of the response structure, whether the TFI is appropriately suited to the population, and whether the items work invariantly across individuals. None of these issues have been examined previously.

A secondary aim is to provide, where possible, linear transformation scores for the subscales and overall structure to use in parametric analysis and in clinics. To maximise retention of the original structure of the TFI, only changes that were deemed necessary to improve scale functioning were made.

6.2. METHOD

6.2.1. Participants

Rasch analysis was conducted on the data described in Chapters 4 and 5. In total 540 people with tinnitus completed the TFI (Table 6.3), 255 clinical patients and 285 members of the public who volunteered for tinnitus research. Demographic data regarding gender and age were collected for both studies. Hearing thresholds (audiometric pure tone average) were collected for the research population only, whilst the clinical population answered a single global question on their hearing (6.2.6.7). This data was used to examine the potential differences in the group responses (Differential Item Functioning).

Table 6.3. Demographic data for each dataset and the full dataset.

		Full dataset	Dataset A	Dataset B
N		540	261	279
Population	Clinical	255	124 (47%)	131 (47%)
	Research	285	137 (53%)	148 (53%)
Gender	Male	256	176 (68%)	180 (65%)
	Female	183	85 (32%)	98 (35%)
Age	Mean (Std)	53yrs,1 mth (13yrs)	52 yrs (13yrs)	54 yrs, 2 mths (13yrs)
	Range	18 – 84 yrs	18 – 76 yrs	18 – 84 yrs

6.2.1.1 Sample size estimates

The sample size of 540 participants is adequate by any recommendations. A sample size of 243 would give 99% confidence that item calibrations (i.e. the mathematical transformations to a linear scale) are within ± 0.5 logits and stable on a questionnaire that has problems with measurement and targeting the population of interest (Linacre, 1994). A well-targeted questionnaire would require a smaller sample (i.e. 108 participants) to give the same calibrations (Linacre, 1994). Similarly, Chen et al. (2014) found that sample sizes below 250 were more likely to have unstable response structures and person and item calibrations. A smaller sample size over 250 also reduces the known problems with the chi-squared estimates and probabilities. Sample sizes > 500 produce inflated significant chi-square estimates, which indicate an apparent misfit in data, and therefore are less informative and reliable (Andrich et al., 2009; Hobart & Cano, 2009; Chen et al., 2014). If the conservative approach is taken then the optimal sample size needed for stable estimates is 250 participants. To check the reliability of the item calibrations between different samples, larger sample sizes are recommended divided into smaller groups (Linacre, 1994). Consequently,

in order to obtain robust estimates of the construct fit to the Rasch model, the full dataset (n=540) was divided into two similar sized independent groups; dataset A (48%; n=261) and dataset B (52%; n=279) (Table 6.3). The groups were stratified to obtain similar numbers of participants from the clinical and research populations and similar numbers of males and females.

6.2.2. Statistical software

Data were entered into SPSS version 22 for initial analysis before exporting to RUMM2030 (Andrich et al., 2009) for Rasch analysis.

6.2.3. Estimation method

Analysis of the TFI data requires a polytomous Rasch model. A significant Likelihood-Ratio ($p < 0.0001$) indicated that, although the TFI has an 11-point response option for every item, the distance between the response categories were not equal for every item, therefore the unrestricted Partial Credit model was chosen as the polytomous Rasch model.

6.2.4. Data screening

6.2.4.1 Missing data

Rasch measurement makes no assumptions about missing data and computes the estimates of person latent trait levels from the observed data that is available (Wright, 2000). The precision of estimates can be slightly biased (with larger SEs) if a large amount of data is missing (50%; Hobart & Cano 2009; Doganay Erdogan et al. 2013). Within this study the amount of missing data (<0.2%) was negligible.

6.2.4.2 Extreme scores

Extreme scores were excluded from item estimations since the examination of relative item difficulty could be biased with these scores. It is impossible to compare

across items and accurately estimate the item difficulty locations when the model cannot predict how far above (10) or below (0) the extreme the scores are. However, individuals with floor and ceiling effects provide important information about the possible range in scores for a given sample. Therefore, for extreme scores, the person location estimates are extrapolated values.

6.2.5. Analysis plan

The two datasets were used to identify any consistently misfitting items or subscales and to assess the stability of item calibrations, providing evidence of convergent validity of the construct. Therefore, the overall model fit, individual item and person fit, the validity of category ordering and reliability of the targeting for the subscales and the overall TFI second-order structure were examined in both datasets. To evaluate the second-order construct, the items in each subscale were combined together into a single 'testlet' and subjected to Rasch analysis as eight items within a questionnaire scale (uniform structure). The full dataset was used to provide the final confirmation of the data fit and structures (subscales and second-order). The larger dataset meant that person-item parameters (calibrations) could be further stabilised since there were more data points to provide the basis for those calibrations. This reduced the potential error variability (noise) in the parameter estimates (stabilising the estimates) that are sometimes associated with smaller sample sizes (Chen et al., 2014). At this stage, problematic items or subscales were recalibrated when possible or recommended to be removed. Finally the possible invariance between different clinical groups were examined (Differential Item Functioning) and the rescaled logits were transformed into understandable scores for use in clinics and research for the subscale and second-order structure that conformed to the Rasch model.

6.2.6. Assessing the fit of the TFI data to the Rasch model expectations

The Rasch model derives best estimates of person ability (level of severity) and item difficulty from the total scores. The model then backtracks and uses these estimates as the basis to obtain expected responses parameters (log odds values (logit scale)) that satisfy the Rasch model (Hagquist, 2001; Hobart & Cano, 2009). The expected values entirely depend on the person ability and item difficulty location estimates on the latent trait (Hagquist & Andrich, 2004). These expected responses are then compared to the observed responses, in which the fit to the Rasch model is evaluated based on the difference between the two (observed – expected = residual). This provide estimates of the degree of unexplained variance left over given the data fit to the model (if the residual was "0" then the data would be a perfect fit to the Rasch model) and in turn estimates for item difficulty and person abilities (Wright 1977).

One important element in calculating the fit statistics are class intervals. Based on the person location distribution (i.e. different levels of severity), the sample is divided into a series of subgroups; class intervals (Hobart & Cano, 2009). For example, all persons that scored below 10, irrespective of the item difficulty, would be categorised into one class interval. These class intervals are used for within data comparisons across different severity groupings which inform the fit statistics and the item characteristic curves. RUMM2030 software automatically identifies the number of class intervals needed based on the total sample size, the larger the sample size, the more class intervals. The sample size in each class interval should however be assessed externally; small class interval sample sizes ($n<30$) limit the amount of comparisons that can be made using the class intervals (Hagquist & Andrich, 2004). Preferably the sample size in each class interval should exceed 50. Finally, as usual with modelling statistics, any changes to the questionnaire (i.e. removing any items)

should be supported by conceptual foundations such as the clinical importance and usefulness of the items (Bond & Fox, 2007; Hagquist et al., 2009; Mayhew et al., 2011).

6.2.6.1 Overall summary fit statistics

The overall summary fit statistics provide an initial overview of data fit indicating sources of misfit in the entire model, and/or items and/or persons. Two classes of summary fit statistics were considered; (i) item-trait interaction (overall model fit) and (ii) item-person interaction (overall fit of items and overall fit of persons).

Item-trait interaction

Item-trait interaction reflects the hierarchical ordering of items across the trait; in particular it measures the extent to which invariance is reflected across the trait. This is reported as the overall model Chi-square statistic (χ^2), i.e. the summing χ^2 across the entire data matrix (Smith, 2000). χ^2 is computed from comparisons of the residuals across different ability groupings (class intervals) for the trait. So, for every given item, χ^2 statistics are summed for each class interval and then summed together to produce an overall χ^2 statistic for that item. Following this, the χ^2 for all items is summed to produce the item-trait interaction (Smith, 2000; Tennant & Conaghan, 2007). Multiple comparisons, such as these, can escalate the possibility of rejecting the model due to Type 1 error, therefore Bonferroni corrections for alpha values at the level of 0.05 were applied. For example, for subscales with three items, the criterion level for alpha were set to 0.05/3 = 0.017, for subscales with four items, the criterion level was set to 0.05/4 = 0.013. If a significant interaction occurs then the ordering of the items are unexpectedly interacting with person locations across the continuum, indicating there might be poor discrimination across the trait being

measured and individual item and person fit statistics should be examined (Andrich & Van Schoubroeck, 1989).

Item-person interaction

Item-person interaction statistics essentially provide values for the level of consensus between all the items across all persons within the cells of the data matrix (Hobart & Cano, 2009). Therefore, the degree to which each item relates to all other items, each person responds to each item and whether these fit the model expectations are examined. The fit residuals for item-person interactions were transformed to approximate z-scores and represent a standardised normal distribution. The overall mean item fit residual and in particular the SD provides a summary of the degree of divergence between the observed and expected value for all the items from each person-item interaction. Similarly, the overall mean person fit residual (and SD) provided a summary of the divergence over all the persons in relation to a given item. Given that the items and persons fit the model expectations then these residuals would have a mean of approximately zero and a SD of one (Wright, 1977; Smith, 2000; Hobart & Cano, 2009). A residual SD of more than ±1.5 provides the first indication that there are possible issues within the item and person observed values, whilst a residual SD of ±2.5 indicates misfit to the model expectations (Bond & Fox, 2007; Mulcahy & Vaughan, 2015). Again, to isolate the source of the misfit, individual residuals were examined.

6.2.6.2 Individual item fit statistics

Fit residuals and χ^2 statistics for each individual item provide evidence on the extent to which the individual items relate to the underlying construct that is being measured and identify any deviations or problematic items (Smith, 2000).

Individual item fit summarises the residuals from all person responses to each item, indicating the degree in which each individual item responses are coherent with the model expectations. χ^2 provides an estimate of the extent that all the responses (grouped in class intervals) for each item deviate from expectation (Andrich & Van Schoubroeck, 1989). Significant χ^2 estimates indicate that the responses to the item were not consistent with expectation relative to other items within the scale and that the deviations from the expected values were large, relative to chance. Again the same Bonferroni corrections were applied as above at the level of 0.05.

Fit residuals are defined as the standardised sum of all differences between observed and expected values summed over all persons. An individual item fit residual between ± 2.5 were deemed to be satisfactory. Fit residuals exceeding ± 2.5 indicate clear deviations from expectation. High positive values (>2.5) indicate that the item is measuring an alternative construct than the other items in the scale and therefore there is a lack of fit between item and the model. Conversely, high negative fit residuals indicate an overlap in item content and therefore item redundancy (Smith, 2000). Removal of items should be guided by theory as well as these statistics, although misfitting items may possibly reduce validity of the scale, they may not dramatically reduce the precision of the measurement.

Item characteristic curves (resembles a sigmoidal (s-shaped) curve) show the monotonic relationship between the expected values and observed values using the class intervals. The data conform to the model if the class intervals (χ^2 values) are reasonably spaced along the continuum and are closely situated to or on the expected curve. The observed values should not cluster at the asymptotes of the curve (upper and lower bounds) as this indicates a lack of ability to separate individuals (Hagquist et al., 2009). A sign of item misfit, which can correspond to high negative fit

residuals, is over-discrimination in the responses, where the observed values form a steeper line than the expected curve. Limited discrimination is detected when the observed values under-discriminate and form a flatter line than the expected curve (Hagquist, 2001; Pallant & Tennant, 2007; Hobart et al., 2007).

6.2.6.3 Individual person-fit statistics

Individual person fit summarises the residuals from each person on all items, indicating the extent to which an individual's responses unexpectedly diverge from the rest of the responses and from model expectations.

The fit residuals for all persons were examined. Any residual that exceeds ± 2.5 indicates deviation from model expectation and can cause misfit at item level (Smith, 2000; Pallant & Tennant, 2007). These deviations could possibly be a reflection of poor measurement or more likely the variability between symptoms in tinnitus (patients may have problems in only one aspect being measured) or other external factors such as co-morbidities or cognitive deficits. The proportion of individual residuals exceeding ± 2.5 are reported, but were not removed as this would reduce the external construct validity of the scale.

6.2.6.4 Validity of category response ordering (thresholds)

To ensure that the item scores were working the way they were intended, the ordering of response categories for each item was investigated by examining the thresholds and the category probability curves.

Thresholds (τ) are the intersection point in which the probability of endorsing two adjacent categories is equal, i.e. there is a 50% chance that a person will select either category option 1 or category option 2 (Hagquist, 2001; Tennant & Conaghan, 2007). The number of thresholds is directly related to the number of response options

(categories). Hence the TFI has eleven response options (0 to 10) for each item therefore there are ten thresholds (τ1 to τ10) for each item.

Category probability curves provided a visual interpretation of whether the categories were working or not, which was supported with examination of the threshold values. These thresholds should show the intended increasing levels of severity reflecting the same order as the manifest categories. Each response option should at some point have the highest probability of being endorsed, when this does not occur the thresholds become disordered (Hagquist, 2001; Hagquist et al., 2009; Mulcahy & Vaughan, 2015).

Disordered thresholds are when there is a low probability of the response category being endorsed or that participants with a wide range of severity endorse the response option. Disordered thresholds were taken to indicate that there are too many response options or that participants are unable to discriminate between the levels in the response options (Linacre, 2002; Lamoureux et al., 2006; Pallant et al., 2006; Tennant & Conaghan, 2007). Response categories were only collapsed if the disordered categories were apparent in both datasets (A and B) and the model fit improved. Collapsing categories with data that conforms to Rasch (i.e. good fit) can inadvertently impact on person-item parameters and the response structure (Rasch, 1966a).

6.2.6.5 The ability of the TFI to appropriately target the population of interest and reliably distinguish between persons

To ensure that the TFI subscales and second-order structure are appropriately targeted at the population of interest, the person and item location distributions, i.e. the spread and ordering of the items and persons across the continuum and the error

values associated with both, and the Person Separation Index estimates were examined.

The distribution of person location scores relative to the fixed value of zero set for items provides evidence of the ability of the TFI to target the given population in the sample. The item locations should correspond to the distribution of the person difficulties, and vice versa the person locations should be represented by the severity of the items that range the continuum (Hagquist et al., 2009). Severe misalignment of item difficulties with person locations lead to high standard errors and therefore imprecise estimates (Hagquist et al., 2009). The items should be evenly distributed along the continuum with discernible levels of increasing difficulty (severity/impact). Large gaps between item locations indicate limited measurement of person information at those locations and that key information is unaccounted for by the questionnaire (Hobart & Cano, 2009). The item location order should make sense clinically and theoretically, although within tinnitus this is hard to interpret as the severity of the problems can be associated to different degrees with a variety of symptoms.

The person locations are expected to be normally distributed as the measure should, for the most part, account for the variability in the population if the measurement is optimal. The inverse standard error associated with the measurement (represented by an information curve on person-item distribution) provides information on the precision of the questionnaire along the underlying trait, where the questionnaire is performing best. The steeper the curve, the greater item information available (higher information value), the less error associated with the questionnaire and in turn more precision. The majority of the information for both persons and items should be within the curve (Hays et al., 2000; Hobart & Cano,

2009). The mean person location should be as close as possible to zero. If the person mean location is vastly higher than zero (positive value) then the questionnaire is mistargeting the population (Pallant & Tennant, 2007). A large number of extreme scores at either the upper or lower bounds of the logit scale are a very clear indication of poor targeting (Pallant & Tennant, 2007; Hagquist et al., 2009; Kurtaiş et al., 2011).

The item map is a visual representation of the ordering of items and provides an alternative view of the alignment of persons and items locations. The item threshold locations are presented in order of difficultly aligned to person locations, the item thresholds presented first represent the items that would first capture the problems with tinnitus, whilst those presented last indicate items that capture higher levels of tinnitus impact. Item maps were used in the assessment of the second order structure.

The Person Separation Index (PSI) examines the extent to which combined pooled items within the questionnaire can detect and separate the differences between individuals. PSI was taken to indicate the precision of each person estimate, allowing us to identify how reliably persons measured are separated, the consistency in which the person locations are ordered, as well as how a set of items conform to a unidimensional structure.

PSI use linear transformations of raw score but exclude extreme scores from the calculations since the associated SEs are exceptionally large and provide very little information about precision (Clauser & Linacre, 1999; Schumacker & Smith, 2007). PSI was calculated by adjusting the variance in each person location estimates on the logit scale for the average measurement error variance of the sample (based on individual SEs). PSI scores range 0 to 1 representing the ratio of (adjusted) true

variance (not due to error) in the sample (Schumacker & Smith, 2007). Low scores (<0.7) indicate all participants have similar scores making it hard to distinguish individual differences. High scores (>0.7) indicate the ability of the questionnaire to distinguish persons (Fisher, 1992; Lamoureux et al., 2006; Schumacker & Smith, 2007; Chen et al., 2014).

To enable assessment of different levels in ability, the PSI score were transformed to strata statistics, to estimate the number of distinct levels of person ability useful for examining and determining group differences (Fisher, 1992; Wright, 1996; Wright & Masters, 2002; Schumacker & Smith, 2007). PSI scores were first transformed into G_p with a range from zero to infinity, given by;

$$G_p = \frac{SA_p}{SE_p} = [R_p/(1 - R_p)]^{1/2} \qquad (6.4)$$

where SA_p is the person variance adjusted for error, SE_p is the root mean square measurement error and R_p is the PSI value. This G_p value was then used to calculate the number of strata levels;

$$\big(4G_p + 1\big)/3 \qquad (6.5)$$

In short, the questionnaire is considered well-targeted for the sample if (i) person and item locations correspond to each other; (ii) the majority of the data is within the information curve; (iii) the mean person location is close to zero; and (iv) the PSI value is >0.7, preferably > 0.8 (Hobart & Cano, 2009).

6.2.6.6 Assessing local independence

Local independence (assumed in Rasch) is the expectation that item pairs across persons are uncorrelated at every point along the variable (Linacre, 1998). However,

if the item pairs are correlated, then local dependency is assumed, reflecting either violations in unidimensionality or response dependency.

Response dependency where responses for the same person to one item directly influences the response on another item can result in overestimations of person locations and artificially elevated reliability estimates (Smith, 2002; Marais & Andrich, 2008). Alongside statistical tests, previous evidence and knowledge of the questionnaire responses were taken into account when assessing response dependency.

To assess potential violations in local independence, the residuals associated with each item were examined using correlation analysis, Principal Components Analysis (PCA) and independent t-tests in RUMM 2030. Item residual inter-correlations show patterns among the residuals and whether subgroups of items cluster together. Pairs of item residuals that correlate ≥ 0.2 either indicate the presence of other latent dimensions, beyond the construct the questionnaire reportedly measures, or potential response dependency (Smith, 2002; Kurtaiş et al., 2011). Residual correlations larger than the average residual correlation indicate the potential item/items that are the source of response dependency.

PCA detects any potential patterns among the item residuals, by assessing the relationships between the first component and the items in the scale. Items were divided into two subsets based on the loading values on the first component, i.e. the highest positive loading versus the highest negative loading. This is because the items with the strongest loadings in either direction will always be the set that breaches unidimensionality. Person location estimates derived from these subsets were individually compared for each person using independent t-tests.

The proportion of significant t-tests outside the range of ± 1.96 should preferably not exceed 5%, as this is a breach in the assumption of local independence, particularly unidimensionality (Smith, 2002; Tennant & Conaghan, 2007; Hagquist et al., 2009).

The test for local independency was conducted initially on the 25 items to confirm multidimensionality then on the second-order construct to confirm a unidimensional underlying construct and no response dependency. However, due to the small number of items in each of the subscales (3 to 4) there was not enough data to conduct meaningful correlations and PCA. The real concern with multidimensionality is when the response patterns indicate more than two dimensions (Linacre, 1998). The TFI subscales do not raise this concern, it is assumed that more than two dimensions is improbable based on the small number of items in each subscale and the fact that there is no previous evidence of response dependency or multidimensionality (Chapters 4 and 5). Unless the individual fit item residuals indicated otherwise (i.e. high negative residuals indicate dependency), local independence of these subscales was assumed.

6.2.6.7 Assessing the Differential Item Functioning between subsets of individuals

Differential Item Functioning (DIF) occurs when the probability of endorsing an item significantly differs between groups of persons despite having equal levels of the underlying trait being measured (Hays et al., 2000; Lamoureux et al., 2006). Two types of DIF, uniform and non-uniform DIF were examined. Uniform DIF is detected when group responses to an item consistently and systematically differ across every level of the health problem (i.e. tinnitus impact) being measured, whilst non-uniform DIF is detected when these differences between the groups vary across

every level of the health problem (Hagquist, 2001; Hagquist & Andrich, 2004; Tennant & Conaghan, 2007). Therefore, to identify DIF within the sample, the standardised residual of each person to each item was classified according to class interval and person factors. Comparisons between the groups were examined both graphically and statistically. Item characteristic curves were conducted and inspected to visually assess the distance of the mean class intervals for each person factor from the expected curve.

A two-way analysis of variance of standardised residuals for each person factor (class intervals × groups in person factor) was calculated to determine any significant differences between the groups in person factor. A significant main effect (person factor) irrespective of class interval represents uniform DIF whilst a significant interaction (person factor × class interval) represents non-uniform DIF. A main effect of class interval, irrespective to person factor, is equivalent to the χ^2 estimates of fit, therefore was not examined. Once again, multiple tests of fit were conducted and so Bonferroni corrections for alpha at the level of 0.05 were applied. Adjustments were based on the number of items/testlets within the subscale or second-order structure and the number of test probabilities (3): (1) main effect of person factor; (2) main effect of class interval and (3) the interaction between the two (person X class interval). For example, for subscales with three items, the criterion level for alpha were set to 0.05/(3x3) = 0.006, for the second-order construct with 6 testlets, the criterion level for alpha were set to 0.05/(6x3) = 0.003.

The person factors investigated were; (i) population (clinical, research), (ii) gender (male, female), (iii) age groups (three categories based on hearing age: >50, 51 – 69 and <70 years of age), and (iv) hearing loss. For the clinical population, this was defined as five categories based on self-reported hearing loss (no problem, small

problem, moderate problem, big problem, very big problem). For the research population, this was defined by four categories of audiometric pure tone average thresholds based on British Society of Audiology guidelines (normal hearing >20 dB; mild hearing loss 20–40 dB; moderate hearing loss 40–70 dB; severe hearing loss <71 dB) (see Table 6.4 for sample size in each person factor). These person factors related to clinically important groups in tinnitus.

6.2.6.8 Transforming the raw scores into linear interval measurement

An important element of the Rasch modelling process is the conversion of the raw ordinal TFI scores (subscales and second-order structure) into linear interval measurement points (logit scale). The implication is that the interval-level change between each raw score can be estimated, which provides a better understanding of the actual change that occurs between each ordinal measurement point and how closely these raw scores approximate the interval-level change.

In order to provide understandable scores that can be used in parametric statistical analysis and can be easily calculated in clinics and research, the logit interval scales were converted, using linear transformations, into metric estimate scores that reflect the range of the original raw scores, given by;

$$y = m + (s \times logit) \qquad (7.6)$$

where *s* is the wanted range/current range and m is the: wanted minimum score ‐ (current minimum × s). To conduct these transformations the data must conform to the Rasch model, otherwise the scores would reflect the model misfit rather than accurate change. Therefore only the subscales and second-order structure that conformed to the Rasch model were provided with the metric scores.

Table 6.4. Sample size frequencies in each person factor group.

Person factor	Person factor groups	N
Population	Clinical	255
	Research	285
Gender*	Male	356
	Female	183
Age	<50 yrs	195
	50 – 69 yrs	297
	70 + yrs	48
Self-defined hearing (Are you having any problems hearing speech or other sounds?)#	No problem	69
	Small problem	76
	Moderate	77
	Big problem	27
	Very big problem	6
BSA hearing thresholds (PTA)*	Normal hearing	181
	Mild loss	72
	Moderate loss	28
	Severe loss	3

***** missing data = 1; # Hearing question from baseline assessment in clinical population study (Chapter 4). BSA = British Society of Audiology. PTA = Pure Tone Audiometry.

It is important to note that the transformation scores provided are for use with complete data only. The location estimates and consequently the metric scores depend on the items that have been endorsed. For instance, for people who have completed the scale (e.g. all 25 items), the conversion score would be a true reflection of their level of tinnitus impact as measured by all the items. People who do not respond to all 25 items could still have the same raw score as those above, but the conversion score would be biased. They will effectively have "0" scores for the items that were not endorsed.

6.3. RESULTS

6.3.1. Response frequency distributions of raw scores

Response distributions for all raw item data (within subscales) are presented in Table 6.5 (dataset A), Table 6.6 (dataset B) and Table 6.7 (full dataset).

Table 6.5. Dataset A item response distributions for all items within designated subscales.

	% of responses within the response categories												
	0	1	2	3	4	5	6	7	8	9	10	Mean	(±SD)
INTR1	0.4	4.6	6.5	7.7	6.2	7.7	5.8	14.2	15	13.5	**18.1**	6.63	(2.81)
INTR2	0	0.8	3.1	7.7	11.9	9.2	10.8	19.2	20	10	6.9	6.45	(2.19)
INTR3	4.2	20.4	10	13.1	5.8	12.7	7.3	6.9	7.3	4.6	6.5	4.29	(2.97)
SOC4	2.7	3.8	5.4	6.9	6.2	8.8	10	8.5	13.5	7.3	**26.2**	6.64	(2.99)
SOC5	3.1	7.3	8.5	11.9	10.4	19.2	13.5	12.7	6.2	3.1	4.2	4.87	(2.46)
SOC6	0.8	5.8	4.2	8.1	4.6	13.8	10	15.4	10.4	11.2	15.4	6.34	(2.73)
COG7	10.4	9.2	12.3	8.1	6.5	10.4	13.5	12.7	7.7	3.5	5.8	4.56	(2.96)
COG8	**18.5**	9.2	10.4	10.8	8.1	11.2	9.6	8.5	7.3	3.1	3.5	3.88	(2.96)
COG9	12.3	11.9	11.2	13.5	9.6	11.2	10	6.9	6.9	3.8	2.7	3.94	(2.80)
SLP10	15	6.9	9.2	8.1	6.5	4.2	6.2	9.2	13.5	7.3	13.8	5.10	(3.52)
SLP11	**16.5**	8.8	11.2	7.7	2.7	6.5	7.3	9.6	11.2	6.9	11.2	4.74	(3.50)
SLP12	**18.1**	8.8	10	8.5	4.6	8.1	5	7.3	11.2	7.7	10.8	4.60	(3.52)
AUD13	14.2	9.6	10.8	10.4	9.2	10.8	9.6	8.5	10.4	2.7	3.8	4.16	(2.94)
AUD14	**20**	10.4	11.5	8.8	8.1	8.8	8.8	9.2	7.3	4.6	2.3	3.77	(3.02)
AUD15	**17.7**	10	8.8	9.2	10.4	8.1	7.3	8.8	7.3	6.9	5.4	4.18	(3.19)
REL16	6.2	6.2	7.7	6.2	5	7.7	10	13.5	16.5	8.5	12.7	5.88	(3.07)
REL17	6.5	8.5	7.3	6.2	6.9	8.1	11.2	15.4	12.3	8.5	9.2	5.51	(3.04)
REL18	3.1	3.8	5.4	4.2	5	5.4	5.8	10.8	14.2	12.7	**29.6**	7.08	(3.00)
QOL19	**26.9**	10	10.4	7.7	5	9.6	6.2	6.9	8.5	3.5	4.2	3.53	(3.22)
QOL20	**20.4**	10	9.2	11.5	7.7	8.1	6.2	8.8	6.2	5.4	6.5	3.96	(3.24)
QOL21	**30.8**	11.9	9.2	6.5	6.5	9.6	5.4	6.2	6.2	4.6	3.1	3.21	(3.16)
QOL22	**27.3**	13.1	11.5	6.2	7.3	6.5	5.8	8.1	6.2	3.5	4.6	3.34	(3.20)
EMO23	**16.2**	11.9	9.6	13.1	4.2	10	8.5	10	6.2	4.2	6.2	4.07	(3.13)
EMO24	9.6	10	11.5	8.1	5	15	7.3	12.3	6.5	8.1	6.5	4.73	(3.08)
EMO25	**31.5**	13.1	8.5	9.2	4.6	7.7	7.3	5	5.4	1.9	5.8	3.12	(3.19)

Bold values indicate items with the lowest (0) or highest (10) response option selected in <15% of respondents.

Table 6.6. Dataset B item response distributions for all items with designated subscales

	% of responses within the response categories												
	0	1	2	3	4	5	6	7	8	9	10	Mean	(±SD)
INTR1	0	4.7	4.7	7.5	8.2	10.8	7.5	12.2	19	9.3	**15.8**	6.50	(2.66)
INTR2	0	0.7	1.8	7.9	7.9	13.6	12.9	21.1	20.1	7.2	6.8	6.46	(2.04)
INTR3	7.2	16.1	12.2	11.1	7.5	11.8	7.5	11.5	7.5	4.3	3.2	4.20	(2.84)
SOC4	5	4.7	5.7	10.8	4.7	11.5	5.4	7.5	10	10.4	**23.7**	6.24	(3.23)
SOC5	5.4	5.4	8.6	12.9	8.2	17.9	9.7	15.1	11.5	3.2	2.2	4.90	(2.53)
SOC6	2.2	2.2	5	8.2	7.2	13.3	10.4	12.2	18.3	9	12.2	6.28	(2.61)
COG7	12.9	9	10.8	11.5	8.2	9	6.8	15.8	10.8	3.6	1.8	4.30	(2.90)
COG8	**17.2**	10.8	7.9	13.3	6.1	11.8	7.5	11.1	11.8	1.8	0.7	3.91	(2.87)
COG9	14	12.2	10	15.1	6.1	9.7	7.5	13.6	9.7	1.4	0.7	3.88	(2.77)
SLP10	14	8.6	9	6.8	8.2	7.5	3.6	13.6	11.8	5.7	11.1	4.92	(3.38)
SLP11	**17.6**	9.3	7.2	9	6.8	7.9	4.7	11.8	10	5.7	10	4.60	(3.42)
SLP12	**19**	7.5	11.5	8.6	5	8.2	3.2	12.5	8.2	6.1	9.7	4.44	(3.44)
AUD13	**19.4**	12.2	8.2	10.8	8.2	10.4	7.2	10.8	7.5	2.2	3.2	3.76	(2.98)
AUD14	**26.2**	10	9	10	9.3	6.5	9	7.9	8.6	1.4	2.2	3.40	(2.98)
AUD15	**24**	10.4	7.9	10	7.9	5.7	6.1	11.1	5.7	6.8	3.9	3.79	(3.25)
REL16	7.5	4.7	8.6	10	6.1	10	8.6	12.5	15.8	7.2	8.6	5.45	(3.01)
REL17	7.5	6.5	12.2	9.3	7.5	11.5	5.4	12.9	12.9	7.2	7.2	5.08	(3.03)
REL18	5	3.6	5	7.9	3.9	8.6	6.8	9.7	12.2	14	**23.3**	6.61	(3.11)
QOL19	**31.2**	9.7	10	9.3	3.9	9	5.4	10	5	2.5	3.6	3.21	(3.12)
QOL20	**21.1**	12.9	11.5	10.4	6.1	7.2	7.2	7.2	7.2	3.9	5.4	3.67	(3.18)
QOL21	**31.5**	14	10.4	8.6	4.7	10.8	3.2	5.7	5.4	3.2	2.2	2.87	(2.97)
QOL22	**30.8**	13.6	9.7	7.9	4.3	8.2	5	7.9	6.8	3.6	1.8	3.06	(3.07)
EMO23	**17.9**	14	15.4	9	3.2	6.1	4.3	9	10	3.9	7.2	3.91	(3.33)
EMO24	9.7	12.9	13.6	7.9	8.6	9.7	5	9	7.9	6.8	9	4.53	(3.22)
EMO25	**35.1**	9.3	11.1	7.5	3.9	9.3	3.9	5.4	6.5	4.7	3.2	3.04	(3.19)

Bold values indicate items with the lowest (0) or highest (10) response option selected in <15% of respondents.

Table 6.7. Full dataset item response distributions for all items within designated subscales

	% of responses within the response categories												
	0	1	2	3	4	5	6	7	8	9	10	Mean	(±SD)
INTR1	0.2	4.6	5.6	7.6	7.2	9.3	6.7	13.2	17.1	11.3	**16.9**	6.56	(2.73)
INTR2	0	0.7	2.4	7.8	9.8	11.5	11.9	20.2	20	8.5	6.9	6.46	(2.11)
INTR3	5.8	18.2	11.1	12.1	6.7	12.2	7.4	9.3	7.4	4.5	4.8	4.24	(2.90)
SOC4	3.9	4.3	5.6	8.9	5.4	10.2	7.6	8	11.7	8.9	**24.9**	6.43	(3.12)
SOC5	4.3	6.3	8.5	12.4	9.3	18.6	11.5	13.9	8.9	3.2	3.2	4.88	(2.49)
SOC6	1.5	3.9	4.6	8.2	5.9	13.5	10.2	13.7	14.5	10	13.7	6.31	(2.67)
COG7	11.7	9.1	11.5	9.8	7.4	9.6	10	14.3	9.3	3.5	3.7	4.43	(2.93)
COG8	**17.8**	10	9.1	12.1	7.1	11.5	8.5	9.8	9.6	2.4	2	3.89	(2.91)
COG9	13.2	12.1	10.6	14.3	7.8	10.4	8.7	10.4	8.3	2.6	1.7	3.91	(2.78)
SLP10	14.5	7.8	9.1	7.4	7.4	5.9	4.8	11.5	12.6	6.5	12.4	5.01	(3.45)
SLP11	**17.1**	9.1	9.1	8.3	4.8	7.2	5.9	10.8	10.6	6.3	10.6	4.67	(3.46)
SLP12	**18.6**	8.2	10.8	8.5	4.8	8.2	4.1	10	9.6	6.9	10.2	4.52	(3.48)
AUD13	**16.9**	10.9	9.5	10.6	8.7	10.6	8.3	9.6	8.9	2.4	3.5	3.95	(2.97)
AUD14	**23.2**	10.2	10.2	9.5	8.7	7.6	8.9	8.5	8	3	2.2	3.58	(3.00)
AUD15	**21**	10.2	8.3	9.6	9.1	6.9	6.7	10	6.5	6.9	4.6	3.98	(3.23)
REL16	6.9	5.4	8.2	8.2	5.6	8.9	9.3	13	16.1	7.8	10.6	5.66	(3.04)
REL17	7.1	7.4	9.8	7.8	7.2	9.8	8.2	14.1	12.6	7.8	8.2	5.29	(3.04)
REL18	4.1	3.7	5.2	6.1	4.5	7.1	6.3	10.2	13.2	13.4	**26.3**	6.84	(3.06)
QOL19	**29.1**	9.8	10.2	8.5	4.5	9.3	5.8	8.5	6.7	3	3.9	3.36	(3.17)
QOL20	**20.8**	11.5	10.4	10.9	6.9	7.6	6.7	8	6.7	4.6	5.9	3.81	(3.21)
QOL21	**31.2**	13	9.8	7.6	5.6	10.2	4.3	5.9	5.8	3.9	2.6	3.04	(3.06)
QOL22	**29.1**	13.4	10.6	7.1	5.8	7.4	5.4	8	6.5	3.5	3.2	3.2	(3.13)
EMO23	**17.1**	13	12.6	10.9	3.7	8	6.3	9.5	8.2	4.1	6.7	3.99	(3.23)
EMO24	9.6	11.5	12.6	8	6.9	12.2	6.1	10.6	7.2	7.4	7.8	4.63	(3.15)
EMO25	**33.4**	11.1	9.8	8.3	4.3	8.5	5.6	5.2	5.9	3.3	4.5	3.07	(3.19)

Bold values indicate items with the lowest (0) or highest (10) response option selected in <15% of respondents.

The response frequency distributions are the first indication of potential mistargeting within the subscales. Most notably, all the items within the Auditory, QoL subscales and two items in the Emotional subscale showed skewed distributions towards the lower values (floor effects). These subscales potentially result in poor targeting within the Rasch model, since this would indicate that the person locations will be lower on the scale than items. Although three other items did show some floor effects, with the largest percentage of participants selecting the lowest severity option (0), there was an even spread in the data across the other response categories. In contrast, three items (INTR1, SOC5 and REL18) showed skewed distributions towards the higher values, but the rest of the items in the subscale showed reasonably normal distributions across the trait, therefore the effect on the targeting should be small.

6.3.2. Dimensionality

All 25 items were submitted as a unidimensional structure to the Rasch model using the full dataset (n=540). The results were as expected; data fit to model expectations was poor. Chi-square for overall fit was significant (χ^2 = 846.4 (150), p >0.0001; Class interval: 7) and overall item fit residual was extreme (Mean = 0.48; SD = 4.77). The pattern of residuals revealed that the majority of the items were linked; residual correlations associated with each item grouped together representing the subscales of the TFI (Table 6.8). One item did not conform to the expected pattern. SOC4 item residual did not correlate with the expected items in the subscale or any other item in the questionnaire. This indicates this item is behaving differently to other items in the scale, either because it is not measuring the same construct or because the measurement format differs. This was kept under consideration during subscale analysis (6.3.3).

Table 6.8. Residual correlations for all 25 items

ITEM	INTR1	INTR2	INTR3	SOC4	SOC5	SOC6	COG7	COG8	COG9	SLP10	SLP11	SLP12	AUD13	AUD14	AUD15	REL16	REL17	REL18	QOL19	QOL20	QOL21	QOL22	EMO23	EMO24	EMO25
INTR1	1																								
INTR2	**0.32**	1																							
INTR3	**0.29**	0.19	1																						
SOC4	0.07	-0.06	-0.05	1																					
SOC5	-0.11	0.12	0.15	0.09	1																				
SOC6	0.11	**0.20**	0.11	0.07	**0.22**	1																			
COG7	-0.13	-0.05	-0.07	-0.13	0.16	-0.07	1																		
COG8	-0.20	-0.08	-0.10	-0.09	0.12	-0.13	**0.65**	1																	
COG9	-0.22	-0.06	-0.08	-0.06	0.17	-0.06	**0.55**	**0.61**	1																
SLP10	-0.10	-0.06	-0.05	-0.10	-0.09	-0.05	-0.11	-0.10	-0.10	1															
SLP11	-0.09	-0.05	-0.05	-0.15	-0.13	-0.06	-0.12	-0.14	-0.11	**0.77**	1														
SLP12	-0.03	-0.06	-0.02	-0.04	-0.06	-0.10	-0.16	-0.15	-0.15	**0.55**	**0.70**	1													
AUD13	0.11	-0.02	-0.17	-0.06	-0.33	-0.13	-0.20	-0.15	-0.24	-0.29	-0.31	-0.26	1												
AUD14	0.09	-0.04	-0.19	-0.06	-0.36	-0.20	-0.17	-0.12	-0.23	-0.30	-0.29	-0.20	**0.86**	1											
AUD15	0.09	-0.04	-0.21	-0.07	-0.38	-0.20	-0.18	-0.15	-0.23	-0.29	-0.28	-0.20	**0.77**	**0.85**	1										
REL16	-0.19	-0.06	-0.05	-0.09	0.02	0.07	-0.01	-0.07	0.01	0.12	0.08	0.01	-0.28	-0.34	-0.31	1									
REL17	-0.25	-0.12	-0.06	-0.09	0.06	0.03	0.10	0.01	0.12	0.12	0.09	0.03	-0.39	-0.44	-0.40	**0.65**	1								
REL18	-0.09	-0.05	-0.08	0.05	-0.06	0.02	-0.13	-0.17	-0.11	0.05	-0.02	-0.06	-0.15	-0.21	-0.16	**0.54**	**0.51**	1							
QOL19	-0.05	-0.11	-0.10	-0.13	-0.12	-0.17	-0.07	0.00	-0.04	-0.31	-0.27	-0.25	0.13	0.16	0.19	-0.12	-0.14	-0.16	1						
QOL20	-0.26	-0.12	-0.01	-0.13	0.11	-0.15	0.02	0.06	0.10	-0.28	-0.23	-0.24	-0.18	-0.16	-0.14	-0.04	0.07	-0.06	**0.39**	1					
QOL21	-0.16	-0.23	-0.04	-0.19	-0.07	-0.17	-0.01	0.06	0.00	-0.29	-0.24	-0.24	0.05	0.07	0.05	-0.18	-0.08	-0.16	**0.41**	**0.40**	1				
QOL22	-0.21	-0.20	-0.18	-0.12	-0.06	-0.22	0.18	0.19	0.12	-0.21	-0.18	-0.19	0.04	0.11	0.11	-0.19	-0.18	-0.19	0.15	0.18	**0.28**	1			
EMO23	-0.20	-0.14	-0.01	-0.06	0.18	0.03	0.00	-0.03	-0.01	-0.09	-0.07	-0.07	-0.36	-0.34	-0.35	-0.01	0.09	-0.02	-0.07	0.19	0.10	0.05	1		
EMO24	-0.20	-0.07	0.04	-0.01	**0.25**	0.14	0.06	0.00	0.12	-0.12	-0.12	-0.09	-0.39	-0.42	-0.40	0.00	0.10	-0.08	-0.09	**0.21**	0.03	0.00	**0.63**	1	
EMO25	-0.20	-0.05	0.05	-0.09	0.17	0.01	-0.02	-0.03	0.04	0.00	-0.02	-0.05	-0.26	-0.26	-0.27	-0.04	0.02	-0.09	-0.14	0.16	0.02	-0.05	**0.39**	0.38	1

n = 540; Red = high residual correlations (<0.2).

The PSI value was high (0.96) indicating that all the items do discriminate individuals, although this may be inflated due to the large amount of overlap in item content. PCA identified a pattern in the residual loadings, in which one subset comprised of strong positive loaded residuals whilst the other represented strong negative loaded residuals. The percentage of significant independent t-tests outside the acceptable range was vastly higher (33%) than the recommended value (5%). Overall, the 25 items together violate the assumptions of local independence and a multidimensional questionnaire structure was clearly apparent.

6.3.3. Subscales analysis

Analysis results for the eight TFI subscales are the same in both datasets (A and B), unless otherwise stated. For both datasets, the numbers of class intervals were 4, which guaranteed that more than 50 persons were in each class interval.

6.3.3.1 Overall summary fit statistics

The overall summary fit statistics are presented in Table 6.9. Inspection of the fit of data revealed non-significant item-trait interactions (χ^2) for the majority of the subscales, therefore the data at this level conforms to the model expectations. Having said this, the item-trait interactions for the Sense of control subscale (χ^2 = 21.17 (9), p = 0.01) in dataset A, the Emotional subscale (χ^2 = 21.86 (9), p = 0.009) in dataset B were significantly larger than the other subscales. These large χ^2 values may indicate problems of under- or over-discrimination at individual item level.

Inspection of the item-person interaction revealed that the residual mean value for persons is reasonably close to zero for all of the subscales and the SD are within the critical values, therefore at this level these statistics do not indicate any serious deviations from model expectations.

Table 6.9. Overall summary fit statistics for the eight TFI subscales using datasets A and B

			Item fit residual		Person fit residual		Item-trait interaction		
	Subscale	Items	Mean	SD	Mean	SD	χ^2 (*df*)	*p**	No ext
Dataset A (n = 261)	Intrusiveness	INTR1,INTR2, INTR3	-0.08	1.48	-0.53	1.00	10.00 (9)	0.675	5
	Sense of Control	SOC4,SOC5, SOC5	0.16	0.62	-0.50	1.05	**21.17 (9)**	**0.036**	12
	Cognition	COG7,COG8, COG9	0.16	0.75	-0.56	1.06	8.85 (9)	1.000	26
	Sleep	SLP10,SLP11, SLP12	0.01	1.73	-0.61	1.10	10.99 (9)	0.828	48
	Auditory	AUD13,AUD14 AUD15	-0.05	**2.38**	-0.79	1.20	12.56 (9)	0.551	33
	Relaxation	REL16,REL17, REL18	-0.52	1.25	-0.52	0.91	12.40 (9)	0.575	25
	QoL	QOL19,QOL20 QOL21,QOL22	0.09	1.60	-0.60	1.22	22.50 (12)	0.129	34
	Emotional	EMO23,EMO24 EMO25	-0.2	0.88	-0.47	0.94	10.92 (9)	0.843	30
	Sense of control-ord	SOC4,SOC5, SOC5	-0.28	**2.13**	-0.49	0.98	15.64 (9)	0.225	12
Dataset B (n = 279)	Intrusiveness	INTR1,INTR2, INTR3	-0.13	0.07	-0.48	1.03	15.41 (9)	0.240	6
	Sense of Control	SOC4,SOC5, SOC5	-0.08	1.16	-0.51	1.06	13.05 (9)	0.481	7
	Cognition	COG7,COG8, COG9	0.15	1.36	-0.77	1.18	12.95 (9)	0.494	25
	Sleep	SLP10,SLP11, SLP12	-0.30	1.76	-0.79	1.19	14.51 (9)	0.315	49
	Auditory	AUD13,AUD14 AUD15	-0.06	**2.01**	-0.7	1.05	7.04 (9)	1.000	57
	Relaxation	REL16,REL17, REL18	-0.08	0.79	-0.5	0.90	7.33 (9)	1.000	28
	QoL	QOL19,QOL20 QOL21,QOL22	-0.16	1.46	-0.60	1.30	21.35 (12)	0.182	35
	Emotional	EMO23,EMO24 EMO25	-0.44	1.08	-0.54	0.99	**21.86 (9)**	**0.028**	28
	Sense of control-ord	SOC4,SOC5, SOC5	-0.13	**2.05**	-0.52	1.04	11.63 (9)	0.705	7

* corrected for multiple comparisons. Bold = exceed recommended criterion; Underlined = marginally below criteria; No ext = number of extremes; χ^2 = Chi-square.

The residual mean for items in each subscale were also close to zero, and for the most part the SD also reflects good fit to model expectations for subscales. However, the residual SD was unexpectedly large for the auditory subscale (>2) across both datasets, indicating misfit to model expectations at item level. For both the Sleep and QoL subscales (dataset A only) SDs are ≥ 1.5, suggesting problems with some items within these subscales.

6.3.3.2 Individual item fit statistics

Individual item fit statistics were examined in each subscale for both datasets A (Table 6.10) and B (Table 6.11). There were some small variations, but most were within acceptable criteria. In fact, all the items in the Intrusiveness, Sense of Control, Cognition, and Relaxation subscales conformed to model expectation.

The χ^2 estimates and fit residuals were within the acceptable range indicating that the observed scores for each item closely adhered to the expected scores. Item characteristic curves for the items confirmed the good fit to model expectations and although, the class intervals for SOC4 do indicate a slightly flatter curve, all intervals were reasonably well distributed along the curve (see Figure 6.1 for example of one item from each subscale). That said, it should be noted that three out of the four class intervals for COG7 were located below zero possibly highlighting some mistargeting with the persons located below the item locations.

Despite being flagged in the overall item-person interaction as having potential problematic items, the item data in both the Sleep and QoL subscales fitted the Rasch model. However, both subscales had one item with residuals that only just fall within the range in the different datasets.

Table 6.10. Summary of item fit statistics for dataset A

	Item	Location	SE	Fit residual	χ^2	df	*p**
Intrusiveness	INTR1	-0.18	0.04	-0.84	5.90	3	0.349
	INTR2	-0.41	0.04	1.64	0.45	3	1.000
	INTR3	0.59	0.04	-1.01	3.84	3	0.839
Sense of Control	SOC4	-0.26	0.03	0.88	3.96	3	0.797
	SOC5	0.51	0.04	-0.28	9.23	3	0.079
	SOC6	-0.26	0.04	-0.11	7.98	3	0.139
Sense of control - ordered	SOC4	-0.47	0.06	2.18	4.25	3	0.708
	SOC5	0.67	0.04	-1.52	5.99	3	0.337
	SOC6	-0.20	0.04	-1.50	5.40	3	0.433
Cognition	COG7	-0.46	0.06	-0.02	1.53	3	1.000
	COG8	0.27	0.06	-0.50	4.92	3	0.534
	COG9	0.19	0.06	0.98	2.40	3	1.000
Sleep	SLP10	-0.19	0.05	1.55	0.98	3	1.000
	SLP11	0.06	0.05	-1.87	8.40	3	0.116
	SLP12	0.13	0.05	0.34	1.62	3	1.000
Auditory	AUD13	-0.11	0.06	1.59	5.43	3	0.428
	AUD14	0.30	0.06	**-2.78**	5.29	3	0.455
	AUD15	-0.19	0.06	1.05	1.84	3	1.000
Relaxation	REL16	0.19	0.05	-1.54	5.66	3	0.387
	REL17	0.45	0.05	-0.88	5.16	3	0.482
	REL18	-0.64	0.05	0.87	1.58	3	1.000
QoL	QOL19	0.03	0.04	-0.55	6.29	3	0.394
	QOL20	-0.25	0.04	-0.08	5.63	3	0.525
	QOL21	0.17	0.04	-1.36	9.88	3	0.079
	QOL22	0.05	0.04	2.36	0.72	3	1.000
Emotional	EMO23	-0.03	0.05	-0.55	5.61	3	0.397
	EMO24	-0.42	0.05	-0.86	5.00	3	0.515
	EMO25	0.45	0.05	0.80	0.31	3	1.000

* corrected for multiple comparisons. n= 261. Bold = exceed recommended criterion. Underlined = marginally below criteria. χ^2 = Chi-square. Sense of control-ordered = Sense of control with ordered thresholds. Class interval = 4.

Table 6.11. Summary of item fit statistics for dataset B.

	Item	Location	SE	Fit residual	χ^2	df	*p**
Intrusiveness	INTR1	-0.27	0.04	-0.08	4.30	3	0.692
	INTR2	-0.54	0.05	-0.12	5.75	3	0.373
	INTR3	0.81	0.04	-0.21	5.37	3	0.441
Sense of Control	SOC4	-0.13	0.03	0.37	2.13	3	1.000
	SOC5	0.39	0.04	-1.41	9.43	3	0.072
	SOC6	-0.26	0.04	0.78	1.49	3	1.000
Sense of control - ordered	SOC4	-0.29	0.06	1.99	3.29	3	1.000
	SOC5	0.51	0.04	-2.10	6.11	3	0.320
	SOC6	-0.22	0.04	-0.29	2.23	3	1.000
Cognition	COG7	-0.33	0.06	0.14	5.32	3	0.450
	COG8	0.15	0.06	-1.21	6.96	3	0.220
	COG9	0.19	0.06	1.51	0.68	3	1.000
Sleep	SLP10	-0.21	0.05	0.51	4.14	3	0.740
	SLP11	0.05	0.05	-2.32	7.24	3	0.194
	SLP12	0.15	0.05	0.90	3.13	3	1.000
Auditory	AUD13	-0.18	0.06	1.11	0.04	3	1.000
	AUD14	0.41	0.06	**-2.38**	5.37	3	0.440
	AUD15	-0.23	0.06	1.10	1.63	3	1.000
Relaxation	REL16	0.22	0.05	-0.90	5.60	3	0.399
	REL17	0.43	0.05	-0.04	1.35	3	1.000
	REL18	-0.65	0.05	0.68	0.38	3	1.000
QoL	QOL19	-0.01	0.04	0.53	3.07	3	1.000
	QOL20	-0.25	0.04	-1.57	7.40	3	0.240
	QOL21	0.16	0.04	-1.15	6.20	3	0.409
	QOL22	0.10	0.04	1.57	4.67	3	0.789
Emotional	EMO23	-0.04	0.04	-0.36	**10.28**	3	**0.049**
	EMO24	-0.44	0.04	-1.56	4.65	3	0.598
	EMO25	0.48	0.04	0.60	6.93	3	0.223

* corrected for multiple comparisons. n= 279. Bold = exceed recommended criterion. Underlined = marginally below criteria. χ^2 = Chi-square. Sense of control-ordered = Sense of control with ordered thresholds. Class interval = 4.

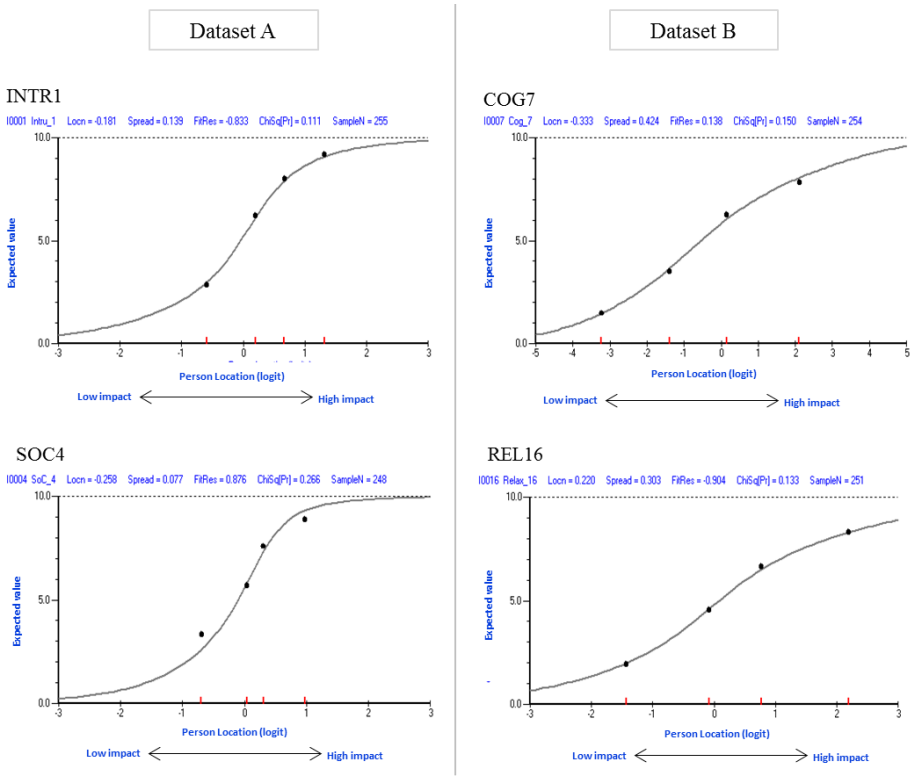

Figure 6.1. Item characteristic curves for example items (INTR1, COG7, SOC4, REL16) from the subscales with acceptable fit.
The close alignment of mean observed score class intervals (●) against the expected sigmoidal curve indicate reasonably good fit to model expectations. Although acceptable, the class intervals for SOC4 show some discrepancy in the model fit that was not reflected in the fit statistics. The flatter curve for COG7 indicates a larger range in person-item distribution. The person locations (logits) represents the continuum of tinnitus impact, with lower impact below "0" logits and higher above "0" logits. ● = mean class interval observed scores.

In dataset A, the fit residual for QOL22 ("*difficulty performing your work or other tasks*") was +2.4 and in dataset B, the fit residual for SLP11 ("*difficulty in getting as much sleep as needed*") was -2.3, indicating misfit (Tables 6.10 and 6.11). Although all class intervals for QOL22 were distributed along the curve, the majority were below zero, indicating some mistargeting (Figure 6.2). There was also a noticeable discrepancy between the observed scores and expected with one class interval located below the curve. This suggests that QOL22 is under-discriminating and potentially measuring an alternative construct than the other items in the subscale.

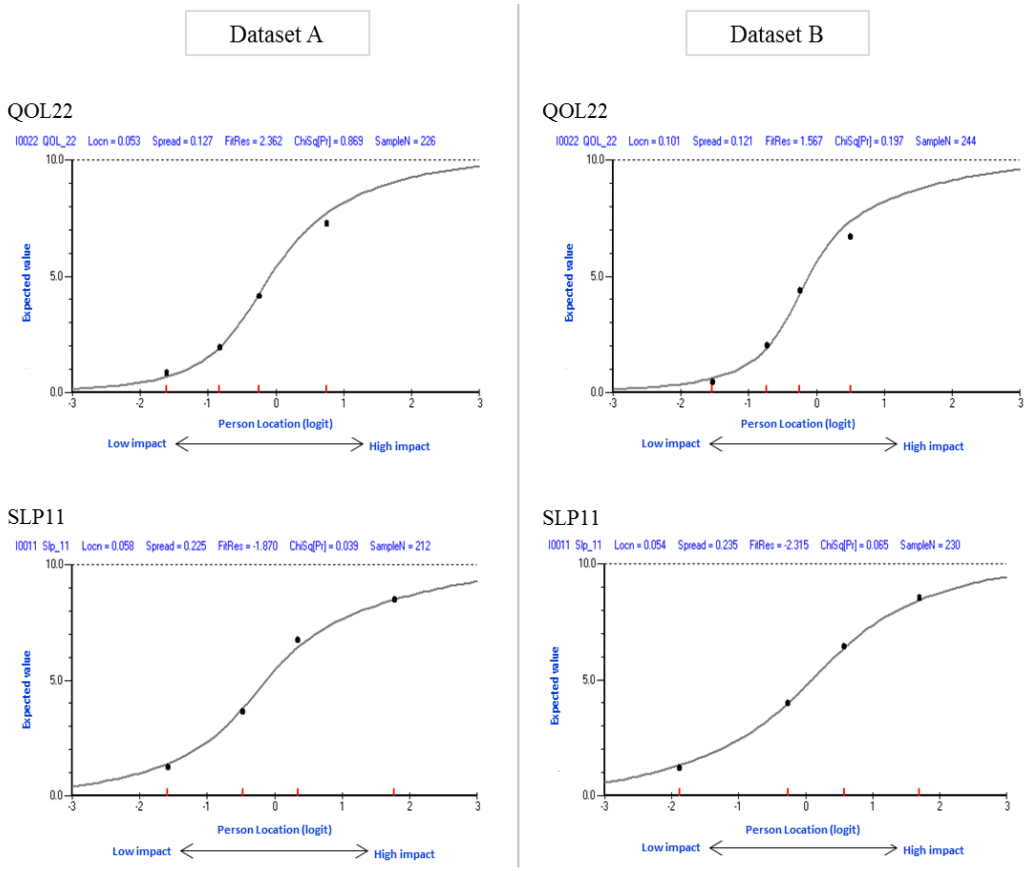

Figure 6.2. Item Characteristic curves for QOL22 and SLP11 that were flagged at summary fit for potential bad fit at item level.
The mean class intervals for QOL22 showed a slightly flatter curve to the observed scores than the expected sigmoidal curve indicating discrepancy in the model fit. The mean class intervals for SLP11 are closely aligned to the expected curve, indicating reasonably good fit to model expectations. The person locations (logits) represents the continuum of tinnitus impact, with lower impact below "0" logits and higher above "0" logits. ● = mean class interval values.

The discrepancy for SLP11 was visibly small (Figure 6.2), the class intervals were well-distributed and close to the expected curve and comparable to the items that did not show large residuals (see Figure 6.1). Items in the Auditory subscale all had non-significant χ^2 estimates. The fit residuals for two items were within the criteria, whilst AUD14 ("*ability to understand people who are talking*") item residuals were outside the established criteria range in dataset A (-2.8) and were just within dataset B (-2.4). Although, this discrepancy visually appears to be relatively small in both datasets (Figure 6.3).

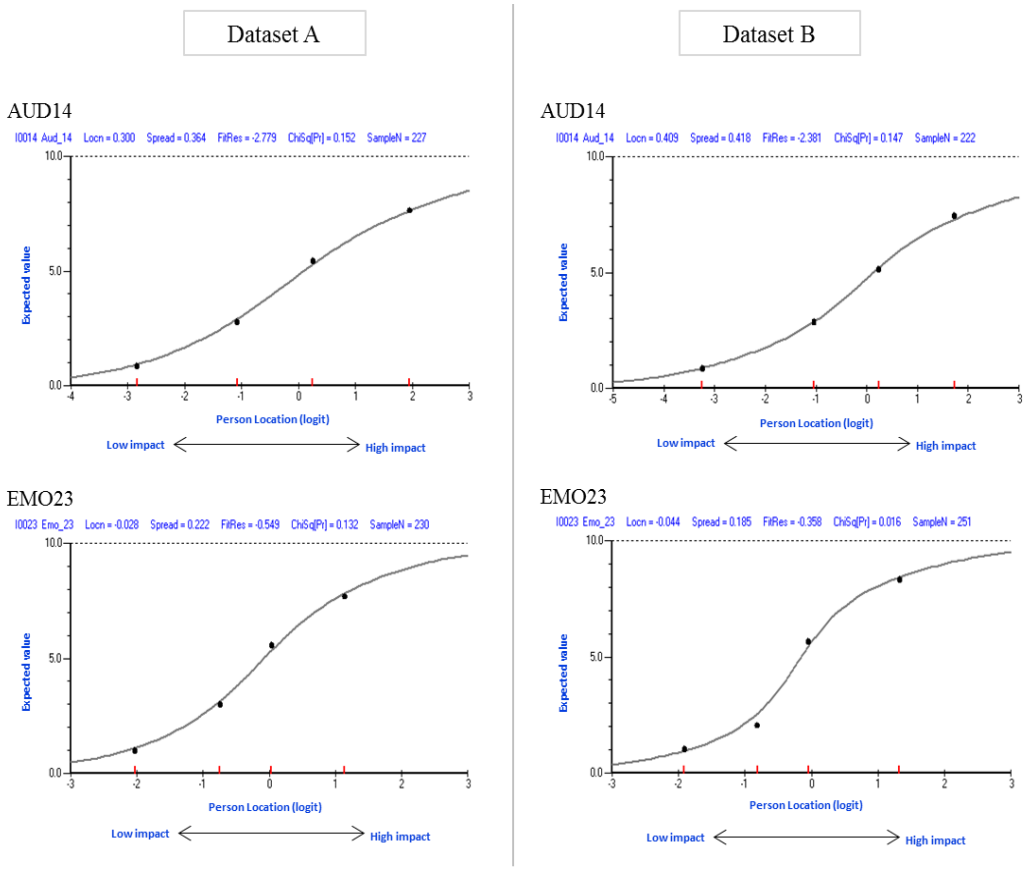

Figure 6.3. Item characteristic curves for the AUD14 and EMO23 that were flagged at summary fit for deviating from the Rasch model at item level.
The high fit residual observed for AUD14 is not apparent in the alignment of the class intervals to the expected curve. EMO23 clearly shows one class interval that was not conforming to the model expectations indicating possible misfit. The person locations (logits) represents the continuum of tinnitus impact, with lower impact below "0" logits and higher above "0" logits. ● = mean class interval values

These high negative residuals suggest that this item is potentially not contributing anything extra to the subscale but is artificially inflating the subscale score. There is overlap in content of the items that could indicate some response dependency, with responses for AUD14 potentially depending on the responses to AUD13 ("*ability to hear clearly*").

The Emotional subscale was the only scale with an item that displayed a significantly large χ^2 estimate (dataset B); all fit residuals were within the acceptable range. EMO23 ("*how anxious or worried*") showed borderline deviations from the expected scores. Inspection of Figure 6.3 showed a small anomaly for one class

interval, in which the observed responses were not conforming to the expected probabilities. The class interval is located at the lower bound of the logit scale (i.e. milder tinnitus) and the response pattern could in fact be a reflection of the distributions (floor effects) that were observed above (6.3.1). This suggests that the item is poorly targeted, at this level, in which participants are scoring slightly lower than expected. At this point in the analysis, as per *a priori* criterion, none of the items are removed from their designated subscale.

6.3.3.3 Individual person-fit statistics

Individual person fit statistics revealed that in each subscale a small percentage of participants had negative residuals outside the acceptable range (Table 6.12). Interestingly, the QoL subscale had the highest percentage across the two datasets, reflecting misfit. No participants had positive residuals outside the range. None of the participants were removed from the analysis as these deviations could reflect the nature occurrences in tinnitus.

6.3.3.4 Validity of category response ordering (thresholds)

The ordering of category responses was examined for the items within their designated subscale in both dataset A (Figure 6.4) and dataset B (Figure 6.5). The pattern of item thresholds indicated that for all items except SOC4 and QOL22, category responses were working as intended and all were utilised. The thresholds order directly corresponded to the order of response categories. Persons located at the lower (negative) points on the logit scale (low TFI scores) have a high probability of scoring the lowest value on the item, whilst persons with high severity scores have a high probability of scoring the highest values.

Table 6.12. Extreme person fit residuals in each subscale for both datasets.

Subscale	No of extremes residual (%)	
	Dataset A	Dataset B
	- 2.5	- 2.5
Intrusiveness	8 (3)	12 (4)
Sense of control	7 (3)	16 (6)
Cognition	6 (2)	–
Sleep	5 (2)	24 (9)
Auditory	15 (6)	9 (3)
Relaxation	5 (2)	7 (3)
Quality of life	17 (7)	20 (7)
Emotional	3 (1)	8 (3)

For the ordered items, each response category threshold was a defined point along the continuum, in which the probability of attaining a higher category score increased as overall tinnitus impact increased (Figures 6.4 and 6.5). For example, category threshold curves show that COG9 has working category thresholds that are defined and systematically ordered along the continuum, i.e. $\tau 1$ is followed by $\tau 2$, which is followed by $\tau 3$ and so on (Figure 6.6). The shape and location differed slightly between items and datasets, with some items the shape of the curves are steeper and the thresholds are less evenly distributed across the continuum, but are still ordered (Figure 6.6). EMO25 item, for example, displays ordered thresholds, but the range of threshold locations along the continuum are smaller, with the majority of early response category thresholds clustering together around the same location (0 logits; Figure 6.6). This indicates that the majority of the items and persons are located centrally with extremes at the lower bounds, corresponding with floor effects in the item (Tables 6.5 and 6.6).

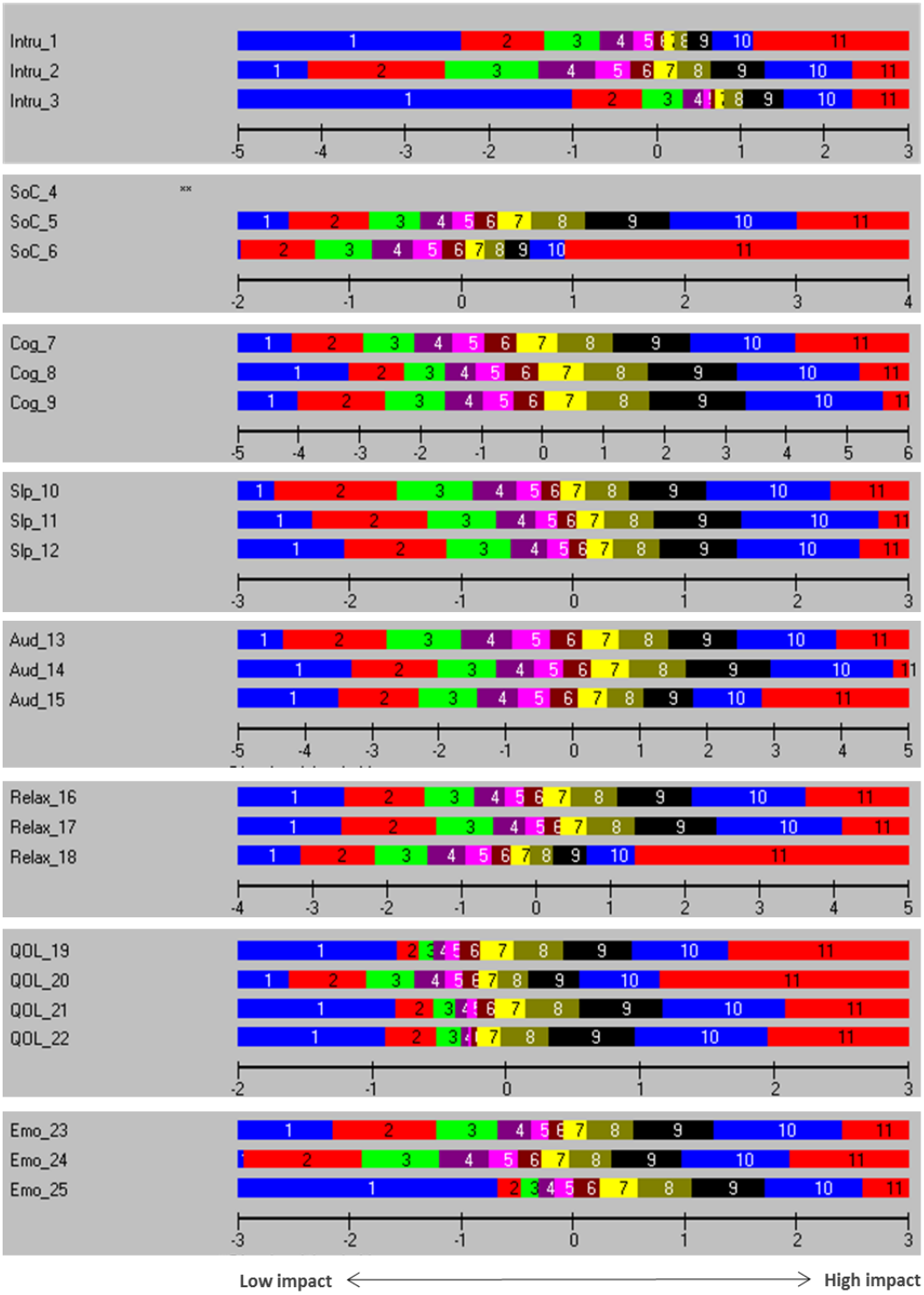

Figure 6.4. Thresholds distribution for all items within designated subscales using dataset A.

** disordered category thresholds. Logit scale continuum presented below each subscale, with lower impact indicated by negative logit values and higher by positive logit values.

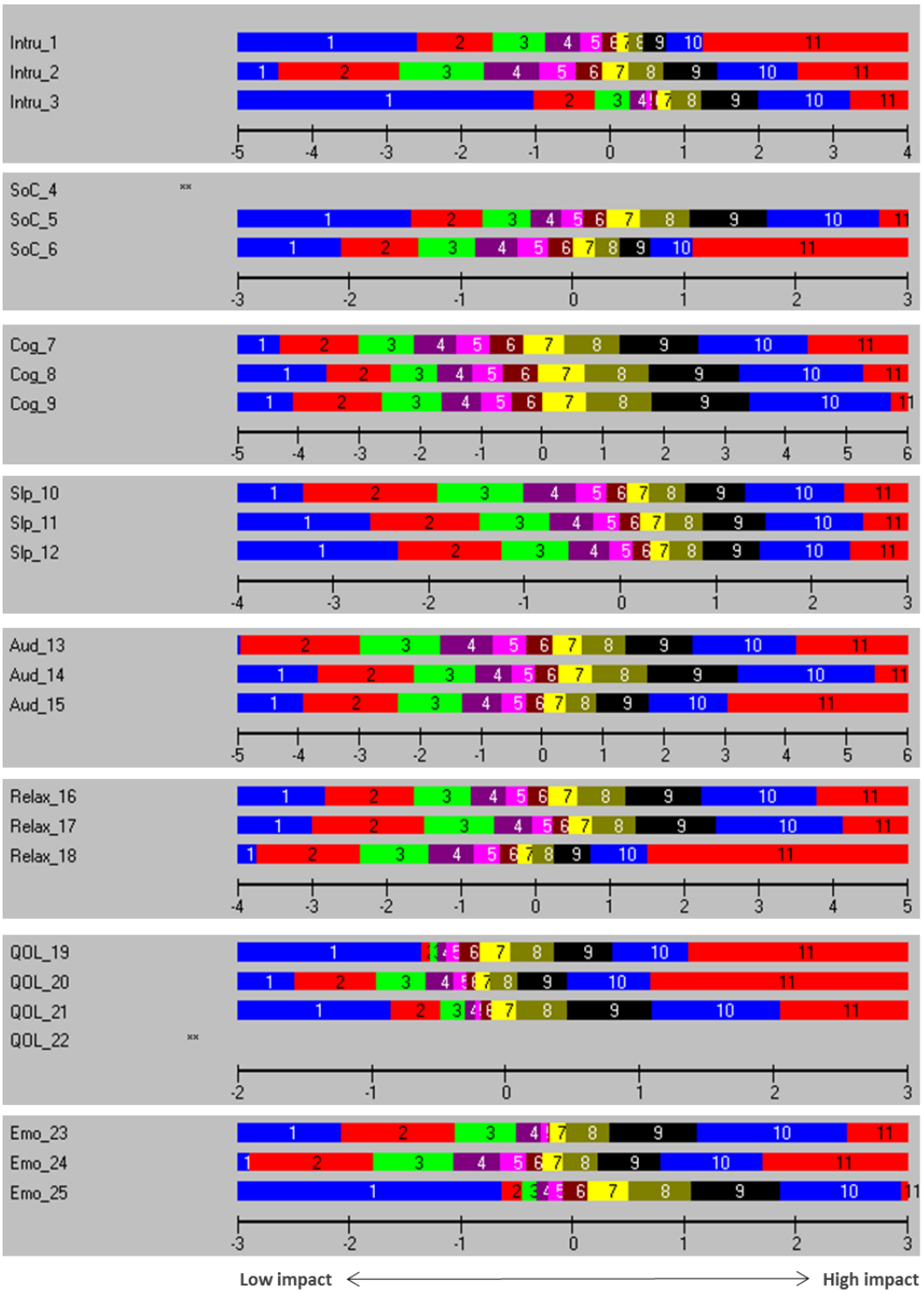

Figure 6.5. Thresholds distribution for all items within designated subscales using dataset B.

** disordered category thresholds. Logit scale continuum presented below each subscale, with lower impact indicated by negative logit values and higher by positive logit values.

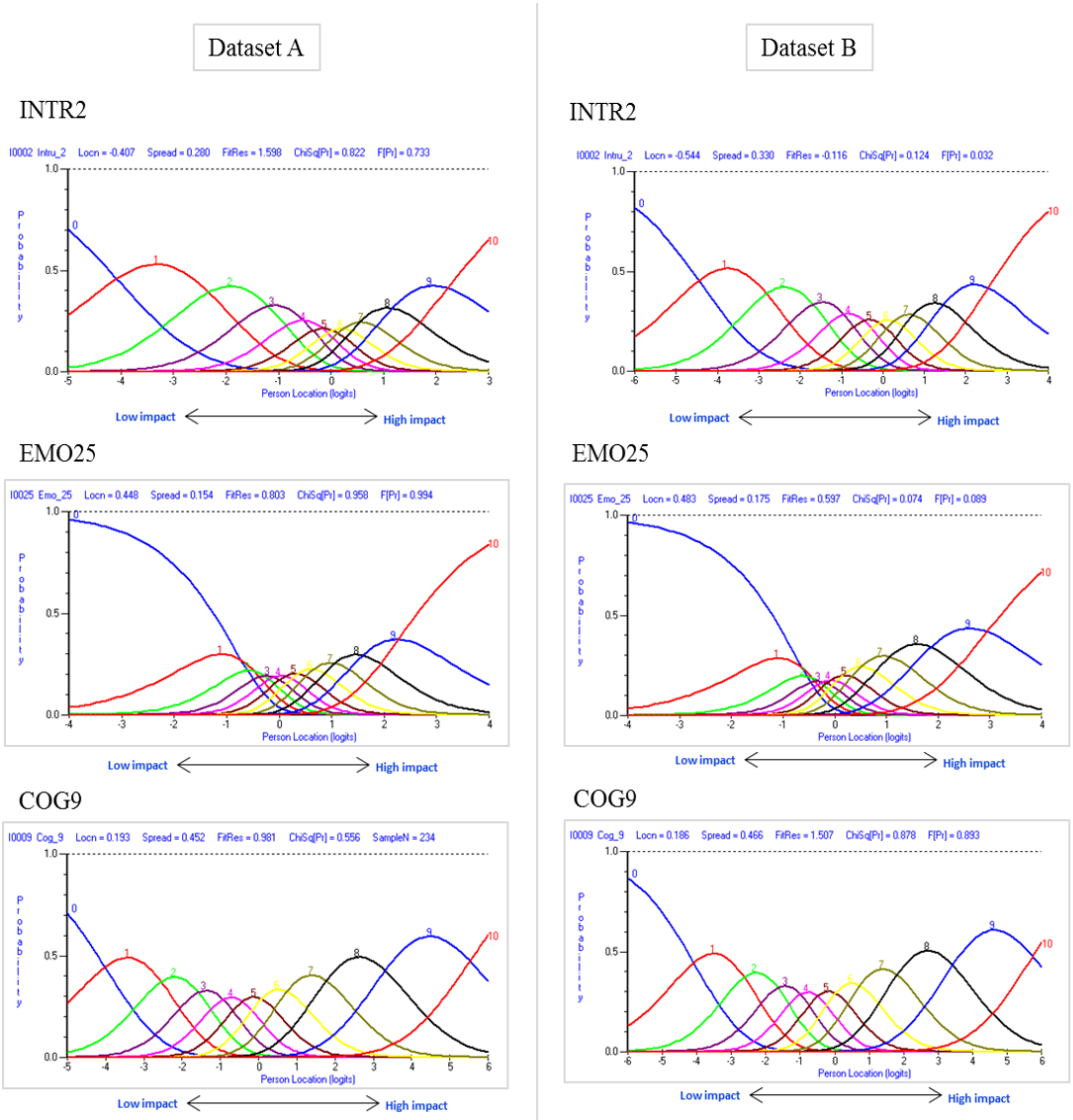

Figure 6.6. Category response curves showing ordered thresholds for INTR2, COG9, and EMO25 in datasets A and B.
The different coloured threshold curves represent the 11 response categories, all of which are ordered indicating that each category has the highest probability of being endorsed at some point. INTR2 and COG9 thresholds are reasonably evenly distributed, whilst the thresholds for EMO25 are mainly located centrally. Person location (logits) represent the continuum of tinnitus impact, with response curves located below "0" logits indicating lower levels of impact and those above "0" logits indicating higher levels of impact.

The threshold estimates for SOC4 and QOL22 were not ordered sequentially (Figure 6.7). SOC4 showed disordered threshold estimates in both datasets; the threshold estimates for τ9, τ6, τ7, and τ8 were all more than the estimates for τ10 (Tables 6.13 and 6.14). Participants were unable to reliably discriminant between the higher severity levels in the categories.

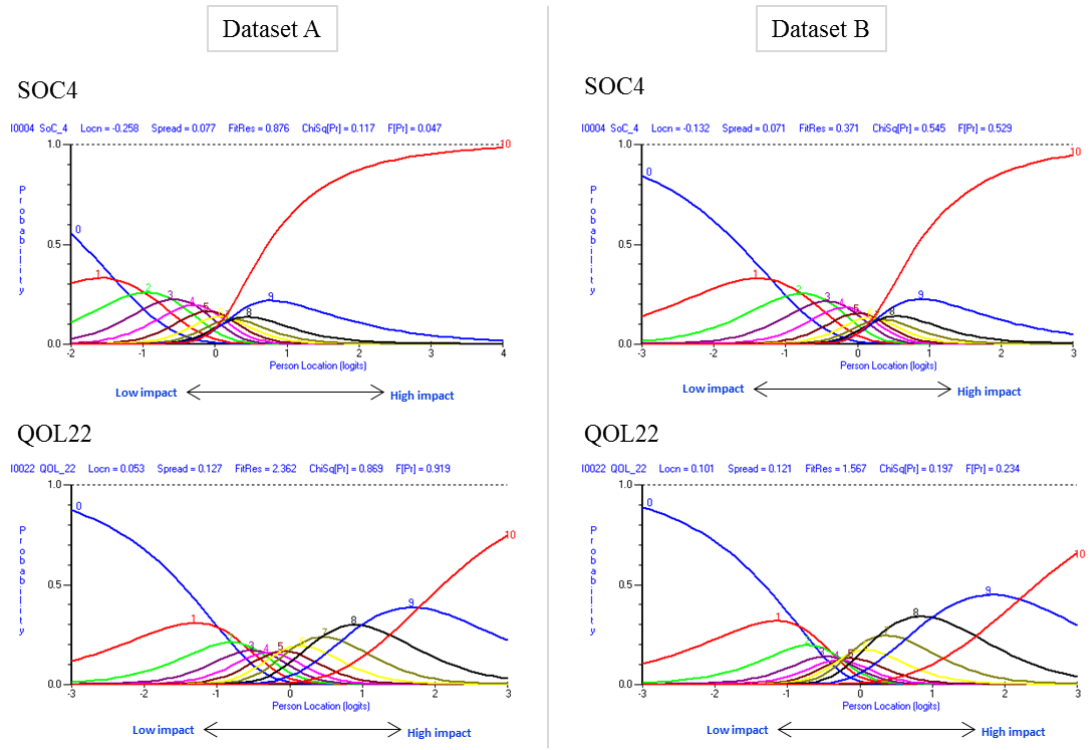

Figure 6.7. Category response curves showing disordered thresholds for SOC4 in datasets A and B and QOL22 in dataset B.
The higher categories thresholds were disordered for SOC4, with the red curve for the last category overlapping all the last four thresholds indicating that participants were unable to discriminant higher levels of severity. For QOL22, the middle thresholds were disordered indicating that tinnitus either impacts on the ability to do tasks or does not. Person location (logits) represent the continuum of tinnitus impact, with response curves located below "0" logits indicating lower levels of impact and those above "0" logits indicating higher levels of impact.

It appears that participants either felt that they were reasonably in control of their tinnitus or they were not in control of their tinnitus at all.

For QOL22, threshold estimates in dataset B were less for τ6 (-0.29) than the estimates for τ5 (-0.28), τ3 (-0.24) and τ4 (-0.23). There was a lower probability of scoring either 2, 3, 4 or 5 than the probability of scoring higher (6 to 10), therefore participants either felt their tinnitus does not impact on their ability to do tasks or that it does definitely impacts (moderately to severely) on their ability. Once again, QOL22 showed misfit to the Rasch model. These results suggest that the lower severity categories are hard to differentiate.

Table 6.13. Summary of item threshold estimates for all items within subscales in dataset A.

Code	Loc	*t*1	*t*2	*t*3	*t*4	*t*5	*t*6	*t*7	*t*8	*t*9	*t*10
INTR1	-0.18	-2.13	-1.16	-0.53	-0.17	0.01	0.09	0.16	0.29	0.56	1.07
INTR2	-0.41	-4.08	-2.33	-1.20	-0.52	-0.16	0.04	0.23	0.55	1.17	2.22
INTR3	0.59	-1.04	-0.17	0.36	0.63	0.74	0.75	0.78	0.89	1.19	1.75
SOC4	-0.26	-1.40	-0.98	-0.60	-0.28	-0.02	**0.16**	**0.26**	**0.25**	**0.13**	-0.10
SOC5	0.51	-1.71	-0.86	-0.35	-0.06	0.12	0.31	0.64	1.22	2.18	3.64
SOC6	-0.26	-1.97	-1.27	-0.75	-0.38	-0.12	0.06	0.21	0.35	0.53	0.78
COG7	-0.46	-3.99	-2.90	-2.11	-1.52	-1.02	-0.49	0.16	1.05	2.28	3.97
COG8	0.27	-2.88	-2.09	-1.51	-1.05	-0.59	-0.04	0.70	1.75	3.21	5.16
COG9	0.19	-4.02	-2.55	-1.57	-0.92	-0.43	0.07	0.75	1.77	3.31	5.52
SLP10	-0.19	-2.23	-1.33	-0.81	-0.55	-0.41	-0.30	-0.08	0.35	1.13	2.35
SLP11	0.06	-2.17	-1.19	-0.64	-0.38	-0.26	-0.15	0.10	0.63	1.57	3.07
SLP12	0.13	-1.85	-1.05	-0.58	-0.33	-0.18	-0.04	0.23	0.72	1.55	2.83
AUD13	-0.11	-3.88	-2.64	-1.71	-1.00	-0.42	0.13	0.73	1.48	2.46	3.78
AUD14	0.30	-3.06	-1.97	-1.20	-0.64	-0.19	0.28	0.87	1.68	2.82	4.41
AUD15	-0.19	-3.15	-2.28	-1.56	-0.95	-0.41	0.11	0.63	1.20	1.86	2.65
REL16	0.19	-2.34	-1.36	-0.76	-0.41	-0.19	0.03	0.39	1.00	1.99	3.50
REL17	0.45	-2.25	-1.20	-0.57	-0.21	0.01	0.24	0.63	1.31	2.43	4.14
REL18	-0.64	-2.66	-1.99	-1.47	-1.06	-0.72	-0.41	-0.11	0.23	0.64	1.16
QOL19	0.03	-1.05	-0.78	-0.60	-0.47	-0.35	-0.17	0.10	0.51	1.11	1.96
QOL20	-0.25	-1.68	-1.15	-0.79	-0.54	-0.36	-0.19	0.00	0.28	0.67	1.25
QOL21	0.17	-0.82	-0.62	-0.48	-0.37	-0.24	-0.05	0.24	0.67	1.29	2.13
QOL22	0.05	-0.98	-0.65	-0.45	-0.33	-0.24	-0.12	0.09	0.44	0.98	1.78
EMO23	-0.03	-2.30	-1.42	-0.84	-0.47	-0.21	0.02	0.31	0.75	1.43	2.44
EMO24	-0.42	-3.07	-2.02	-1.32	-0.86	-0.55	-0.29	0.02	0.48	1.18	2.23
EMO25	0.45	-0.72	-0.48	-0.28	-0.09	0.11	0.35	0.66	1.07	1.59	2.27

SOC4 (presented in red) has disordered thresholds. Bold = disordered thresholds.

Table 6.14. Summary of item threshold estimates for all items within subscales in dataset B.

Code	Loc	*t*1	*t*2	*t*3	*t*4	*t*5	*t*6	*t*7	*t*8	*t*9	*t*10
INTR1	-0.27	-2.59	-1.56	-0.85	-0.39	-0.10	0.10	0.26	0.46	0.77	1.26
INTR2	-0.54	-4.45	-2.82	-1.69	-0.94	-0.45	-0.09	0.26	0.74	1.45	2.54
INTR3	0.81	-1.01	-0.18	0.27	0.49	0.57	0.64	0.83	1.24	2.01	3.24
SOC4	-0.13	-1.20	-0.78	-0.43	-0.13	0.09	**0.24**	**0.32**	**0.31**	**0.22**	0.04
SOC5	0.39	-1.45	-0.80	-0.37	-0.10	0.11	0.32	0.61	1.06	1.75	2.76
SOC6	-0.26	-2.07	-1.38	-0.86	-0.49	-0.21	0.01	0.21	0.43	0.71	1.08
COG7	-0.33	-4.30	-3.00	-2.08	-1.41	-0.86	-0.30	0.37	1.29	2.59	4.37
COG8	0.15	-3.52	-2.47	-1.72	-1.15	-0.63	-0.05	0.70	1.77	3.25	5.28
COG9	0.19	-4.09	-2.62	-1.65	-1.00	-0.50	0.02	0.73	1.81	3.42	5.74
SLP10	-0.21	-3.31	-1.91	-1.00	-0.45	-0.13	0.08	0.31	0.68	1.32	2.35
SLP11	0.05	-2.61	-1.47	-0.73	-0.27	0.01	0.22	0.47	0.86	1.52	2.54
SLP12	0.15	-2.32	-1.23	-0.53	-0.11	0.14	0.31	0.52	0.87	1.47	2.42
AUD13	-0.18	-4.94	-2.97	-1.65	-0.81	-0.25	0.18	0.66	1.37	2.48	4.18
AUD14	0.41	-3.67	-2.08	-1.09	-0.49	-0.10	0.27	0.82	1.74	3.23	5.47
AUD15	-0.23	-3.92	-2.35	-1.31	-0.66	-0.26	0.05	0.39	0.91	1.75	3.07
REL16	0.22	-2.82	-1.62	-0.86	-0.40	-0.09	0.18	0.57	1.21	2.24	3.79
REL17	0.43	-3.00	-1.48	-0.54	-0.03	0.24	0.45	0.76	1.37	2.43	4.14
REL18	-0.65	-3.75	-2.36	-1.42	-0.82	-0.46	-0.23	-0.03	0.26	0.74	1.52
QOL19	-0.01	-0.62	-0.56	-0.51	-0.44	-0.34	-0.19	0.04	0.36	0.80	1.37
QOL20	-0.25	-1.57	-0.96	-0.59	-0.38	-0.28	-0.22	-0.11	0.09	0.47	1.08
QOL21	0.16	-0.85	-0.48	-0.30	-0.22	-0.18	-0.10	0.09	0.46	1.09	2.05
QOL22	0.10	-0.85	-0.41	**-0.24**	**-0.23**	**-0.28**	-0.29	-0.14	0.25	1.00	2.20
EMO23	-0.04	-2.07	-1.05	-0.50	-0.27	-0.22	-0.20	-0.06	0.33	1.13	2.47
EMO24	-0.44	-2.89	-1.78	-1.06	-0.64	-0.41	-0.26	-0.08	0.23	0.80	1.71
EMO25	0.48	-0.63	-0.44	-0.32	-0.22	-0.08	0.15	0.51	1.06	1.86	2.95

SOC4 and QOL22 (presented in red) have disordered thresholds. Bold = disordered thresholds.

However, the disordered thresholds were not reflected across both datasets (Table 6.13), therefore as per the *a priori* criterion these categories were not collapsed.

In terms of SOC4 disordering is apparent in both datasets A and B. Although, the item-trait interactions χ^2 estimate was only significant in dataset A (Table 6.9) this disordering could be inflating this value as the response structure is not behaving as expected. Accordingly, SOC4 response categories were collapsed into 6 response categories (0 to 5) rather than the original 10 categories (Figure 6.8). As a consequence of this, all the summary fit statistics improved, except for the item residual SD which indicated misfit at item level (Table 6.9). The individual item fit residuals for SOC4 and SOC5 (dataset B) were inflated, although they did not exceed the established range (Tables 6.10 and 6.11).

6.3.3.5 Targeting and reliability

As anticipated all the items in all subscales followed a logical order, in which the first indication of the problem is identified by the first threshold in all three/four items.

All subscales had a PSI score above criterion (Table 6.15). The G_p values and the number of distinct strata levels are given in Table 6.15. These values suggest that each subscale can reliably differentiate between individuals in relation to the trait being measured by the subscale. Examination of strata levels indicated that the subscales are capable of discerning two to four different levels of individual differences. For example, the Sense of Control subscale can reliably distinguish between two levels of person ability, whilst the Relaxation subscale can distinguish four levels of person ability.

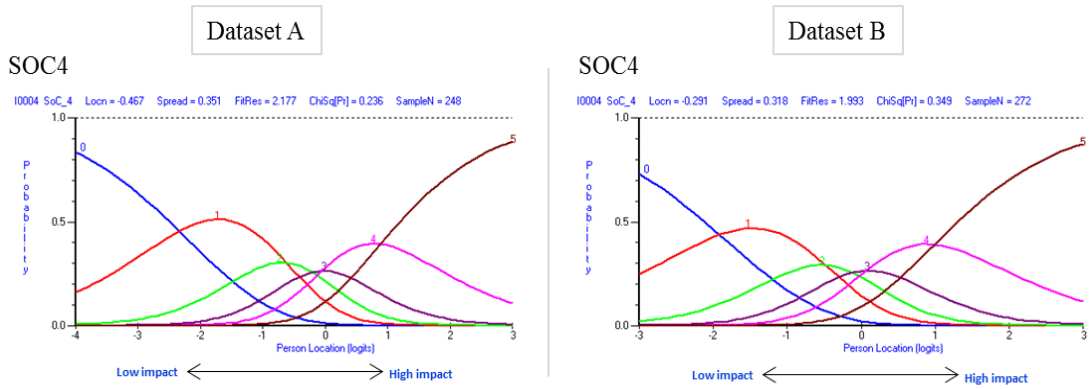

Figure 6.8. Category characteristic curves for the collapsed category thresholds for SOC4.
The 11 response categories for SOC4 were collapsed to six ordered category response thresholds, where each category has the highest probability of being endorsed at some point. Person location (logits) represent the continuum of tinnitus impact, with response curves located below "0" logits indicating lower levels of impact and those above "0" logits indicating higher levels of impact.

Logically, subscales with larger distribution of the persons across the continuum more easily differentiate between person ability levels, since there are fewer people at each measurement point. However, this number does not always reflect the large gaps in measurement points and the appropriateness of the questionnaires targeting.

6.3.3.5.1 Intrusiveness

The alignment of person location to item locations on the intrusiveness subscale is reasonably well-targeted (Figure 6.9). Only one item threshold (INTR2 τ1) did not target any person locations at -4 logits, but it does provide the initial indication of problems with intrusiveness. Otherwise the item threshold locations were reasonably well-targeted, with only small gaps between item thresholds, such as between the item locations at -1 to -2 logits. The persons cover the continuum reasonably well (-3 to +4 logits).

Table 6.15. Summary of reliability statistics, Cronbach's alpha (α), Person separation index (PSI), Gp and Strata for each subscale in both datasets.

	Subscale	α	PSI	G_p	Strata*
Dataset A	Intrusiveness	0.79	0.76	1.77	2.69
	Sense of Control	0.76	0.78	1.93	2.91
	Cognition	0.95	0.92	3.46	4.96
	Sleep	0.95	0.88	2.69	3.93
	Auditory	0.95	0.91	3.23	4.65
	Relaxation	0.93	0.88	2.75	4.00
	QoL	0.93	0.83	2.19	3.26
	Emotional	0.92	0.86	2.52	3.69
Dataset B	Intrusiveness	0.81	0.81	2.03	3.04
	Sense of Control	0.80	0.77	1.82	2.76
	Cognition	0.95	0.92	3.29	4.72
	Sleep	0.95	0.89	2.80	4.06
	Auditory	0.95	0.91	3.21	4.61
	Relaxation	0.93	0.90	2.93	4.24
	QoL	0.90	0.77	1.81	2.75
	Emotional	0.92	0.85	2.34	3.46

*Strata values should be truncated to whole numbers. α = Cronbach's alpha.

The mean person location of 0.45 logits was only slightly above the item mean location (0 logits; Figure 6.9). Only persons with extreme high scores (>3 logits) were not covered by any item thresholds, but the rest of persons were covered. Understandably, there was always less confidence in the ability estimates for persons who score at the extremes since it is impossible to say exactly how far above or below the ceiling or floor that person really lies. Therefore, the reduced measurement points in these locations are not necessarily a big problem as the reliability of differentiating between participants with extremely severe tinnitus is always limited.

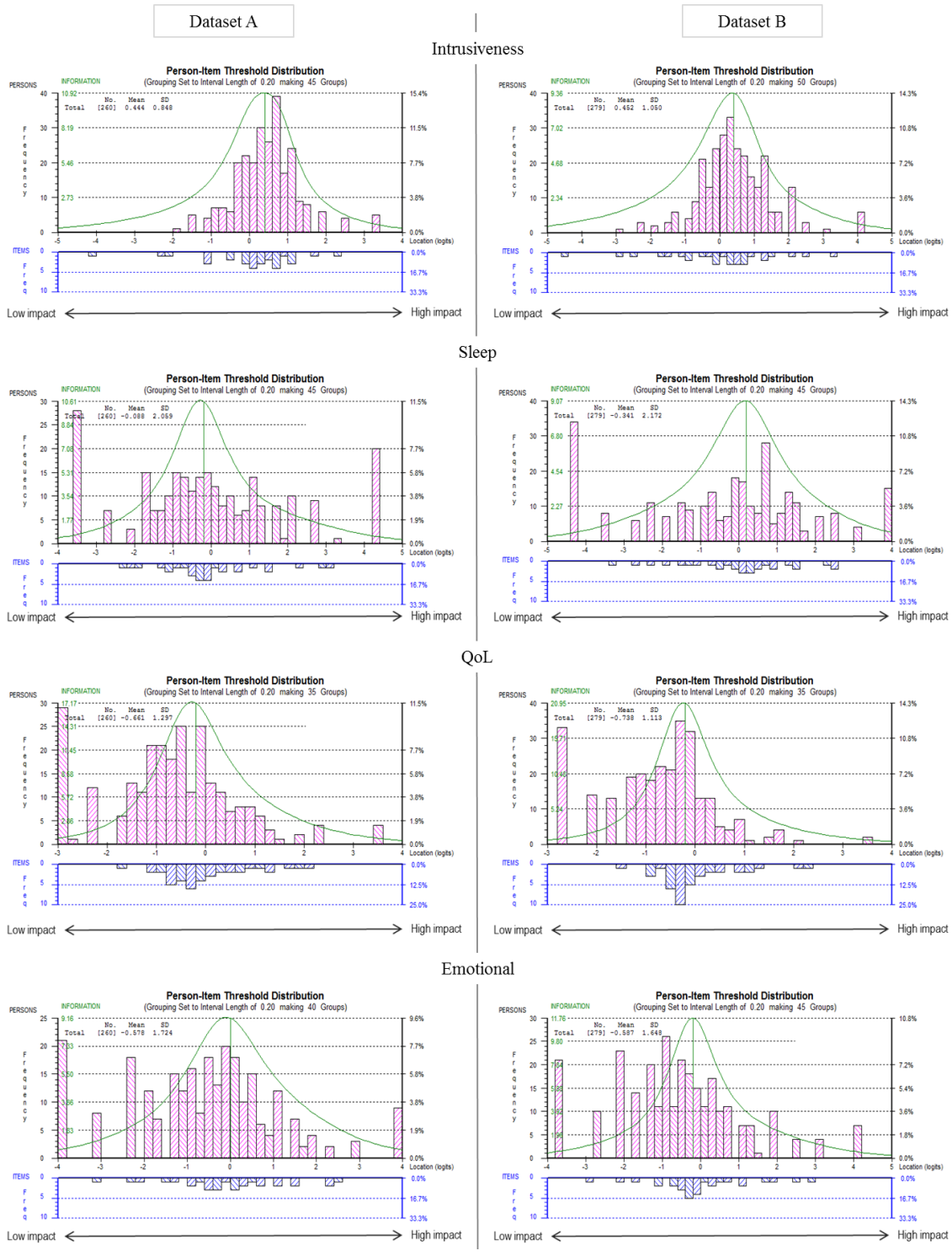

Figure 6.9. Targeted person-item distributions for Intrusiveness, Sleep, QoL and Emotional subscales in datasets A and B.
Person and item locations are reasonably well-aligned, but with extreme scores at lower end of the scale. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

More importantly, the TFI was able to identify those persons with severe tinnitus problems. The steep information curve and high value showed that the majority of the persons and items were located within the curve. Therefore there was less error and more precision of measurement associated with those locations. In summary, the persons are reasonably well-targeted and can be distinguished by the items, which are capturing the range of problems on the trait being measured.

6.3.3.5.2 Sense of Control

The person and item locations were evenly distributed along the continuum, with person locations ranging from -3 to +3 logits (Figure 6.10). Despite the disordered thresholds in one item (Figure 6.7) there was good alignment between person locations to item threshold locations. The mean person location of 0.3 was just above 0 logits (mean item location). However, having collapsed the response categories for SOC4 (Figure 6.8), the mean person location marginally increased (0.5), with the person distribution skewed towards the higher levels alignment (Figure 6.10). Furthermore, the alignment of person and item locations slightly altered with the targeting at the extremes for persons reduced. Key information could be unaccounted for because there were no corresponding measurement points for persons experiencing milder levels of tinnitus and noticeable gaps in measurement for persons experiencing severe tinnitus. In spite of this, the information value was reasonably high (8), although somewhat smaller than the unaltered subscale information curve (13). The person and item locations fit near-perfectly within the information curve within minimal error associated with the person locations.

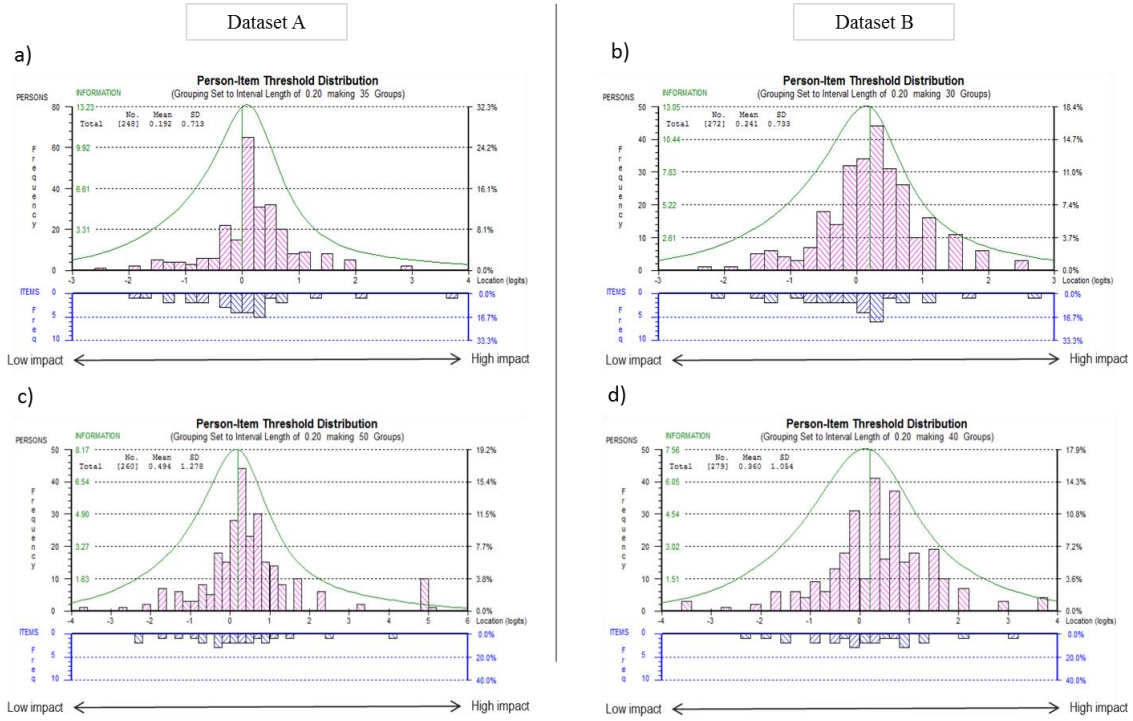

Figure 6.10. Person-item distribution for Sense of control subscale before (a/c) and after (b/d) collapsing thresholds in datasets A and B.
Person and item locations are reasonably well-aligned, although after collapsing thresholds there are more gaps at the higher end of the scale, indicating limited measurement at these points. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

Although, the loss of the additional item thresholds did reduce the measurement precision for the subscale, it still had reasonably high precision and alignment which would indicate that this subscale effectively targeted the variability on the Sense of Control domain.

6.3.3.5.3 Cognition

Despite the large spread in the person locations on the continuum (-6 to +7 logits), the distribution of the person locations was unevenly skewed towards the lower values (Figure 6.11). Measurement precision was slightly limited at the centre of the continuum, reflected in the low information values (≥5).

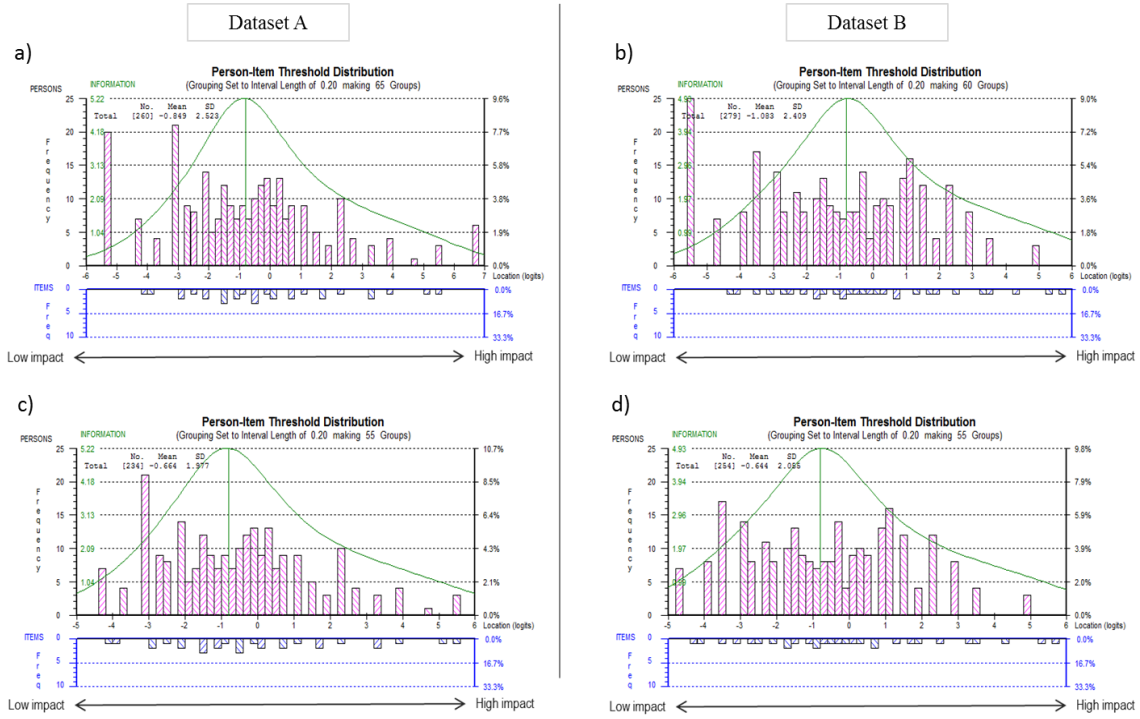

Figure 6.11. Person-item distribution for Cognition subscale with extreme score (a/b) and without (c/d) in both datasets.
The Cognition subscale covers a large range in the person locations on the continuum, but measurement precision was slightly limited at the centre of the continuum, reflected in the low information values. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

The mean person locations (-0.849 and -1.083) were less than that of the items (0), indicating potential mistargeting at these points. Participants were, on average, measuring a lower level of tinnitus impact than the items cover. Compared to the higher values, there are fewer measurement points at the lower values per number of persons. This implies that information about participants that did not have a problem with tinnitus was more limited and harder to differentiate than those who did have a problem. There were once again a number of small gaps in measurement particularly at the higher extreme values, for example between +5 and +7 logits (dataset A). Yet, when extreme scores were removed, the persons located at the higher values were generally covered by the items and there was large spread of item

thresholds across the continuum relative to the person distribution (-4 to +6 logits). The Cognition subscale is reasonably well-targeted.

6.3.3.5.4 Sleep

The Sleep subscale was also reasonably well-targeted (Figure 6.9). The person and item locations generally corresponded, with all item thresholds being utilised. Person locations were spread across the continuum (-5 to +5 logits), although dataset A did appear to have a slightly flatter person location distribution than dataset B. Again there were some small gaps in the measurement with the largest gap (+1.6 to +2.2 logits) in the higher values. Indeed, in accordance with the other subscales, the targeting at the extremes was suboptimal, with no items to differentiate individuals at these levels. There is limited information available at these locations. The mean person location is marginally below that of the items, indicating that the participants on average were experiencing sleep problems at a lower level than the measurement average (Figure 6.9). However, the impact on the targeting was relatively small, only extreme levels would be slightly harder to discriminate. The rest of the person locations were covered by items indicating a high reliability to differentiate between individuals. The majority of the sample was within the information curve, and the values were high in comparison to the other subscales, indicating good precision of measurement and little error.

6.3.3.5.5 Auditory

At first glance, the person locations appeared to be generally covered by the item threshold locations and vice-versa, the items mirror the person locations (Figure 6.12). The person locations were spread over a large range on the continuum (-5 to +6 logits). However, there was skew in the person locations.

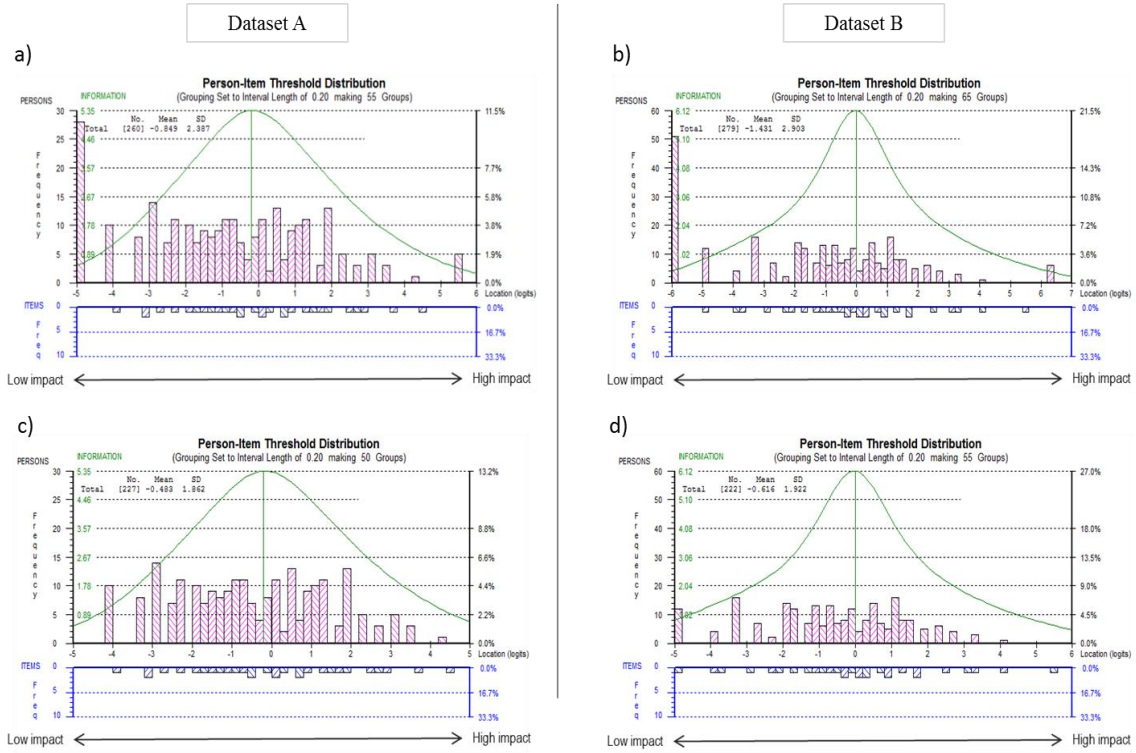

Figure 6.12. Person-item distribution for Auditory subscale with extreme score (a/b) and without (c/d) in both datasets.
The subscale is poorly targeted. There is an extremely flat distribution of the persons across the continuum indicating that the Auditory subscale may not be measuring a construct associated with tinnitus. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

The mean person locations indicated that the majority of participants were measuring lower levels of hearing problems than the average item locations. There was mistargeting which could reflect the high item residuals observed in individual item fit statistics. Additionally, although some person and item locations were within the curve (Figure 6.12), a closer inspection revealed an extremely flat distribution of the persons across the continuum. Removing the extreme scores, there was no point at which the persons or items locations peak in the expected normal distribution (Figure 6.12). The extremely flat distribution indicates poor measurement precision.

The low information value (5.35 – 6.12), in comparison to the other subscales, and the wide breath of the curve indicated large amounts of error associated with the

measurement. This distribution could also be seen in the item characteristic curves for each item in the subscale (Figure 6.3). The expected value curve was flat rather than the anticipated monotonic curve (e.g. Figure 6.1). Thus, although the high PSI value did indicate discrimination between individuals, the ability of the subscale to discriminate tinnitus severity in relation to hearing in the sample population was limited. The Auditory subscale was poorly targeted. Together, these results would indicate that the three items in the Auditory subscale are not necessarily measuring a construct associated with tinnitus at all.

6.3.3.5.6 Relaxation

The mean person location was only slightly above the mean item location in both datasets (Figure 6.13). Person locations distributions were unevenly spread along the continuum (-5 to +6 logits), and for dataset B, the locations were clearly skewed towards the higher values. Again this suggests participants were experiencing higher levels of severity than the items measure. Similar to the other subscales, there was some mistargeting between item location and person location at these higher locations, with noticeable gaps in measurement. The precision of measurement at these locations was poor. Key information was not being properly measured by the limited number of item threshold locations. By removing extreme scores, the majority of persons were covered by measurement points. However, in some cases there were a limited number of thresholds at the location measuring a large number of participants (i.e. 3.5 logits), possibly reducing discrimination at these locations. Even though the information curve values were not as high as the other subscales, the majority of the person and item locations fell within the curve in dataset A (and slightly less in dataset B).

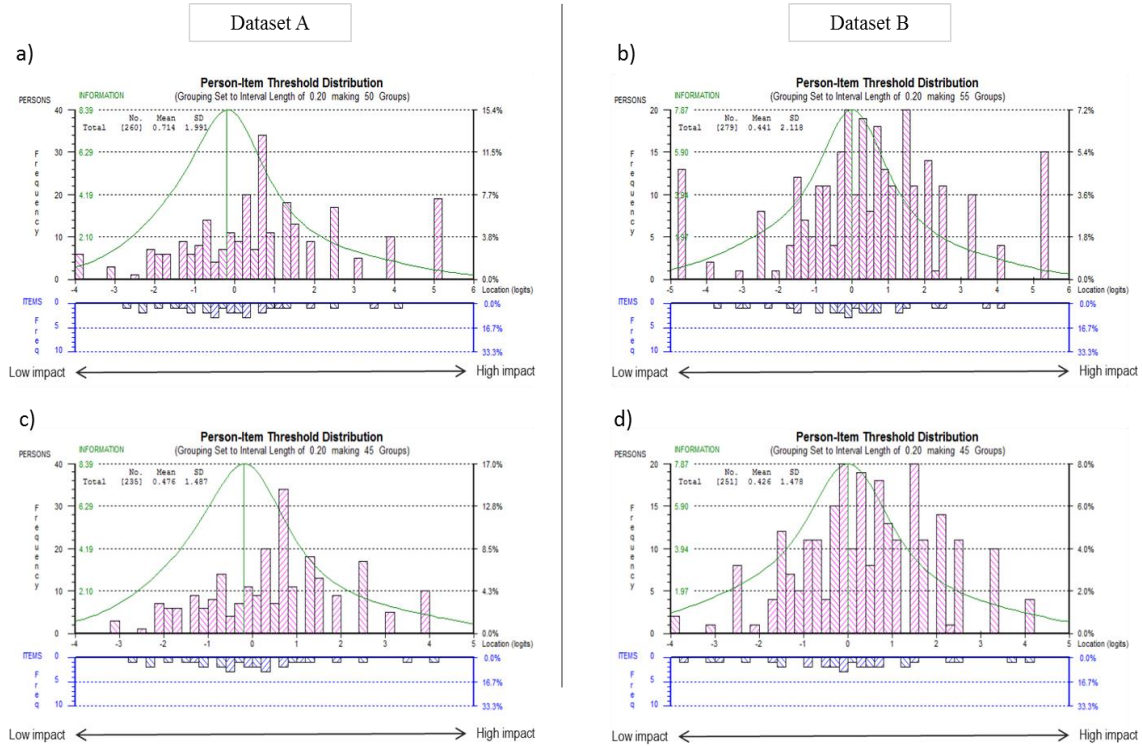

Figure 6.13. Person-item distribution for Relaxation subscale with extreme score (a/b) and without (c/d) in both datasets.
Person locations were unevenly spread along the continuum and skewed towards the higher values with misalignment with item locations, indicating mistargeting and loss of information. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

This indicates some measurement error associated with the person locations, but there was precision of measurement at the centre of the curve. Therefore the Relaxation subscale is reasonably targeted to the population.

6.3.3.5.7 Quality of life (QoL)

Despite concerns over large skew in the raw data towards the lower values in the scale, the QoL subscale was well-targeted (Figure 6.9). Person locations were distributed along the continuum (-3 to +4 logits), and the mean person location (-0.7 logits) was only slightly below the mean item location (0 logits). Therefore the person locations did in general correspond to the item locations. Nevertheless, targeting at lower extremes of the person distribution was suboptimal. There was

limited information available about lower levels of tinnitus impact, since no items are located at these levels, therefore it is harder to discriminant individuals with little problems with tinnitus. The information about the rest of sample, in particular the higher levels of tinnitus impact were covered by item thresholds with only small gaps in measurement, again at the extreme level. Item locations were therefore evenly distributed along the continuum. The information curve value was exceptionally high for both datasets (A= 17.17, B = 20.95) compared to the other subscales. The majority of persons and item locations also fell within this curve, indicating precision of measurement at these locations and therefore good targeting.

6.3.3.5.8 Emotional

Despite the high χ^2 value and floor effects in the raw data, the Emotional subscale was reasonably well-targeted (Figure 6.9). Person locations were spread along the continuum (-4 to +4 logits) and were mirrored by the item threshold locations, suggesting that all the items were being utilised. The person mean location (-0.6 logits) was slightly below the mean item location. The impact on the targeting was relatively small, only extremely mild levels of tinnitus would be slightly harder to discriminate. There were some small but noticeable gaps in the item distribution which potentially limited the measurement accuracy at these points, but the number of person locations within the gaps, were relatively small (Figure 6.9). This suggests that the inclusion of additional measurement points would not significantly improve accuracy. The information curve again had a high value (>9), with minimal error associated with the person locations, indicating that a large amount of person information was measured accurately.

6.3.4. The second-order construct of TFI

In order to assess the validity of the overall TFI underlying construct (second-order structure), all eight subscales were transformed into 'testlets' and subjected to Rasch analysis as a uniform structure. Testlets that deviated from model expectations were removed from the overall TFI structure. Following this, the TFI second-order structure was continuously reassessed until a final second-order structure was identified (Figure 6.14).

The overall summary fit statistics for the second-order models are presented in Table 6.16. The PSI values for all the models were above the established criterion (Table 6.17). Examination of the response category ordering is not reported in testlet analysis. Each testlet had thirty thresholds associated with the 10 thresholds for each of the three items. The ordering was relatively meaningless as this was not representative of the response options presented in the TFI. The numbers of class intervals were 4 for both datasets.

6.3.4.1 Evaluation of the eight-factor second-order structure

Overall summary fit statistics indicated substantial deviations from the Rasch model at item level (Table 6.16). The χ^2 estimate for item-trait interactions was significantly large ($p>0.001$) indicating substantial variance in the hierarchical ordering of items across the trait and unexpected interactions between items and persons. Similarly, the item-person interactions indicated severe deviations with an overall item fit residual SD that exceeded the recommended criterion.

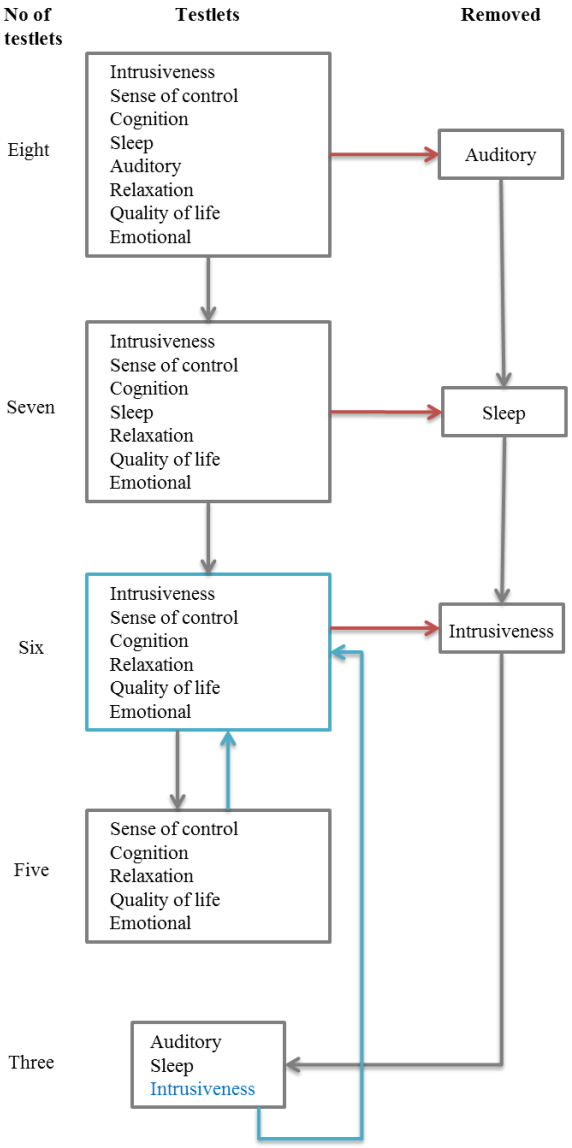

Figure 6.14. A flow diagram of the process of removing testlets from second-order structure.
Red arrows = testlets removed from analysis. Blue = the final second-order structure.

To identify the source of the misfit in the data, individual item fit statistics were examined. For four of the testlets (Sense of Control, Cognition Auditory and Emotional), the individual χ^2 estimates were significantly larger ($p>0.05$) than the other testlets (Figure 6.15). The fit residuals also show deviations from model expectation, with three testlets (Auditory, Sleep, Cognition) greatly exceeding the acceptable criterion.

Table 6.16. Summary fit statistics for second-order construct

			Item fit residual		Person fit residual		Item-trait interaction		No extreme
	No of subscales	Subscales removed	Mean	SD	Mean	SD	χ^2 (df)	*p**	
Dataset A (n = 261)	8		0.60	3.26	-0.21	1.17	107.14 (24)	**>0.001**	1
	7	AUD	0.58	2.3	-0.25	1.09	44.48 (21)	**0.014**	2
	6	AUD/SLP	0.42	2.08	-0.32	1.07	22.46 (18)	1.000	2
	5	AUD/SLP/ INTR	0.36	1.35	-0.34	1.02	12.13 (15)	1.000	2
	3	SOC/COG/ REL/QOL/ EMO	0.42	0.76	-0.25	0.91	22.67 (9)	**0.021**	1
	2	SOC/COG/ REL/QOL/ EMO/INTR	-0.15	0.97	-0.57	1.69	19.63 (18)	0.708	6
Dataset B (n = 279)	8		0.19	3.99	-0.22	1.26	195.57 (24)	**>0.001**	0
	7	AUD	0.29	2.20	-0.31	1.15	46.75 (21)	**0.007**	0
	6	AUD/SLP	0.35	1.58	-0.34	1.08	22.84 (18)	1.000	0
	5	AUD/SLP/ INTR	0.33	1.01	-0.34	0.98	17.19 (15)	1.000	1
	3	SOC/COG/ REL/QOL/ EMO	0.58	1.05	-0.32	1.06	29.87 (9)	**0.003**	0
	2	SOC/COG/ REL/QOL/ EMO/INTR	0.10	0.51	-0.63	1.82	33.05 (18)	**0.040**	7

* corrected for multiple comparisons. Bold = significant values. Class intervals = 4**.**

The Auditory and Cognition testlets exceeded both sets of criterion, violating the assumptions of the Rasch model. Of the two testlets, the Auditory testlet is by far more problematic than the Cognition testlet (Figure 6.15). The Auditory testlet discrepancies were noticeable for both the χ^2 estimate and fit residuals.

Table 6.17. Summary of reliability statistics for second-order construct in datasets A and B.

	Second order					
	N° of subscales	Subscale removed	**α**	**PSI**	**G_p**	**Strata***
Dataset A	Eight		0.90	0.91	3.28	4.71
	Seven	AUD	0.91	0.91	3.21	4.61
	Six	AUD/SLP	0.91	0.91	3.19	4.58
	Five	AUD/SLP/INTR	0.90	0.91	3.20	4.59
Dataset B	Eight		0.89	0.90	2.94	4.25
	Seven	AUD	0.88	0.88	2.74	3.40
	Six	AUD/SLP	0.91	0.91	3.26	4.68

*Strata values need to be truncated to whole numbers.

The large misfit of data to model expectation could therefore be an artefact of the Auditory testlet. Given the previous findings (Chapters 4 and 5) and the extremely unusual targeting problems (6.3.3.5), these results were not unexpected. The Auditory testlet was removed and a second-order model with seven testlets was then examined.

6.3.4.2 Evaluation of the seven-factor second-order structure

Overall fit dramatically improved following the removal of the Auditory testlet (Table 6.16). However, the item-trait interaction remained significant ($p>0.01$) and the SD for item fit residual indicated problems with the item data (SD < 2). Inspection of the individual item fit statistics clearly identified the source of the misfit as the high fit residuals for the Sleep testlet in both datasets and Intrusiveness testlet in dataset A (Figure 6.16). The χ^2 estimate for the Sleep testlet, albeit non-significant, was substantially larger than the other testlets (Figure 6.16).

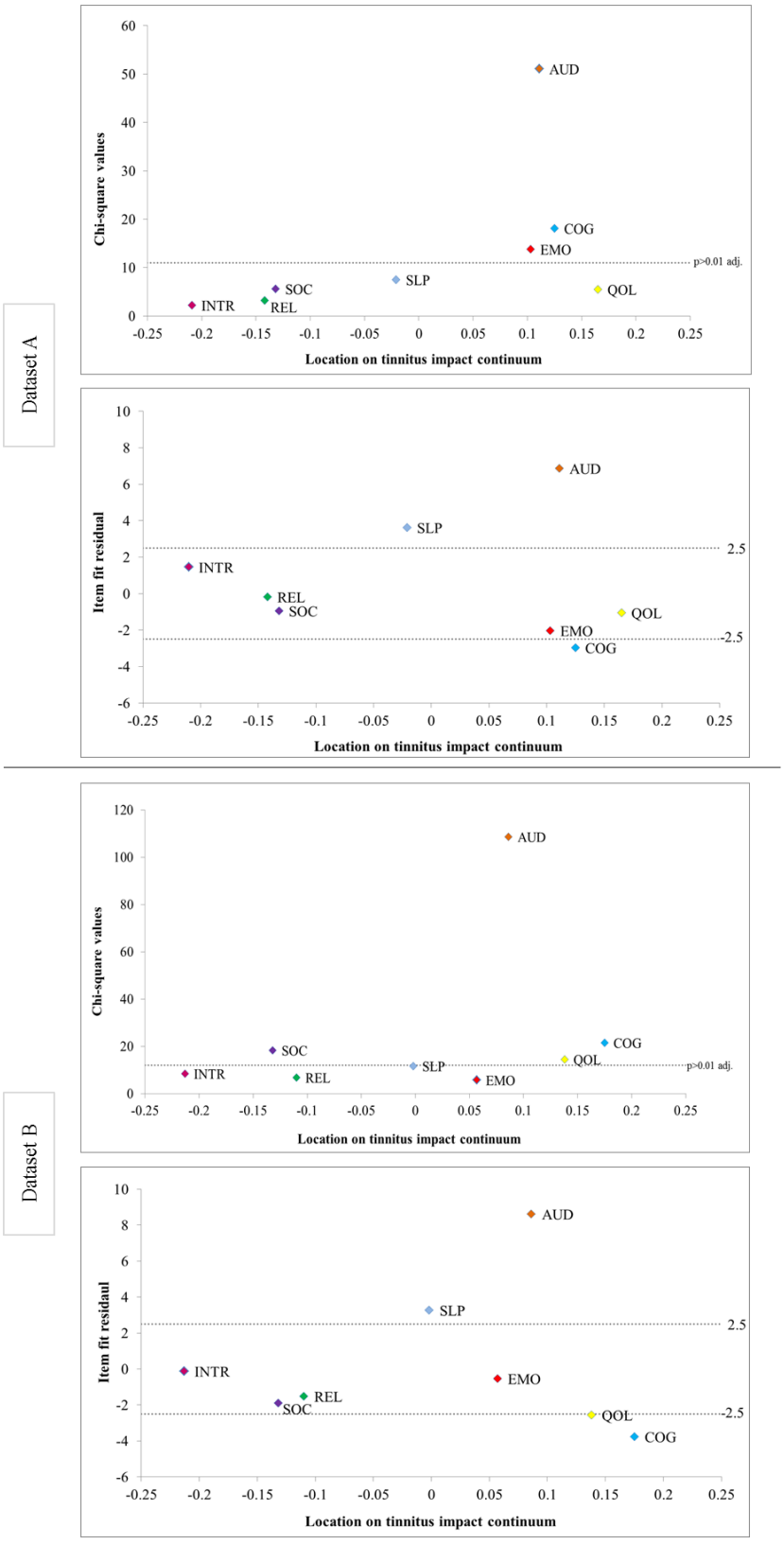

Figure 6.15. Individual item Chi-square values and fit residuals for eight-factor second-order structure in both datasets.

Chi-square (χ^2) values above the dashed lines indicate testlets that significantly deviate from expectation. Testlet fit residuals above and below the grey lines (± 2.5) are exceeding the criterion. INTR = Intrusiveness; SOC = Sense of control; COG = Cognition; SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional.

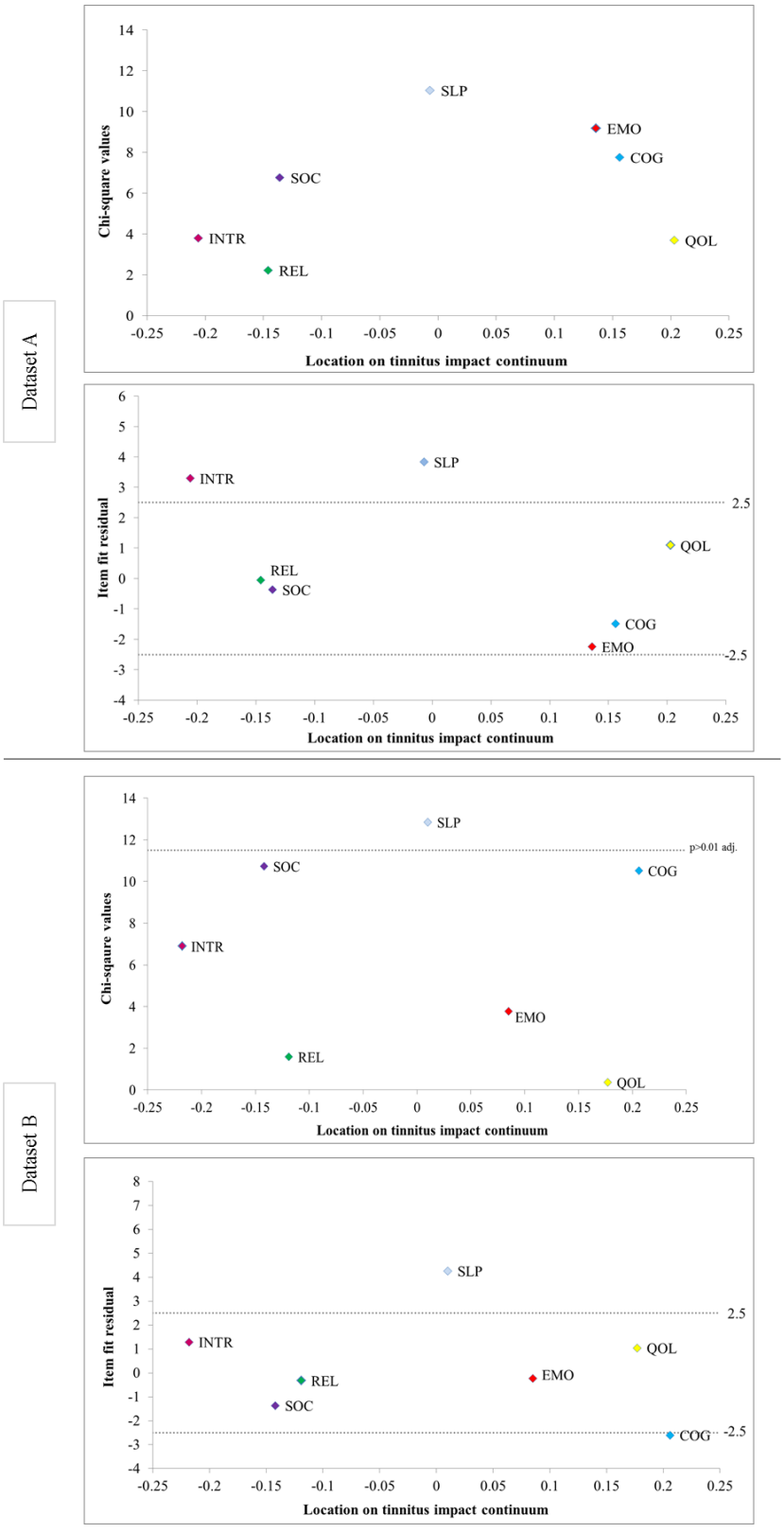

Figure 6.16. Individual item Chi-square values and fit residuals for seven-factor second-order structure in both datasets.

In dataset A, all Chi-square (χ^2) values for the testlets were non-significant. In dataset B, Chi-square (χ^2) values above the dashed line indicate testlets that significantly deviate from expectation. Testlet fit residuals above and below the grey lines (± 2.5) are exceeding the criterion. INTR = Intrusiveness; SOC = Sense of control; COG = Cognition; SLP = Sleep; REL = Relaxation; QOL = Quality of life; EMO = Emotional

To identify the utility of the Sleep subscale in the second-order structure, the item map was examined (Figure 6.17). Unfortunately due to limitations of RUMM software, the item map could only display a finite number of item thresholds. It was however apparent that the Sleep testlet thresholds were located past where the majority of thresholds were grouped and displayed. In other words, the Sleep testlet did not provide any additional measurement points that were not already provided by the other testlets. The Intrusiveness testlet thresholds (01-03) provided the first measurement points that would capture the initial first problem with tinnitus. Again given these findings and those reported in Chapter 4 and 5, the decision was made to remove the Sleep testlet.

The Sleep testlet was removed and a second-order model with six testlets was then examined.

6.3.4.3 Evaluation of the six-factor second-order structure

The overall summary fit statistics for the six-factor structure were now all within the established criterion, although the item fit residual SD still indicated some misfit at item level for dataset A (Table 6.16). Individual item fit statistics indicated that the Intrusiveness testlet fit residual was outside the established boundaries in dataset A whilst in dataset B, none of the testlets deviated from the model expectation (Figure 6.18). High positive residuals such as this can indicate that the testlet is measuring an alternative construct to the other testlets. However, before making the decision on whether to remove or maintain the Intrusiveness testlet, further evaluation of the model fit, the targeting and assumptions of local independence was indicated.

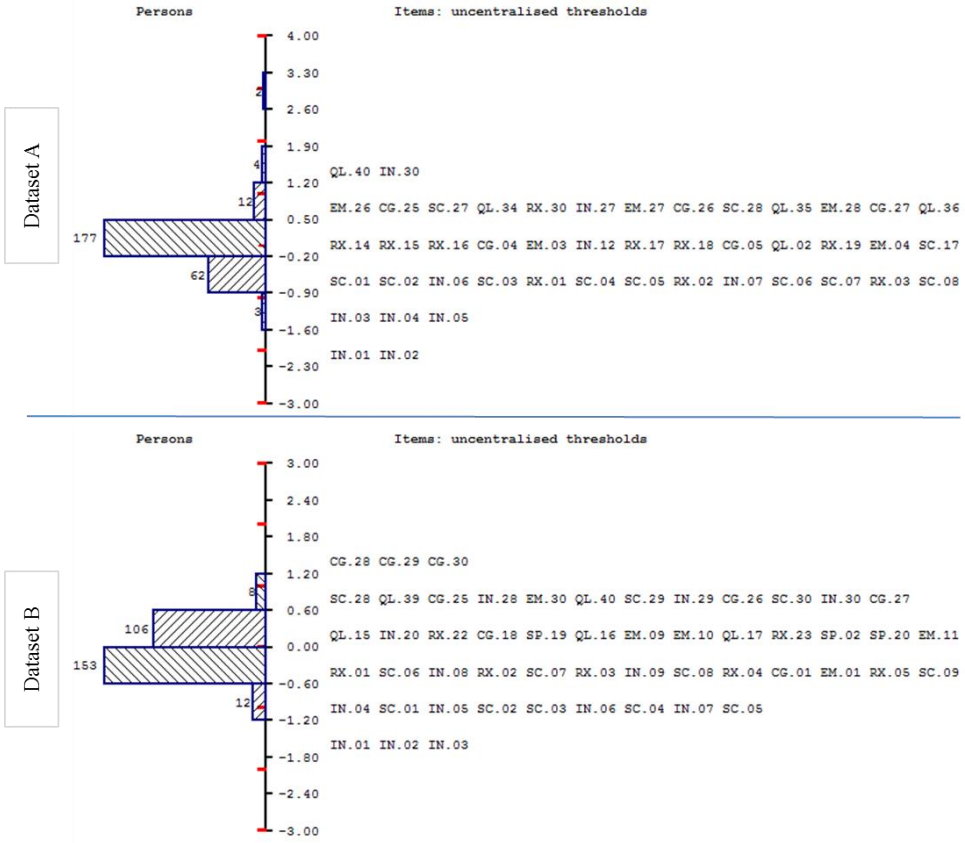

Figure 6.17. Item maps for the seven-factor second-order structure.
Intrusiveness lower thresholds (IN.01) provide the first indication of a problem with tinnitus, whilst the Cognition higher thresholds (CG. 28 – 30) identify severe problems with tinnitus. Left = Person locations. Right = Item thresholds. IN = Intrusiveness; SC = Sense of control; CG = Cognition; SP = Sleep; AD = Auditory; RX = Relaxation; QL = Quality of life; EM = Emotional. The number presented after the letters signifies the threshold parameter.

The χ^2 estimate for the Intrusiveness testlet was similar to the estimates of the other testlets, all of which were non-significant, indicating invariance in the hierarchical ordering that conform to the Rasch model (Figure 6.18). The Intrusiveness testlet provided the first indication of a problem with tinnitus and this did not change with the new six-factor structure (Figure 6.19). The first few thresholds capture persons with milder levels of tinnitus. Although one might expect this to be case given that the Intrusiveness items are the first presented in the TFI, in fact the first response categories (thresholds) for any of the items could have provided the first indication of tinnitus problems. The testlet locations do not have to logically follow the ordering of the items in the questionnaire format.

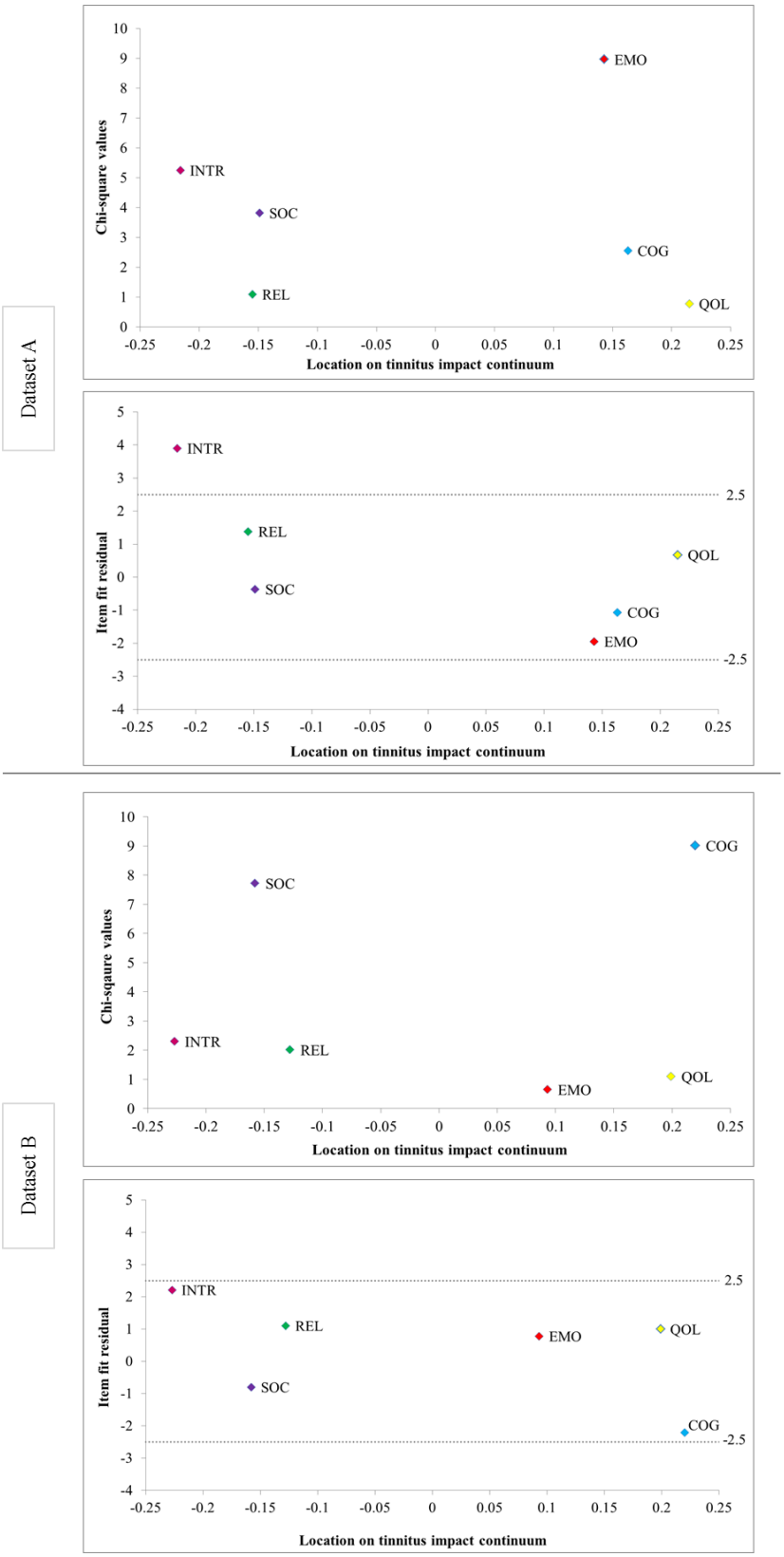

Figure 6.18. Individual item Chi-square values and fit residuals for the six-factor second-order structure in both datasets.
Chi-square (χ^2) values for the testlets were non-significant, indicating acceptable fit to the Rasch model expectations. Testlet fit residuals above and below the grey lines (± 2.5) are exceeding the criterion. All testlets in dataset B are within criteria. INTR = Intrusiveness; SOC = Sense of control; COG = Cognition; REL = Relaxation; QOL = Quality of life; EMO = Emotional

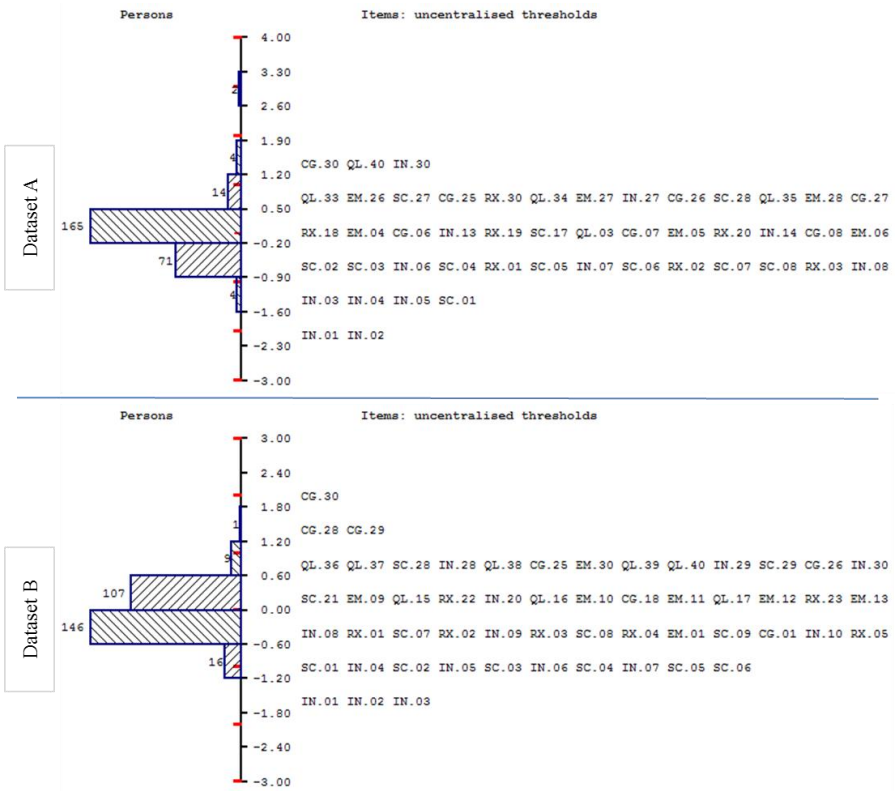

Figure 6.19. Item maps for the six-factor second-order structure.
Intrusiveness lower thresholds (IN.01) were still the first indication of a problem with tinnitus and the Cognition higher thresholds (CG. 28 – 30) identify severe problems with tinnitus. Left = Person locations. Right = Item thresholds. IN = Intrusiveness; SC = Sense of control; CG = Cognition; RX = Relaxation; QL = Quality of life; EM = Emotional. The number presented after the letters signifies the threshold parameter.

For instance, the first Relaxation threshold (displayed as RX.01 in Figure 6.19) provided measurement information before the Cognition testlet, even though this subscale is ordered later in the TFI. The Intrusiveness testlet provided measurement points at -2 logits that were not covered by any other testlets. Removing it could potentially impact on the measurement precision.

Furthermore, data in the six-factor structure conformed to the Rasch model expectation in all other conditions. The six-factor structure was well-targeted with all persons covered by items with the exception of the extreme scores at the higher level (Figure 6.20). The information curve value was exceptionally high indicating less

error and more precision of measurement associated with the locations within the curve.

The PSI score was better than the original eight testlet version (Table 6.17). The *r* value (0.71) indicated high association between the latent domains. The A value (0.90) indicated 90% common shared variance. The assumptions of local independence were not violated. Analysis of the residual correlations revealed no observable pattern between testlets, suggesting that the content of the testlets was invariant (Table 6.18). PCA identified two subsets in the testlet residuals. The Sense of Control and Intrusiveness testlets had strong negative loading values, whilst the QoL, Cognition and Emotional testlets had strong positive loading residuals. Independent t-tests indicated that a small proportion of comparisons fell outside the 95% CI (≥5%). Person locations from each subset of testlets were significantly different for 4.7% of cases in dataset A and for 5.3% of cases in dataset B. The six-factor structure conformed to the assumptions of unidimensionality.

It is worthwhile examining the 5-factor structure to establish the potential improvement of model fit, unidimensionality and the likely impact on the measurement precision and targeting of removing the Intrusiveness testlet. This also provided the unique opportunity to assess potential alternative structures with the removed testlets. The next section will therefore assess the impact of removing the Intrusiveness testlet from the 6-factor structure and the possible bi-factor models; the 5-factor structure (Sense of control, Cognition, Relaxation, QoL, Emotional) with a 3-factor structure (Intrusiveness, Sleep, Auditory) (Fig. 6.14).

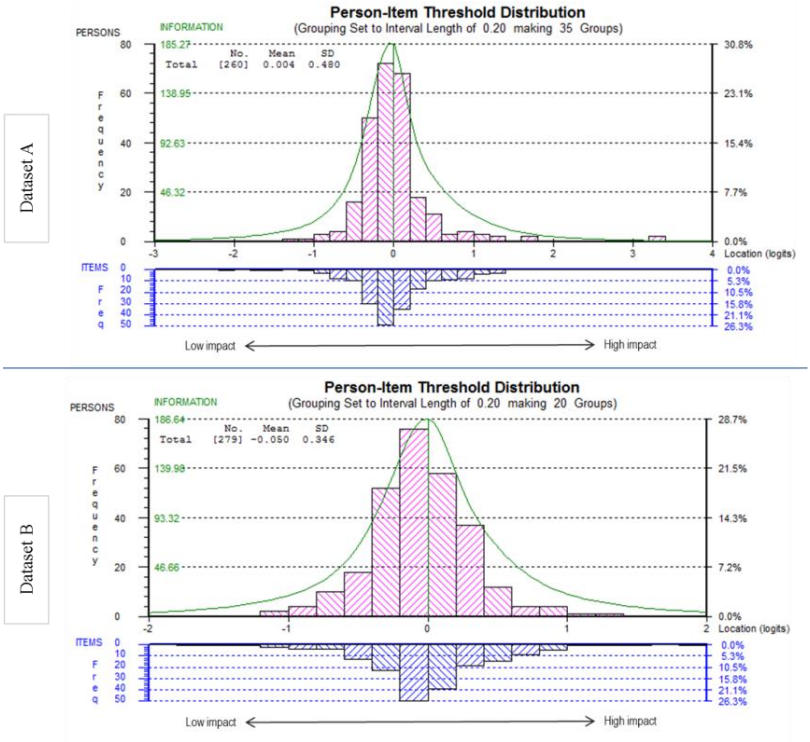

Figure 6.20. Person-item distribution for 6-factor second-order structure
Person and testlet threshold locations were aligned, indicating that the 6-factor structure was well-targeted for the population. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus.

6.3.4.4 Evaluation of the five-factor and three-factor second-order structures

The 5-factor model slightly improved the model fit in both datasets (Table 6.16). The residual correlations showed no discernible pattern (Table 6.19), and although the proportion of significant independent t-tests has increased slightly (5.4%), it is just within the recommended criteria. The 5-factor structure is assessing a unidimensional underlying construct. However, the information previously provided by the Intrusiveness testlet was lost within this structure and measurement precision was reduced. Closer inspection of the item map show that the Sense of control testlet now provides the first measurement points (Figure. 6.21). However, the first person locations are not covered by any testlet thresholds and measurement at these points does not exist.

Table 6.18. Residual correlations between the 6 testlets

	TESTLET	INTR	SOC	COG	REL	QoL	EMO
Dataset A	INTR	1					
	SOC	-0.028	1				
	COG	-0.307	-0.196	1			
	REL	-0.199	-0.197	-0.175	1		
	QoL	-0.28	-0.404	0.022	-0.329	1	
	EMO	-0.265	0.101	-0.171	-0.117	-0.205	1
	TESTLET	INTR	SOC	COG	REL	QoL	EMO
Dataset B	INTR	1					
	SOC	0.087	1				
	COG	-0.115	-0.124	1			
	REL	-0.233	-0.053	-0.194	1		
	QoL	-0.333	-0.433	-0.067	-0.291	1	
	EMO	-0.273	-0.16	-0.22	-0.225	-0.103	1

The six-factor structure would therefore appear to provide the most information on the impact of tinnitus increasing measurement precision. However, before making a final recommendation of the inclusion of Intrusiveness testlet and new TFI structure, the three-factor structure first needs to be examined. If this data conforms to the Rasch model, then the information from the Intrusiveness testlet will be available within this three-factor structure and therefore the five-factor structure would be considered an acceptable fit of the data with sufficient measurement precision. This however was not the case. The overall summary fit statistics for the three-factor structure indicated deviations from the Rasch model at item level (Table 6.16). The χ^2 estimates for item-trait interactions and individual item fit was significant ($p>0.05$) indicating unexpected deviations between observed and expected scores.

Table 6.19. Residual correlations between the 5 testlets

	TESTLET	SOC	COG	REL	QoL	EMO
Dataset A	SOC	1				
	COG	-0.226	1			
	REL	-0.225	-0.245	1		
	QoL	-0.414	-0.09	-0.397	1	
	EMO	0.084	-0.266	-0.169	-0.299	1
	TESTLET	SOC	COG	REL	QoL	EMO
Dataset B	SOC	1				
	COG	-0.118	1			
	REL	-0.061	-0.245	1		
	QoL	-0.456	-0.089	-0.392	1	
	EMO	-0.168	-0.256	-0.303	-0.198	1

The Intrusiveness testlet is measuring a construct that is clearly unrelated to the other two testlets (Auditory, Sleep), and is more suitable with the underlying construct being measured by the remaining five testlets. The 6-factor structure is therefore identified as providing the most reliable information for an overall score.

6.3.4.5 Evaluation of the residual correlations between the two discarded factors

Conceptually, the Auditory and Sleep subscales (testlets) do appear to assess a purely functional component, i.e. the functional impact of tinnitus on hearing or sleep, which could theoretically create an alternative second-order construct. However, previous evidence (Chapters 4 & 5) showed that these two subscales were poorly related. Psychometrics did not support this as a valid construct. High fit residual correlations (> 0.60) between the items within the subscales confirmed the presence of two latent dimensions and indicated possible response dependency in the items.

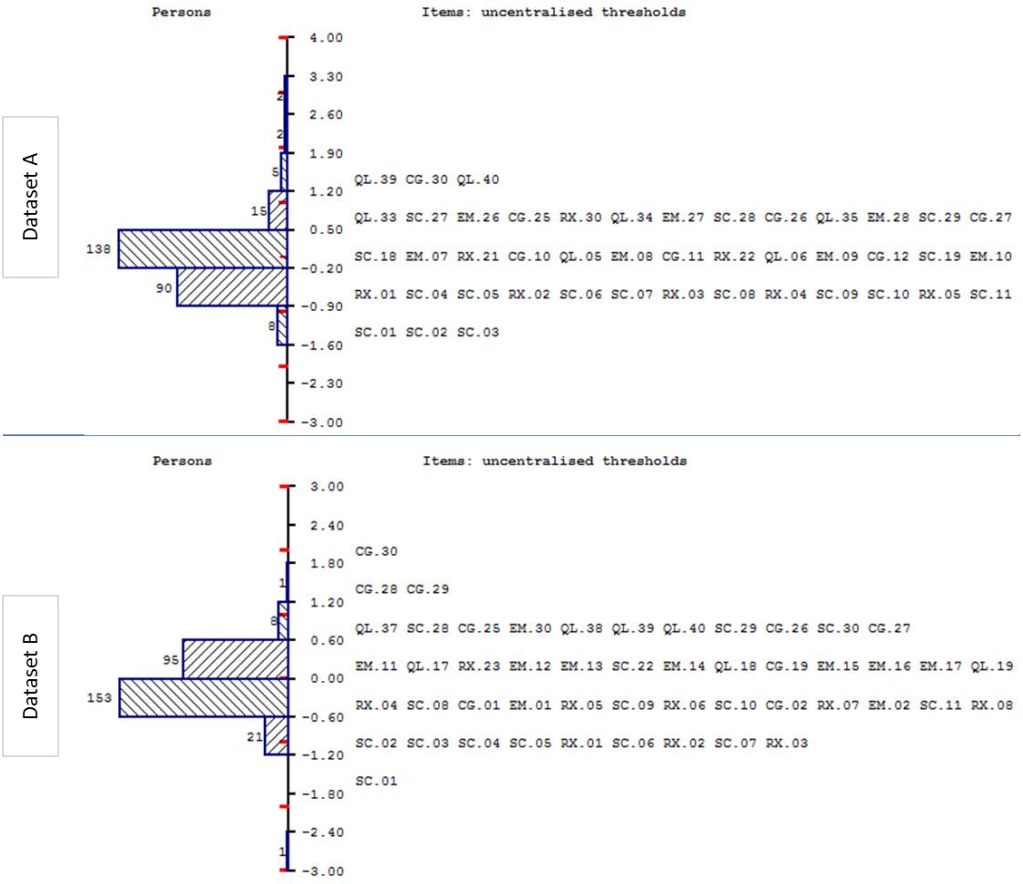

Figure 6.21. Item maps for 5-factor second-order structure.
Without the Intrusiveness subscale, the Sense of control lower thresholds (SC.01) provides the first indication of a problem with tinnitus, but this would not account for all persons at the lower end of the scale. Left = Person locations. Right = Item thresholds. IN = Intrusiveness; SC = Sense of control; CG = Cognition; RX = Relaxation; QL = Quality of life; EM = Emotional. The number presented after the letters signifies the threshold parameter. Logit scale values represent the severity continuum.

6.3.5. The full dataset analysis

To ensure power and stability of the person-item parameters and calibrations for the final recommendations and transformation analysis, the full dataset was subjected to Rasch modelling.

Initially, the model fit for the subscales and 6-factor second-order structure were confirmed and finalised. Problematic items or subscales were recalibrated when possible or recommended to be removed before DIF analysis (6.3.5.2) and score transformations analysis was conducted (6.3.5.3).

6.3.5.1 Confirmation of the subscales and six-factor structure fit to Rasch model expectation

For confirmation purposes, the summary fit statistics (Table 6.20), individual fit statistics (Table 6.21 – 6.23), category ordering (Figure 6.22), targeting (Figures 6.23 – 6.24) and reliability (Table 6.24) for each subscale and the six-factor structure (TFI-19/TFI-18) were examined. The number of class intervals chosen with this larger dataset was 7, so that more than 50 persons were in each class.

Consistent with the previous results, all category thresholds were ordered for all items within their designated subscales (Figure 6.22) and only 2% to 7% participants had negative residuals outside the acceptable range (Table 6.23). The targeting and reliability for all the subscales and 6-factor mirrored previous results, i.e. the same subscales were reasonably well-targeted (Figures 6.23; Table 6.24). Four of the eight subscales (Intrusiveness, Cognition, Relaxation and Emotional) and the six-factor structure showed acceptable data fit to the Rasch model. Although as predicted in all cases the χ^2 estimates were somewhat inflated (see Tables 6.20 – 6.22).

Unexpectedly, the Sense of Control subscale now no longer conformed to the Rasch model. Closer inspection of the item fit statistics revealed that the fit residuals for SOC4 and SOC5 were now above ±2.5, and SOC4 had a large positive residual indicating that it was no longer measuring the same construct as the other items (Table 6.21). These large deviations appeared to be a product of collapsing SOC4 thresholds. The large fit residuals were not present beforehand and it was noted that after SOC4 thresholds were collapsed, the item residuals for SOC4 and SOC5 did increase.

Table 6.20. Summary fit statistics for eight subscales and TFI six-factor (TFI-18/TFI-19) structure using full dataset.

		Item fit residual		Person fit residual		Item-trait interaction		No extreme
Subscale	**Items**	Mean	SD	Mean	SD	χ^2 (df)	*p**	
Intrusiveness	INTR1, INTR2, INTR3	-0.26	1.17	-0.53	1.04	22.17 (18)	0.673	11
Sense of control-ordered	SOC4,SOC5, SOC5	-0.39	**2.91**	-0.51	1.02	29.51 (18)	0.126	19
Sense of control-disordered	SOC4,SOC5, SOC5	-0.05	1.13	-0.51	1.06	34.76 (18)	**0.030**	19
Cognition	COG7, COG8, COG9	0.11	1.44	-0.68	1.1	21.49 (18)	0.765	51
Sleep	SLP10, SLP11, SLP12	-0.29	**2.52**	-0.73	1.17	27.70 (18)	0.198	90
Auditory	AUD13, AUD14, AUD15	-0.17	**3.14**	-0.75	1.14	13.98 (18)	1.000	90
Relaxation	REL16, REL17, REL18	-0.54	1.46	-0.52	0.91	21.59 (18)	0.750	53
QoL-4	QOL19, QOL20, QOL21, QOL22	-0.14	2.04	-0.62	1.27	47.74 (24)	**0.009**	65
QoL-3	QOL19, QOL20, QOL21	-0.21	0.26	-0.63	1.08	29.67 (18)	0.122	89
Emotional	EMO23, EMO24, EMO25	-0.53	1.42	-0.52	0.96	31.80 (8)	**0.070**	58
Second order	**Testlets**							
six-factor TFI-19	INTR, SOC, COG, REL, QOL-4, EMO	0.45	2.34	-0.35	1.08	53.77 (36)	0.168	2
six-factor TFI-18	INTR, SOC, COG, REL, QOL-3, EMO	0.49	2.09	-0.35	1.08	46.49 (36)	0.678	2

* corrected for multiple comparisons. Bold = exceeds recommended criterion. Underlined = marginally below criteria. χ^2 = Chi-square. No extreme = number of extreme. n = 540. Class intervals = 7.

Table 6.21. Individual item fit statistics for the full dataset.

	Item	Location	SE	Fit residual	χ^2	df	*p**
Intrusiveness	INTR1	-0.215	0.026	-0.795	*10.574*	6	0.307
	INTR2	-0.463	0.031	1.091	3.909	6	1.000
	INTR3	0.678	0.025	-1.061	7.690	6	0.785
Sense of Control	SOC4	-0.183	0.022	0.749	7.346	6	0.870
	SOC5	0.429	0.027	-1.343	**19.971**	6	**0.008**
	SOC6	-0.246	0.025	0.436	7.444	6	0.845
Sense of Control -ordered	SOC4	-0.361	0.041	**2.859**	6.619	6	1.000
	SOC5	0.56	0.029	**-2.736**	17.002	6	**0.028**
	SOC6	-0.199	0.027	-1.313	5.898	6	1.000
Cognition	COG7	-0.392	0.041	0.041	8.507	6	0.610
	COG8	0.205	0.041	-1.298	10.815	6	0.283
	COG9	0.187	0.042	1.583	2.169	6	1.000
Sleep	SLP10	-0.189	0.034	1.489	2.686	6	1.000
	SLP11	0.053	0.034	**-3.186**	**20.602**	6	**0.006**
	SLP12	0.136	0.034	0.826	4.421	6	1.000
Auditory	AUD13	-0.134	0.043	1.755	0.633	6	1.000
	AUD14	0.341	0.042	**-3.795**	11.503	6	0.222
	AUD15	-0.207	0.04	1.535	1.848	6	1.000
Relaxation	REL16	0.198	0.035	-1.871	12.334	6	0.165
	REL17	0.443	0.035	-0.782	4.727	6	1.000
	REL18	-0.641	0.035	1.032	4.535	6	1.000
QoL-4	QOL19	0.011	0.027	0.006	7.440	6	0.846
	QOL20	-0.248	0.027	-1.370	13.574	6	0.114
	QOL21	0.165	0.027	-1.881	**18.529**	6	**0.015**
	QOL22	0.072	0.027	**2.671**	8.450	6	0.621
QoL-3	QOL19	0.04	0.03	0.058	8.519	6	0.607
	QOL20	-0.261	0.029	-0.454	7.899	6	0.737
	QOL21	0.221	0.03	-0.240	13.258	6	0.117
Emotional	EMO23	-0.035	0.031	-0.736	14.028	6	0.088
	EMO24	-0.428	0.032	-1.851	11.885	6	0.194
	EMO25	0.463	0.032	0.980	5.890	6	1.000

* corrected for multiple comparisons. Bold = exceeds recommended criterion. χ^2 = Chi-square. n = 540. Class intervals = 7.

Table 6.22. Individual item fit statistics for the six testlets in the six-factor structures (TFI-18/TFI-19) before and after Emotional testlet recalibration for population Differential Item Functioning (DIF).

DIF		Testlets	Location	SE	Fit residual	χ^2	*df*	*p**
	six-factor TFI -19	Intrusiveness	-0.218	0.011	**4.180**	9.142	6	0.996
		Sense of Control	-0.15	0.01	-0.939	13.136	6	0.246
		Cognition	0.185	0.009	-2.400	18.58	6	**0.030**
		Relaxation	-0.138	0.009	1.694	2.667	6	1.000
		QoL - 4	0.209	0.008	0.976	2.692	6	1.000
		Emotional	0.113	0.009	-0.824	7.548	6	1.000
	six-factor TFI -18	Intrusiveness	-0.22	0.011	**3.937**	10.862	6	0.558
		Sense of Control	-0.151	0.011	-1.206	12.599	6	0.300
		Cognition	0.188	0.009	-1.597	8.167	6	1.000
		Relaxation	-0.137	0.009	1.539	1.365	6	1.000
		QoL - 3	0.203	0.009	1.028	4.207	6	1.000
		Emotional	0.117	0.009	-0.741	9.298	6	0.948
Population DIF	six-factor TFI -19	Intrusiveness	-0.24	0.011	**4.274**	9.783	6	0.804
		Sense of Control	-0.171	0.011	-0.858	12.942	6	0.264
		Cognition	0.165	0.009	-2.43	17.193	6	**0.054**
		Relaxation	-0.161	0.009	1.628	3.102	6	1.000
		QoL - 4	0.188	0.008	1.006	1.556	6	1.000
		Emotional (Clinic)	0.03	0.013	1.022	1.347	6	1.000
		Emotional (Res)	0.189	0.013	-2.484	10.887	6	0.552
	six-factor TFI -18	Intrusiveness	-0.242	0.011	**4.020**	10.451	6	0.642
		Sense of Control	-0.173	0.011	-1.135	11.494	6	0.444
		Cognition	0.168	0.009	-1.642	12.547	6	0.306
		Relaxation	-0.16	0.009	1.453	1.552	6	1.000
		QoL -3	0.182	0.009	1.084	2.875	6	1.000
		Emotional (Clinic)	0.034	0.013	1.021	1.662	6	1.000
		Emotional (Res)	0.191	0.013	-2.314	11.296	6	0.480

* corrected for multiple comparisons. Bold = exceeds recommended criterion. χ^2 = Chi-square. n = 540. Class intervals = 7.

Table 6.23. Individual person fit statistics for the subscales and TFI six-factor (TFI-18/TFI-19) structure using the full database.

	N° of extremes residuals (%)
Subscale	**- 2.5**
Intrusiveness	18 (3)
Sense of control	27 (5)
Cognition	12 (2)
Sleep	23 (4)
Auditory	22 (4)
Relaxation	15 (3)
QoL-4	40 (7)
QoL-3	17 (3)
Emotional	15 (3)
Second-order	
six-factor TFI-19	21 (4)
six-factor TFI-18	18 (3)

Given that the measurement precision was also affected by the thresholds collapsing (6.3.3.4), the decision was made to return to the original response structure (0 to 10) with the disordered thresholds. This in turn substantially improved model fit and measurement precision, although the χ^2 estimate was again significant (Table 6.20); SOC4 no longer showed any misfit with the other items (Table 6.21).

The three other subscales (Sleep, Auditory, QoL) remained problematic. The QoL subscale had exceptionally large χ^2 estimates and a high item fit residual. The χ^2 estimate for overall fit was significant ($p = 0.012$ corrected). The estimates for all but one item were large, although only QOL21 is significant ($p = 0.012$ corrected), the others are marginal (Tables 6.20 and 6.21). This could be a product of the large sample size, yet there was some indication previously of potential deviations.

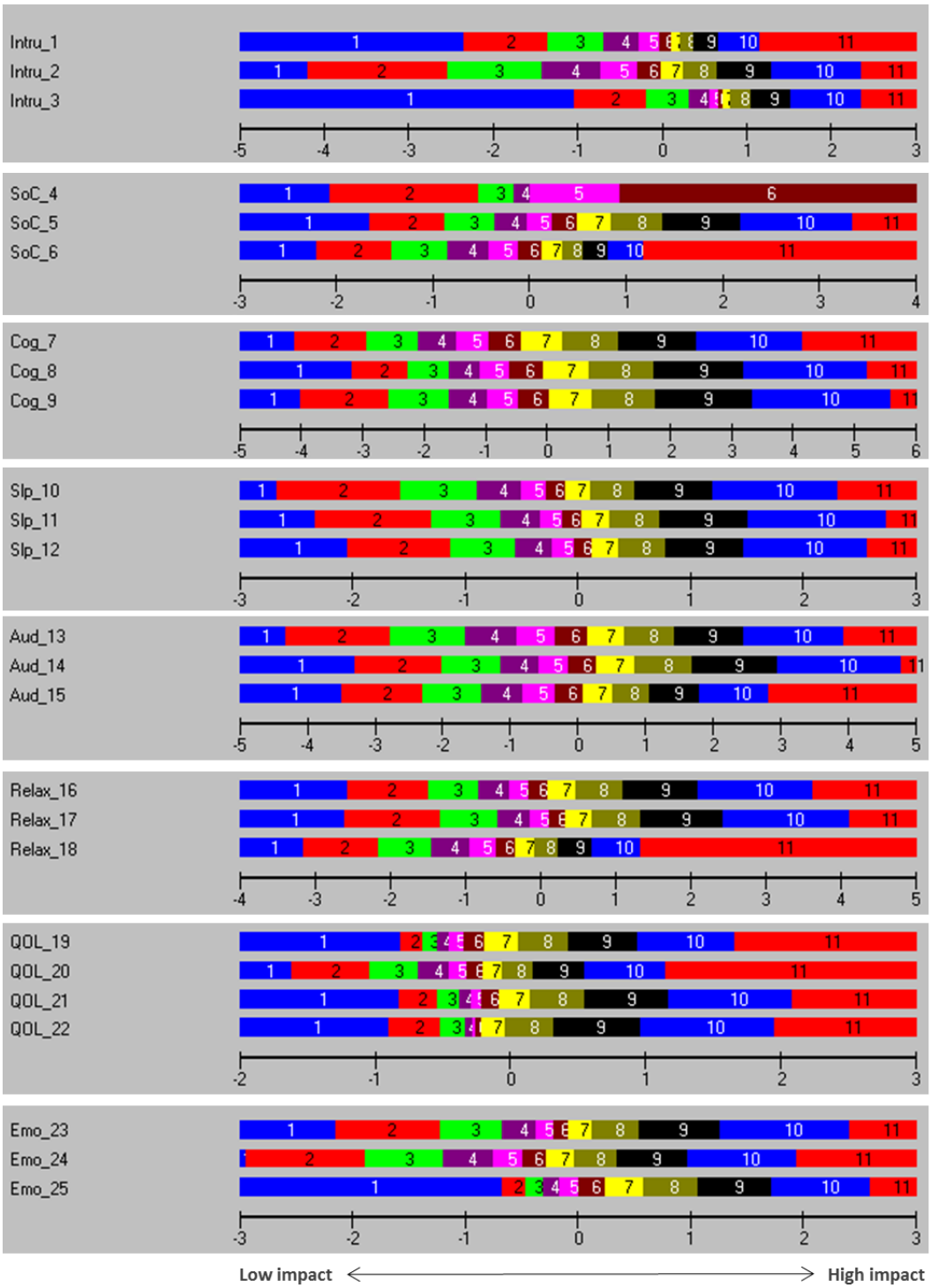

Figure 6.22. Thresholds distribution for all items within designated subscales using full dataset.
Logit scale continuum presented below each subscale, with lower impact indicated by negative logit values and higher by positive logit values.

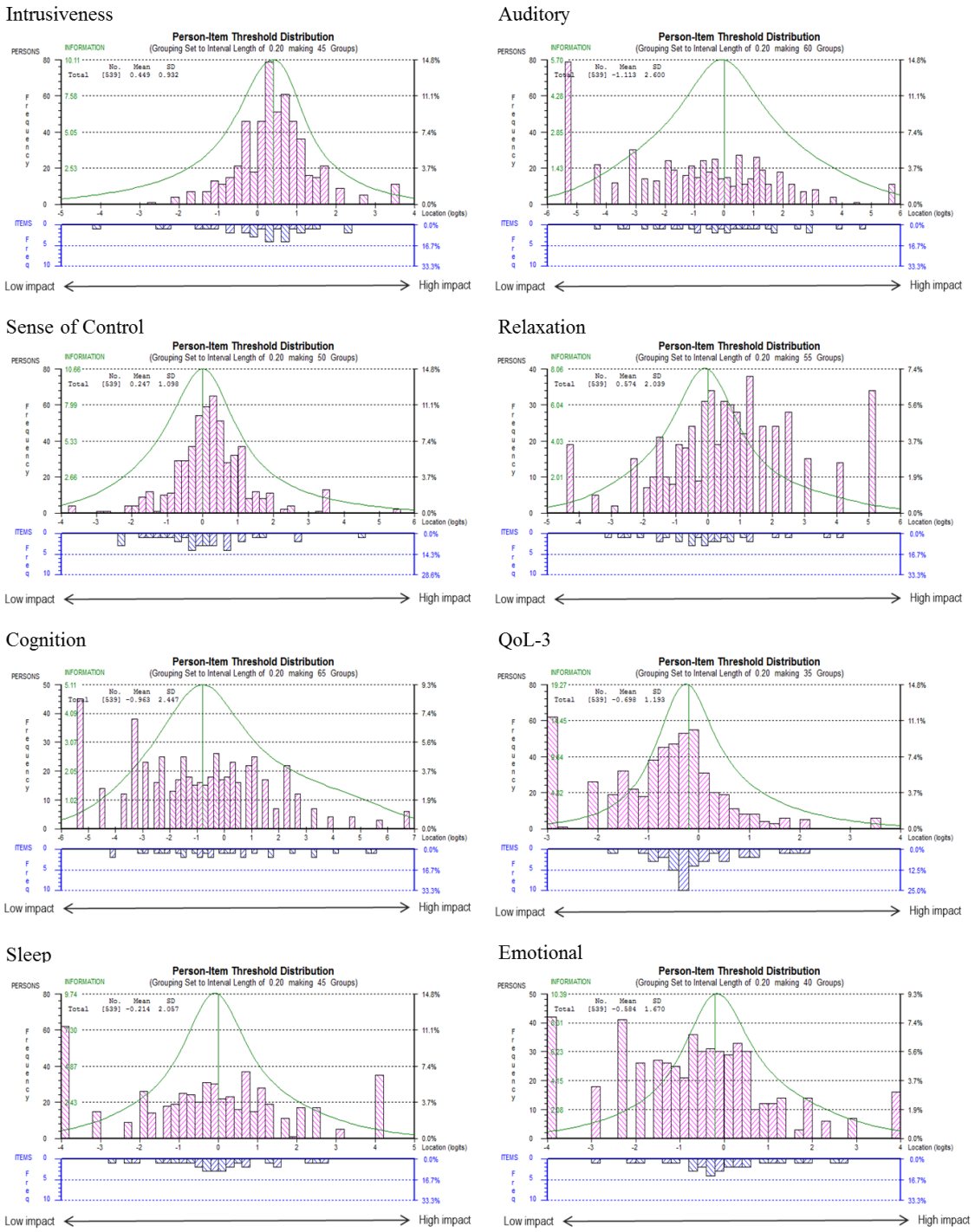

Figure 6.23. Person-item threshold distributions for the eight subscales using full dataset.

Person-item threshold distributions in the full dataset reflect previous distributions. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus. n = 540.

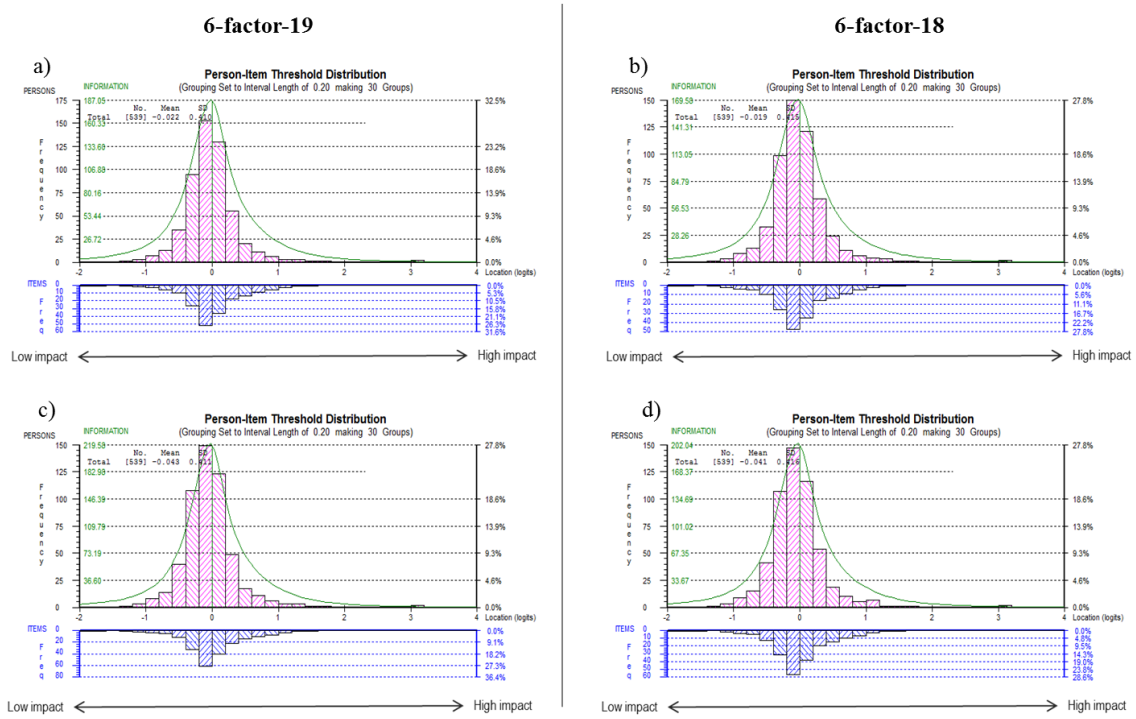

Figure 6.24. Person-item threshold distributions for TFI 6-factor (TFI-19/TFI-18) before (a/b) and after (c/d) Emotional testlet recalibration for population DIF using the full dataset..

The 6-factor structure targets the population. Blue = item locations. Pink = person locations. Green = information curve and represents the inverse standard error. The logit scale represents the severity continuum with negative values indicating low impact and positive values indicating higher impact of tinnitus. n = 540. DIF = Differential Item Functioning

Interestingly, the source of this misfit in fact could potentially be the item that has a very small non-significant χ^2 estimate; QOL22 (Table 6.21). QOL22 has a large positive fit residual which indicates that this item is not tapping into the same underlying construct as the other items subscale. This misfitting item could be shaping the model expectation. This is not the first time the reliability of QOL22 has been called into question. On top of the evidence from this chapter (noticeable under-discrimination, large χ^2 estimates, disordered categories), in Chapters 4 and 5 it was shown that this item cross-loaded with other factors and was regularly missed by participants within the clinical study. Consequently, QOL22 has proven to be problematic and should potentially be removed from the questionnaire.

Table 6.24. Summary of reliability statistics, Cronbach's alpha (α), Person separation index (PSI), G*p* and Strata for each subscale and 6-factor structure in full dataset.

Subscale	α	PSI	G_p	Strata*
Intrusiveness	0.78	0.77	1.90	2.87
Sense of control	0.75	0.76	1.84	2.79
Cognition	0.93	0.91	3.39	4.85
Sleep	0.91	0.86	2.72	3.96
Auditory	0.93	0.91	3.21	4.61
Relaxation	0.91	0.87	2.85	4.14
QoL-3	0.87	0.80	1.98	2.97
QoL-4	0.89	0.80	2.00	3.00
Emotional	0.90	0.86	2.44	2.59
Second-order				
six-factor TFI-18	0.91	0.91	3.12	4.49
six-factor TFI-19	0.91	0.91	3.17	4.56

*Strata values should be truncated to whole numbers. n = 540.

For completeness and to study the consequences of removing the item, the remaining analyses for both the QoL subscale and the six-factor structure were conducted with QOL22 (QoL-4/six-Factor TFI-19) and without QOL22 (QoL-3/six-Factor TFI-18). The model fit following the removal of QOL22 dramatically improved for QoL subscale (Tables 6.20 and 6.21).

The issue with the Sleep subscale was less clear-cut, with a high negative fit residual for SLP11 indicating possible overlap in content and item redundancy (Table 6.21). An inspection of item wording confirmed an overlap in content rather than item response dependency indicated in the correlations. SLP 11 asked "How often did your tinnitus cause you difficulty in getting as much sleep as you needed?", whilst SLP12 asked "How much of the time did your tinnitus keep you from sleeping as deeply or as peacefully as you would have liked?". Both SLP11 and SLP12 are essentially measuring the same aspect of the sleep difficulties. Creating a two-item

subscale by either combining items or removing an item would reduce the overall reliability of Sleep as a standalone scale. Furthermore, the high fit residual was not consistent across all the analyses; there is little to no evidence of the deviation on the item characteristic curve graph for SLP11, and the measurement precision was not adversely affected by this overlap. Although the person distribution was flatter than previously observed, the information curve value was still high and the items and persons were still aligned (Figure 6.23). Therefore, the item remained in the subscale and the subscale was included in the following analyses.

The Auditory subscale has consistently shown poor fit to the Rasch model in all analyses. It had a large negative fit residual for AUD14 (Table 6.21) which, alongside the high residual correlations, indicates overlap in content and redundant items within the scale. A closer examination of the item wording indicates the high residual and high correlations could be a product of response dependency within the subscale. AUD13 asked about "ability to hear clearly?", AUD14 asked about "ability to understand people who are talking?", and AUD15 asked about "ability to follow conversations in a group or at meeting?". The second two items responses appeared to be dependent on the response to the first item. Again removing AUD14 was not a viable option as this would reduce the overall reliability. More importantly, the extremely flat person distribution clearly apparent throughout, indicated a misfit between the health problem being measured by the items and the health concerns of the persons. Further evidence of this misfit in the underlying measurement was apparent within the second-order structure in which the Auditory subscale consistently showed the most extreme large deviations from the second-order underlying construct measured by the other subscales combined. In light of this, the

Auditory subscale was not considered reliable and therefore was excluded from the transformation analysis.

6.3.5.2 Differential Item Functioning in the subscales and six-factor structure

All eight subscales were submitted to DIF analysis. Although the Auditory subscale was recommended for removal, the decision was made to assess this subscale in relation to the person factors as the interactions might provide additional information about the construct underlying the auditory subscale. Within the DIF analysis, the data in each class interval would be separated and compared based on person factors. With this in mind, unless otherwise stated, the class intervals were 5 for this analysis.

6.3.5.2.1 Population

Initially, the distribution of the person locations in the subscales and six-factor structure were examined for population differences using analysis of variance (Table 6.25). In all cases, the mean person locations were significantly lower on the continuum for the research population than the clinical population (Table 6.26). Given the score distributions seen in Chapters 4 and 5, these results are unsurprising. Basically, the items/testlets located higher were more likely to target the majority of the clinical population whilst those located lower on the continuum were more likely to target the research population (Figures 6.25 – 6.26). Therefore the items/testlets needed to be distributed along the continuum to reliably differentiate between both populations. In most instances, this was the case, although some measurement points at the extreme higher end of the continuum are somewhat limited, therefore it may be harder to differentiate the clinical population at these extremes.

Table 6.25. ANOVA results for differential functioning in targeting for population, gender and age groups in the TFI subscales and the six-factor structure (TFI-19/TFI-18).

	Population		Gender		Age	
Items	F	*p**	F	*p**	F	*p**
Intrusiveness*	31.66	**>0.001**	14.22	**>0.001**	14.876	**>0.001**
Sense of control*	35.42	**>0.001**	16.83	**>0.001**	1.933	1.000
Cognition*	29.436	**>0.001**	9.460	**0.018**	0.414	1.000
Sleep*	31.047	**>0.001**	11.539	**>0.001**	0.878	1.000
Auditory*	9.386	**0.018**	1.347	1.000	7.224	0.567
Relaxation*	19.01	**>0.001**	12.11	**>0.001**	0.022	1.000
Quality of life - 4	21.37	**>0.001**	0.035	1.000	1.174	1.000
Quality of life – 3*	24.06	**>0.001**	0.03	1.000	3.92	0.432
Emotional*	71.33	**>0.001**	9.592	**0.018**	0.223	1.000
Second-order						
TFI - 19‡	41.83	**>0.001**	7.755	0.09	2.373	1.000
TFI - 18‡	43.21	**>0.001**	8.49	**0.054**	2.78	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5.

Table 6.26. Mean locations and standard deviations for population, gender and age person factor groups targeting in subscales and the six-factor structure (TFI-19/TFI-18).

	Population				Gender				Age			
	Clinical		Research		Male		Female		>50yrs		50 to 70+yrs	
Items	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Intrusiveness	0.681	1.00	0.241	0.81	0.342	0.94	0.658	0.89	0.246	0.87	0.564	0.95
Sense of control	0.544	1.03	0.070	0.81	0.175	0.95	0.525	0.91	0.218	0.84	0.336	1.01
Cognition	-0.372	2.40	-1.489	2.37	-1.194	2.33	-0.514	2.62	-0.872	2.09	-1.014	2.63
Sleep	0.295	2.02	-0.668	1.99	-0.430	2.02	0.201	2.08	-0.324	1.92	-0.151	2.13
Auditory	-0.752	2.70	-1.434	2.47	-1.019	2.45	-1.294	2.88	-1.510	2.68	-0.887	2.53
Relaxation	0.973	2.08	0.219	1.94	0.354	2.01	0.994	2.03	0.592	1.79	0.564	2.17
Quality of life – 4	-0.451	1.27	-0.918	1.08	-0.693	1.15	-0.713	1.28	-0.772	1.11	-0.656	1.24
Quality of life - 3	-0.449	1.43	-1.013	1.24	-0.741	1.32	-0.763	1.45	-0.901	1.29	-0.660	1.39
Emotional	0.021	1.66	-1.124	1.49	-0.749	1.63	-0.281	1.69	-0.629	1.51	-0.558	1.76
Second-order												
TFI - 19	0.095	0.47	-0.126	0.32	-0.057	0.42	0.046	0.38	-0.058	0.31	-0.001	0.45
TFI - 18	0.101	0.47	-0.126	0.32	-0.057	0.43	0.053	0.38	-0.059	0.32	0.003	0.46

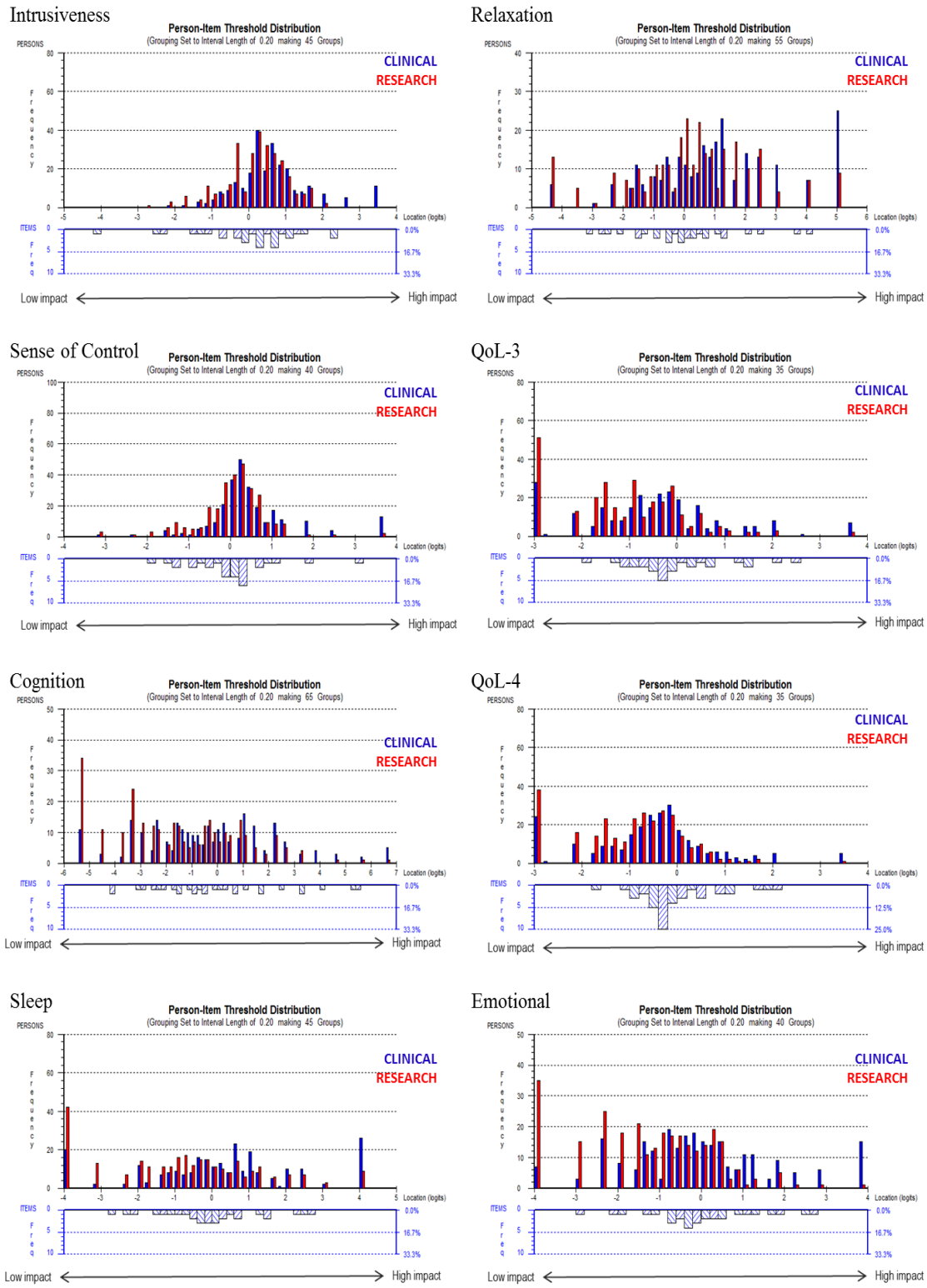

Figure 6.25. Person-item threshold distributions in Intrusiveness, Sense of Control, Cognition, Sleep, Auditory, QoL-3, QoL-4 and Emotional subscales showing differences in targeting locations for clinical and research populations.
Participants from a research population (red) are located lower on the logit scale than participants from a clinical population (blue).

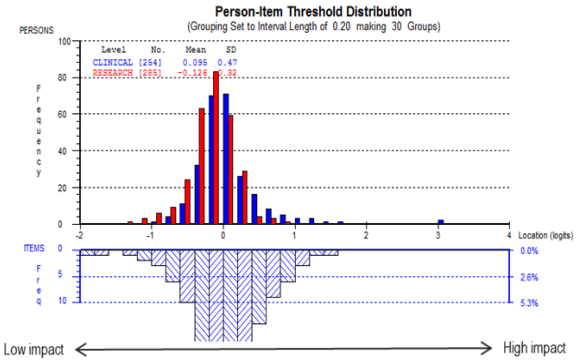

Figure 6.26. Person-item threshold distributions in TFI-18 showing differences in targeting locations for clinical and research populations
Participants from a research population (red) are located slightly lower on the logit scale than participants from a clinical population (blue). Unfortunately RUMM2030 software does not provide the full plot for the TFI-18.

A summary of the population DIF results for the items residuals within each subscale are presented in Table 6.27. INTR3, SOC5, REL17, REL18 and QOL21 (both versions of the subscale) showed main effects of populations with probability values exceeding the adjusted alpha ($p>0.05$). Inspection of the item characteristics curves indicated that given equal levels of tinnitus impact (i.e. location) in four of the items, the research population in general scored slightly lower across the continuum measured by the subscale (Figure 6.27). For example, QOL21 showed consistent uniform differences in population responses at each class interval (Figure 6.27). The Relaxation subscale was the only scale to have more than one item with DIF (REL17, REL18). However, they displayed opposing results. REL17 displayed the same results as above, whilst for REL18, the clinical population consistently scored lower across the continuum, despite equivalent estimated person locations (Figure 6.27). The population differences in this subscale were opposing and as a consequence the population effects were cancelled out when the overall score for subscale was calculated. Therefore there was no need to make any adjustments to the items. The results for the 6-factor structures indicated that only the Emotional testlet, showed significant differences in population (Table 6.28).

Table 6.27. ANOVA results for the differences in item functioning between population and gender factor groups in the designated subscales

		Population		Population X CInt		Gender		Gender X CInt	
	Items	F	*p**	F	*p**	F	*p**	F	*p**
INTR	INTR1	1.387	1.000	0.431	1.000	0.234	1.000	0.832	1.000
	INTR2	0.194	1.000	1.946	1.000	0.871	1.000	1.127	1.000
	INTR3	10.344	**0.009**	0.731	1.000	1.631	1.000	2.588	0.324
SOC	SOC4	0.956	1.000	0.446	1.000	0.139	1.000	0.544	1.000
	SOC5	12.917	**>0.001**	-0.388	1.000	0.752	1.000	1.179	1.000
	SOC6	0.055	1.000	1.337	1.000	1.692	1.000	0.144	1.000
COG	COG7	0.044	1.000	0.500	1.000	0.878	1.000	3.138	0.135
	COG8	0.418	1.000	2.292	0.531	0.359	1.000	0.401	1.000
	COG9	2.573	0.981	1.375	1.000	0.097	1.000	3.007	0.162
SLP	SLP10	4.029	0.405	0.787	1.000	4.655	0.288	0.510	1.000
	SLP11	0.771	1.000	1.843	1.000	0.148	1.000	0.153	1.000
	SLP12	1.176	1.000	1.084	1.000	2.096	1.000	0.363	1.000
AUD	AUD13	3.938	0.432	0.466	1.000	4.740	0.270	1.427	1.000
	AUD14	0.000	1.000	0.776	1.000	0.357	1.000	1.166	1.000
	AUD15	0.979	1.000	0.198	1.000	3.140	0.693	2.914	0.189
REL	REL16	1.345	1.000	1.205	1.000	6.347	0.108	1.064	1.000
	REL17	27.610	**>0.001**	-0.270	1.000	0.419	1.000	2.155	0.657
	REL18	26.072	**>0.001**	0.309	1.000	0.373	1.000	2.225	0.585
QoL-4	QOL19	1.067	1.000	1.530	1.000	0.002	1.000	0.126	1.000
	QOL20	1.875	1.000	1.059	1.000	0.336	1.000	1.026	1.000
	QOL21	15.520	**>0.001**	-1.658	1.000	0.810	1.000	0.373	1.000
	QOL22	2.145	1.000	0.849	1.000	0.095	1.000	1.104	1.000
QoL-3	QOL19	2.910	0.801	1.247	1.000	0.033	1.000	1.146	1.000
	QOL20	0.090	1.000	1.478	1.000	0.170	1.000	0.889	1.000
	QOL21	10.965	**0.009**	-0.048	1.000	0.644	1.000	1.599	1.000
EMO	EMO23	4.228	0.360	0.259	1.000	4.308	0.342	1.116	1.000
	EMO24	0.460	1.000	0.579	1.000	5.023	0.225	1.190	1.000
	EMO25	0.422	1.000	0.606	1.000	8.262	**0.036**	0.831	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5.

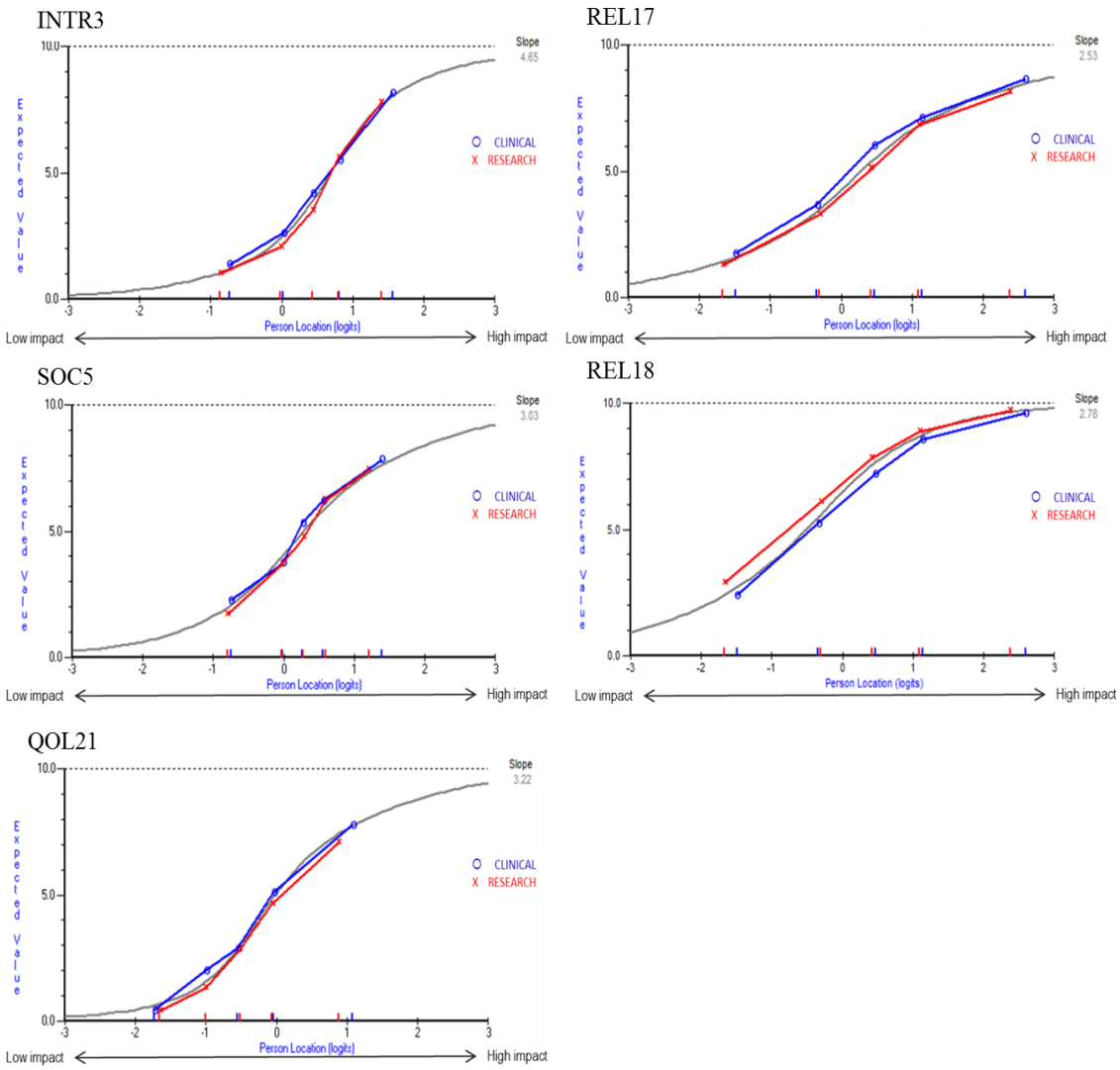

Figure 6.27. Item characteristic curve for INTRU3, SOC5, QOL21, REL17, and REL18 showing differences in item functioning between clinical and research populations.
Given equal levels of tinnitus impact, the clinical population (blue) consistently scored higher than the research population (red) for INTR3, SOC5 and QOL21 (uniform DIF). In REL17, the clinical population consistently scored higher than the research population, whilst in REL18, the clinical population unusually consistently scored lower (uniform DIF).

Again, given equal levels of tinnitus impact, the clinical participants were more likely to endorse higher scores throughout the continuum (Figure 6.28). The solution to the observed population DIF was to separately assess clinical and research populations by splitting and recalibrating the items/testlet based on population. With the exception of the Relaxation items, two new items/testlets were created for each item/testlet that displayed DIF.

Table 6.28. ANOVA results for the differences in item functioning between population and gender factor groups in the six-factor structure (TFI-19/TFI-18).

		Population		Population X C*Int*		Gender		Gender X C*Int*	
	Testlets	F	*p**	F	*p**	F	*p**	F	*p**
TFI -19	Intrusiveness	0.183	1.000	0.717	1.000	1.356	1.000	1.143	1.000
	Sense of control	1.375	1.000	0.564	1.000	11.312	**0.018**	0.291	1.000
	Cognition	0.004	1.000	1.561	1.000	0.803	1.000	0.772	1.000
	Relaxation	4.799	0.522	0.213	1.000	0.709	1.000	0.766	1.000
	Quality of life - 4	1.360	1.000	1.102	1.000	22.914	**>0.001**	1.257	1.000
	Emotional	35.876	**>0.001**	-0.832	1.000	0.989	1.000	0.308	1.000
TFI-18	Intrusiveness	0.276	1.000	0.393	1.000	1.054	1.000	1.469	1.000
	Sense of control	1.110	1.000	0.789	1.000	10.193	**>0.001**	0.060	1.000
	Cognition	0.061	1.000	2.188	1.000	0.305	1.000	0.706	1.000
	Relaxation	5.696	1.000	0.531	1.000	0.296	1.000	1.365	1.000
	Quality of life - 3	0.286	1.000	0.329	1.000	19.094	**>0.001**	1.133	1.000
	Emotional	33.961	**>0.001**	-1.860	1.000	0.427	1.000	0.205	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5.

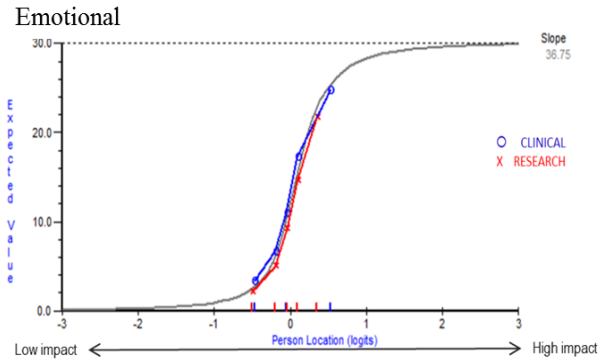

Figure 6.28. Item characteristic curve for the Emotional testlet showing differences in item functioning between clinical and research populations.
Given equal levels of tinnitus impact, the clinical population consistently scored slightly higher than the research population (uniform DIF).

For example, INTR3 became INTR3Cl (clinical) and INTR3Re (research). Consequently, this created two separate subscales for Intrusiveness, Sense of Control and QoL and two separate overall second-order scales with unbiased estimates of tinnitus impact that can be used in clinic or in research. Following this, the overall fit for items and six-factor structure were reassessed. For the most part the fit statistics improved, with the exception of the overall $\chi 2$ estimate which increased with the additional items (Tables 6.22, 6.29 and 6.30). It is worth mentioning that in QoL-4 subscale, QOL20 had a large $\chi 2$ estimate which was previously associated with QOL21, whilst in QoL-3 there were no deviations to report (Table 6.30).

6.3.5.2.2 Gender

Examination of the gender differences in the distribution of the person locations revealed significant differences for six subscales and the six-factor structure (Table 6.25). The mean location for males was significantly lower than females in all cases (Table 6.26). The items lower on the continuum were therefore more likely to target males than the items higher on the continuum (Figure 6.29). In general however, all six subscales items were reasonably well distributed across the continuum and would be appropriately targeted for both genders responses (Figure 6.29).

Table 6.29. Summary fit statistics for Intrusiveness, Sense of control, Quality of life (3/4) and Emotional subscales and 6-factor structure (TFI-18/TFI-19) following item recalibration for population and gender Differential Item Functioning.

			Item fit residual		Person fit residual		Item-trait interaction	
DIF split	**Subscale**	**Items**	Mean	SD	Mean	SD	χ^2 (df)	*p*
Item split for population DIF	Intrusiveness§	INTR1, INTR2, **INTR3Cl, INTR3Re**	-0.28	0.98	-0.51	0.98	32.56 (24)	0.456
	Sense of control§	SOC4,**SOC5Cl, SOCRe**, SOC5	-0.14	0.81	-0.51	1.06	37.69 (24)	0.148
	Quality of life - 4‡	QOL19, QOL20, **QOL21Cl, QOL21Re**, QOL22	-0.24	1.73	-0.61	1.26	49.34 (30)	0.070
	Quality of life - 3§	QOL19, QOL20, **QOL21Cl, QOL21Re**,	-0.16	0.32	-0.62	1.09	39.20 (24)	0.104
Item split for gender DIF	Emotional	EMO23, EMO24, **EMO25M, EMO25F**	-0.388	1.20	-0.547	1.01	39.53 (24)	0.096
	Second order	**Testlets**						
Testlet split for population DIF	TFI – 19*	INTR, SOC, COG, REL, QoL-4, **EMOCl, EMORe**	0.31	2.42	-0.34	1.07	56.81 (42)	0.441
	TFI – 18*	INTR, SOC, COG, REL, QoL-3, **EMOCl, EMORe**	0.36	2.2	-0.34	1.07	51.87 (42)	0.987

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5.

Table 6.30. Individual item fit statistics for Intrusiveness, Sense of control, Quality of life (3/4) and Emotional subscales following item recalibration for population and gender Differential Item Functioning.

DIF		Items	Location	SE	Fit residual	χ^2	*df*	*p*
Population	INTR§	INTRU1	-0.383	0.027	-0.76	11.094	6	0.344
		INTRU2	-0.632	0.031	1.159	5.517	6	1.000
		INTR3Cl	0.39	0.039	-0.981	8.000	6	0.952
		INTR3Re	0.625	0.034	-0.554	7.952	6	0.968
	SOC§	SOC4	-0.308	0.022	0.671	8.344	6	0.856
		SOC5Cl	0.119	0.04	-0.902	9.262	6	0.636
		SOC5Re	0.561	0.037	-0.781	12.944	6	0.176
		SOC6	-0.372	0.025	0.437	7.143	6	1.000
	QoL-4‡	QOL19	-0.023	0.027	0.009	8.074	6	1.000
		QOL20	-0.283	0.027	-1.277	**16.348**	6	0.050
		QOL21Cl	0.051	0.039	-1.284	8.561	6	1.000
		QOL21Re	0.216	0.039	-1.368	9.38	6	0.765
		QOL22	0.039	0.027	**2.678**	6.983	6	1.000
	QoL-3§	QOL19	-0.021	0.03	0.052	14.366	6	0.104
		QOL20	-0.322	0.029	-0.388	11.752	6	0.272
		QOL21Cl	0.072	0.042	0.183	6.588	6	1.000
		QOL21Re	0.272	0.044	-0.473	6.5	6	1.000
Gender	EMO§	EMO23	-0.18	0.031	-0.848	**17.066**	6	0.036
		EMO24	-0.565	0.032	-1.866	11.494	6	0.296
		EMO25M	0.198	0.038	0.445	10.299	6	0.452
		EMO25F	0.547	0.053	0.717	0.669	6	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5.

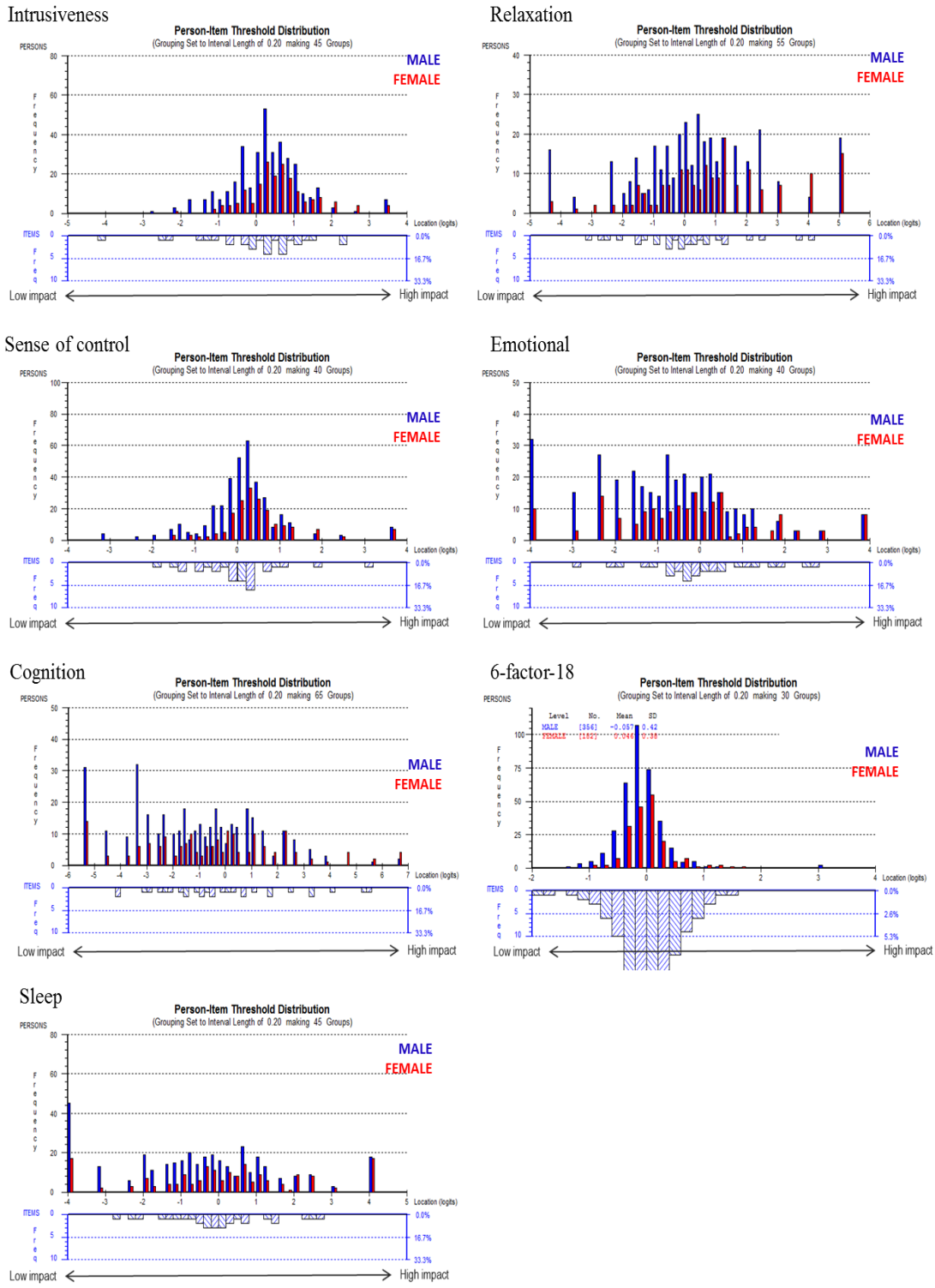

Figure 6.29. Person-item threshold distributions in Intrusiveness, Sense of Control, Cognition, Sleep, Relaxation and Emotional subscales and TFI-18 showing differences in targeting locations for gender.

Males (blue) are located lower on the logit scale than females (red), but items in general cover all person locations for both genders. Unfortunately, RUMM2030 software does not provide the full plot for the TFI-18.

A summary of the gender DIF results for the items residuals within each subscale are presented in Table 6.27. Only EMO25 (*How DEPRESSED were you because of your tinnitus?*) showed significant differences in gender alone. No interaction was observed between class interval and gender. In Figure 6.30 (which plots EMO25), the graph clearly reveals that in general, given similar estimated person locations, males consistently identified higher depression in relation to their tinnitus across all levels of tinnitus impact. For example, males experiencing mild tinnitus endorse higher response categories than females experiencing the same level of tinnitus. This item (EMO25) was split into two separate items that assess gender alone, providing separate scores for each.

Although the summary fit statistics improved (Table 6.29) and there was no change to individual person fit statistics, the individual item statistics were slightly inflated (Table 6.30). Following the item recalibration, the item residuals all increased, all were still within criteria though, and a large significant χ^2 estimate was now observed for EMO23. This could be an indication that this item is not measuring the same construct as the others in the scale. However, the alignment of observed scores to the characteristic curve was slightly over-discriminating, with only two class intervals deviating from the curve (Figure 6.31). This slight deviation was not apparent previously with the full dataset, it is therefore assumed that this deviation is the product of the additional item parameter. Given this problem and the fact that people did not always identify with one particular gender it was decided it would not be feasible to split this item.

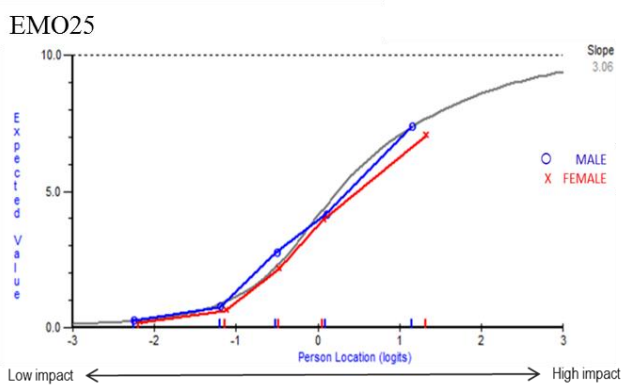

Figure 6.30. Item characteristic curve for the EMO25 showing differences in item functioning between males and females.
Females (red) consistently scored lower than the males (blue), given the same level of tinnitus impact.

In the six-factor structure (TFI-19/TFI-18), there was a significant gender effect for two testlets (Table 6.28). Responses in the Sense of Control testlet varied with males more likely to endorse higher response categories than females, with similar overall level of tinnitus impact (Figure 6.32). However, the responses in the QoL testlet were the opposite with females endorsing higher response categories. Consequently, although these were not invariant in terms of gender, there was no need to split the items since the two opposing effects of gender cancelled each other out in the overall score.

6.3.5.2.3 Age

The sample size in the class intervals for the 70yrs+ age group was too small to make meaningful comparisons with the other age groups. So for the DIF analysis, the data were amalgamated into two age groups (<50 and 50-70+) (Table 6.31). The distributions of person locations were significantly different in age for the Intrusiveness and Auditory subscales only (Table 6.25). The mean person locations for people aged 50 to 70+ years were significantly higher on the continuum than the locations for people <50 years (Table 6.26).

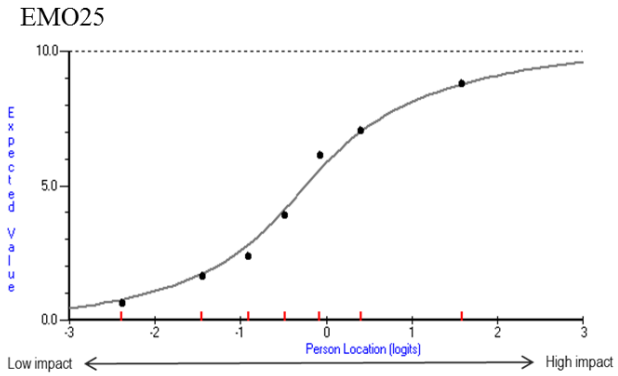

Figure 6.31. Item characteristic curve for EMO23 showing two of seven class intervals slightly deviating from the expected curve following recalibration of EMO25.

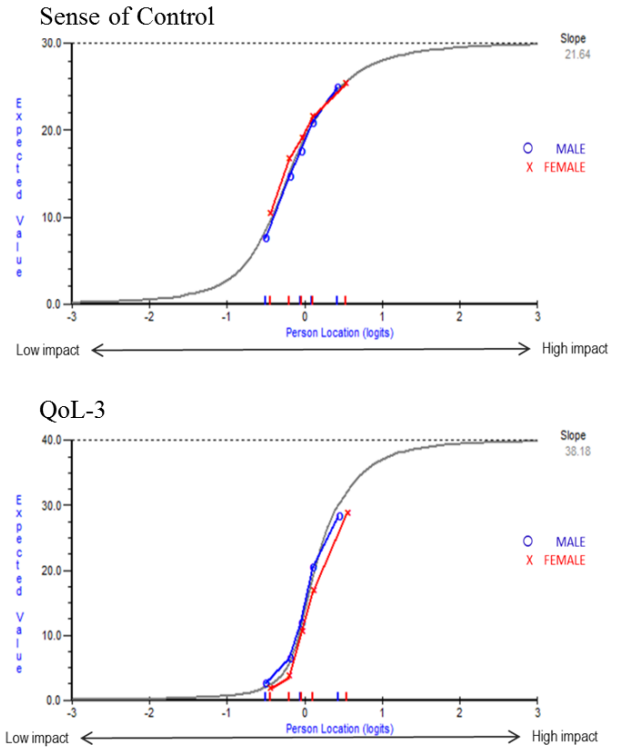

Figure 6.32. Item characteristic curve for the Sense of control and Quality of life-3 testlets showing opposing differences in item functioning between gender.
For Sense of Control testlet males (blue) consistently scored lower than the females (red), given the same level of tinnitus impact, whilst QoL-3 shows the opposite results.

Older people had higher levels of tinnitus intrusiveness than younger people. In terms of targeting, the intrusiveness subscale had limited items located at the higher extremes on the continuum, therefore it can potentially differentiate people below 50 better than those above (Figure 6.33).

Table 6.31. Sample size frequency for initial and collapsed person factor groups for age, self-defined hearing and hearing thresholds.

	Initial		Collapsed groups	
Person factor	Person factor groups	N	Person factor groups	N
Age	<50 yrs	195	<50 yrs	195
	50 – 69 yrs	297	50 – 70+ yrs	345
	70 + yrs	48		
Self-defined hearing	No problem	69	No problem	69
	Small problem	76	Hearing problems	186
	Moderate	77		
	Big problem	27		
	Very big problem	6		
BSA hearing thresholds (PTA)*	Normal hearing	181	Normal hearing	181
	Mild loss	72	Hearing loss	103
	Moderate loss	28		
	Severe loss	3		

The Auditory subscale had poor targeting, so was not considered in this context.

A summary of the age DIF results for the items residuals within each subscale is presented in Table 6.32. Two items within the QoL-4 subscale showed age differences exceeding adjusted probability values. QOL19 and QOL22 displayed opposing age effects (Figure 6.34) and consequently no adjustments were needed. Furthermore, the analysis excluding QO22 (QoL-3) resulted in no evidence of age effects in any items. Therefore, those effects identified could be due to QOL22 as without this item the effects disappear.

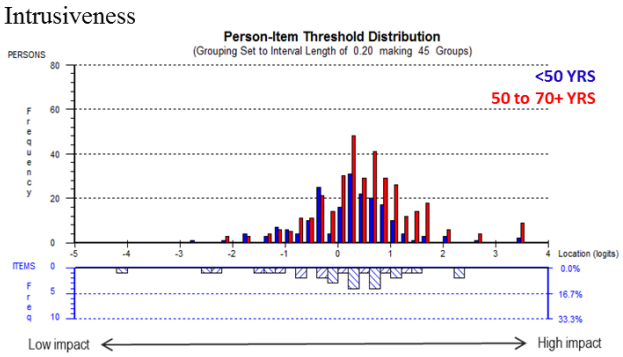

Figure 6.33. Person-item threshold distributions in Intrusiveness subscales showing differences in targeting locations for age groups.
Older participants (red) are located higher on the logit scale than younger participants (blue), but other than the extremes all person locations are covered by item threshold locations.

In the six-factor structure, there were again only two testlets showing significant age effects; Intrusiveness and Cognition (Table 6.33). Closer inspection of the item characteristic curves revealed that these two testlets were reflecting opposing age effects (Figure 6.35). Once again there was no need to adjust for these effects as they did not impact on the overall structure score or violate the assumption of Rasch.

6.3.5.2.4 Hearing

Due to the number of groups in both the self-reported hearing loss and hearing thresholds person factors, the class intervals were reduced from 5 to 2 to enable meaningful comparisons with larger sample sizes. Unfortunately, the sample size in the hearing thresholds groups was not sufficient to maintain the British Society of Audiology categories. The mild, moderate and severe hearing thresholds (>20 dB) were amalgamated into a single group to compare against the normal hearing group (<20 dB) (Table 6.31). The problem with sample size was also evident in the self-reported hearing loss person factor groups. To circumvent the small sample sizes, the five self-reported hearing loss groups were collapsed.

Table 6.32. ANOVA results for the differences in item functioning between age factor groups all items in designated factor.

		Age		Age X CInt	
	Items	F	*p**	F	*p**
INTR*	INTR1	4.685	0.279	0.604	1.000
	INTR2	0.001	1.000	0.546	1.000
	INTR3	2.166	1.000	1.478	1.000
SOC*	SOC4	4.942	0.243	0.497	1.000
	SOC5	2.157	1.000	0.363	1.000
	SOC6	2.067	1.000	0.526	1.000
COG*	COG7	0.117	1.000	0.306	1.000
	COG8	0.181	1.000	1.175	1.000
	COG9	0.023	1.000	0.676	1.000
SLP*	SLP10	2.680	0.918	0.764	1.000
	SLP11	0.672	1.000	1.094	1.000
	SLP12	1.410	1.000	1.593	1.000
AUD*	AUD13	2.902	0.801	0.184	1.000
	AUD14	0.643	1.000	0.285	1.000
	AUD15	3.952	0.423	0.294	1.000
REL*	REL16	0.564	1.000	1.371	1.000
	REL17	1.404	0.657	2.155	0.438
	REL18	0.462	0.585	0.874	1.000
QoL-4§	QOL19	8.417	**0.048**	1.393	1.000
	QOL20	0.000	1.000	0.765	1.000
	QOL21	7.105	1.000	-0.057	1.000
	QOL22	16.815	**>0.001**	0.720	1.000
QoL-3*	QOL19	2.024	1.000	0.872	1.000
	QOL20	3.757	0.477	1.679	1.000
	QOL21	1.720	1.000	0.188	1.000
EMO*	EMO23	0.445	1.000	1.224	1.000
	EMO24	0.053	1.000	0.967	1.000
	EMO25	0.226	1.000	0.755	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 2.

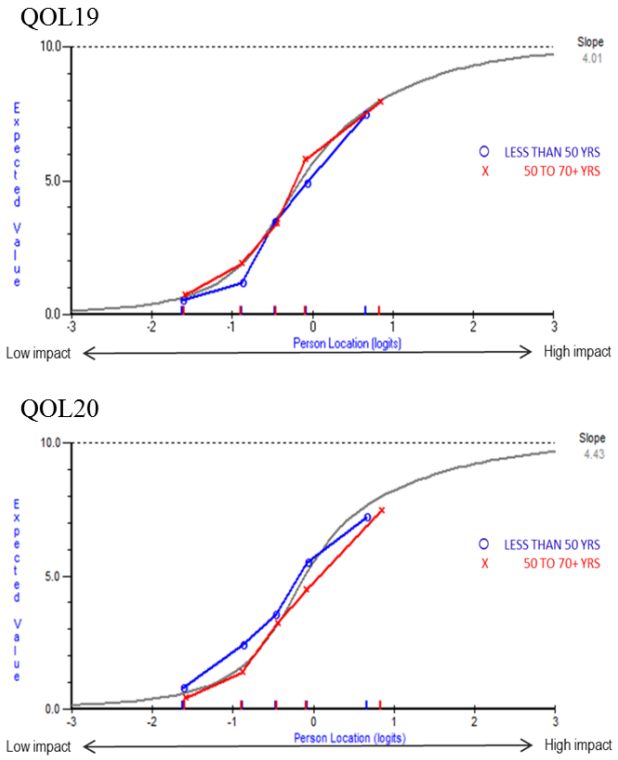

Figure 6.34. Item characteristic curve for QOL19 and QOL22 (QoL-4) showing opposing differences in item functioning responses between age groups.
For QOL19 participants who are less than 50 yrs consistently score lower than those who are older, whilst QOL20 shows the opposite results.

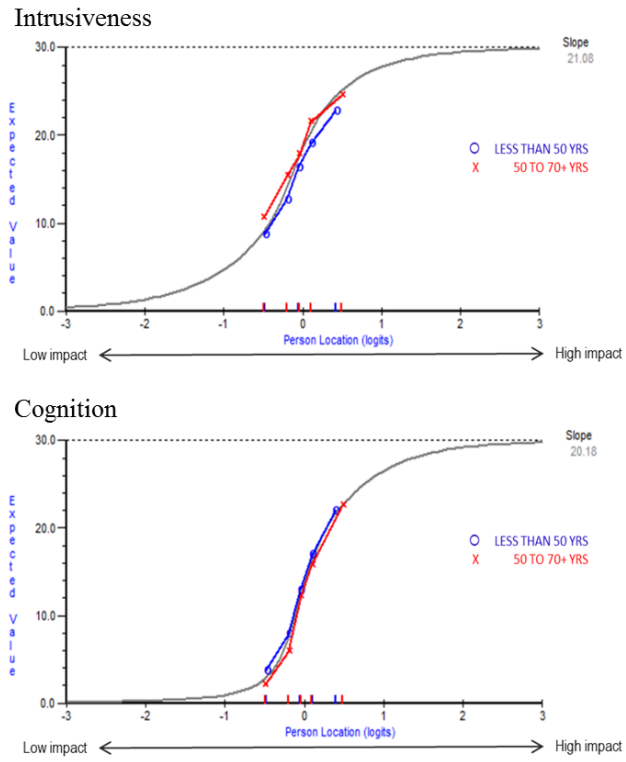

Figure 6.35. Item characteristic curve for Intrusiveness and Cognition testlets showing opposing differences in item functioning responses between age groups.
For the Intrusiveness testlet, participants who are less than 50 yrs consistently score lower than those who are older, whilst the Cognition testlet shows the opposite results.

Table 6.33. ANOVA results for the differences in item functioning between age and hearing factor groups in the 6-factor structure (TFI-19/TFI-18).

						Clinical				Research			
		Age		Age X C*Int*		Hearing		Hearing X C*Int*		Hearing		Hearing X C*Int*	
	Testlets	F	*p*	F	*p*	F	*p*	F	*p*	F	*p*	F	*p*
TFI-19	Intrusiveness	24.575	>**0.001**	0.666	1.000	0.241	1.000	7.379	1.000	7.138	0.144	1.309	1.000
	Sense of control	0.805	1.000	0.758	1.000	1.232	1.000	0.922	1.000	3.597	1.000	-0.001	1.000
	Cognition	11.116	**0.018**	1.039	1.000	9.282	**0.054**	-2.662	1.000	0.036	1.000	0.431	1.000
	Relaxation	3.634	1.000	1.999	1.000	8.148	0.090	1.079	1.000	16.285	>**0.001**	0.089	1.000
	Quality of life - 4	0.145	1.000	1.381	1.000	31.006	>**0.001**	3.290	1.000	1.752	1.000	-0.081	1.000
	Emotional	1.938	1.000	0.609	1.000	19.904	>**0.001**	3.619	1.000	2.862	1.000	1.630	1.000
TFI -18	Intrusiveness	23.499	>**0.001**	0.566	1.000	0.595	1.000	6.150	0.252	6.937	0.162	0.624	1.000
	Sense of control	0.506	1.000	1.068	1.000	0.712	1.000	0.791	1.000	3.411	1.000	-0.044	1.000
	Cognition	12.409	>**0.001**	1.079	1.000	11.841	**0.018**	-1.811	1.000	0.091	1.000	0.712	1.000
	Relaxation	4.530	0.612	1.745	1.000	6.496	0.198	0.754	1.000	17.161	>**0.001**	0.014	1.000
	Quality of life - 3	2.491	1.000	0.948	1.000	20.593	>**0.001**	1.053	1.000	3.628	1.000	-0.012	1.000
	Emotional	2.571	1.000	0.564	1.000	16.660	>**0.001**	2.342	1.000	3.155	1.000	1.496	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 5 for Age comparisons; 2 for hearing comparisons.

Initially, the small hearing problem and moderate problem groups were combined, as were the big problem to very big problem groups creating three hearing groups (no problem, small to moderate problem, big to very big problem). However once again the numbers in the big to very big problem group (n = 33) were not sufficient to conduct the analysis.

Thus, the self-defined hearing groups were also amalgamated into two hearing groups (no problem vs hearing problems). Additionally to reduce the number of comparisons, the hearing person factor groups for the clinical population were examined separately from the hearing thresholds groups for the research population. Closer inspection of the class intervals for each subscale and 6-factor analysis revealed that despite these attempts to improve sample size, in some cases the sample sizes for each class interval were not large enough to make strong conclusions about the differences in item functioning and targeting. Accordingly, the results were indicative rather than conclusive, and as a consequence no adjustments to items were made.

The distributions of person locations were significantly different in hearing groups for the Cognition, Auditory and QoL subscales with the clinical data, whilst only the person locations in the Auditory subscale were significantly different for hearing thresholds with the research data (Table 6.34). The mean person locations for people self-defined as having “no problem” with hearing and those with normal hearing were significantly lower on the continuum than the locations for people self-defined as having “hearing problems” and those with hearing loss (Table 6.35). Predictably, the difference was most apparent for the Auditory subscale, with the average location for the “no problem” group located more than 3.4 logits from the hearing problem groups.

Table 6.34. ANOVA results for differential functioning in targeting for hearing groups in the TFI subscales and 6-factor structure (TFI-19/TFI-18).

	Hearing			
	Clinical		**Research**	
Items	F	*p*	F	*p*
Intrusiveness	2.338	0.666	1.892	1.000
Sense of control	0.752	1.000	0.777	1.000
Cognition	5.354	**0.009**	0.058	1.000
Sleep	0.007	1.000	0.199	1.000
Auditory	58.461	**>0.001**	13.841	**>0.001**
Relaxation	0.270	1.000	0.569	1.000
Quality of life - 4	10.180	**>0.001**	0.290	1.000
Quality of life - 3	8.254	**>0.001**	0.614	1.000
Emotional	0.172	1.000	0.074	1.000
Second-order				
Six-factor TFI - 19	0.117	1.000	3.2116	0.424
Six-factor TFI - 18	0.1438	1.000	2.7456	0.783

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 2.

In fact, the majority of people with extreme low scores (> -5 logits) had normal hearing or "no problems" with hearing (Figure 6.36). This further highlights the problems with the Auditory items not necessarily measuring hearing problems in relation to tinnitus; the responses to the subscale could be solely based on hearing problems alone. In general, the Cognition and QoL items were reasonably well distributed across the continuum and were still covering the different ranges in person locations for hearing (Figure 6.36). The Auditory subscale has poor targeting so is not considered in this context.

Table 6.35. Mean locations and standard deviations for population, gender and age person factor groups targeting.

	Clinical				Research			
	No problem		Problem with hearing		Normal hearing		Hearing loss	
Items	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Intrusiveness	0.523	1.20	0.914	0.98	0.112	0.75	0.334	0.74
Sense of control	0.426	1.09	0.634	0.94	-0.045	0.75	0.120	1.04
Cognition	**-1.451**	2.31	-0.011	2.62	-1.479	2.26	-1.363	2.38
Sleep	0.315	2.35	0.363	2.26	-0.575	1.73	-0.751	2.04
Auditory	**-3.719**	2.25	0.305	2.11	**-2.199**	2.47	-0.313	2.16
Relaxation	0.872	2.37	1.155	2.17	0.332	2.14	-0.010	2.08
Quality of life – 4	**-1.142**	1.24	-0.198	1.19	-0.990	1.12	-0.863	1.05
Quality of life - 3	**-1.157**	1.44	-0.189	1.35	-1.122	1.31	-0.910	1.17
Emotional	-0.101	1.66	0.066	1.63	-0.166	1.58	-1.254	1.44
Second-order								
TFI - 19	-0.047	0.32	0.152	0.49	-0.152	0.35	-0.128	0.31
TFI - 18	-0.034	0.34	0.153	0.50	-0.151	0.35	-0.124	0.31

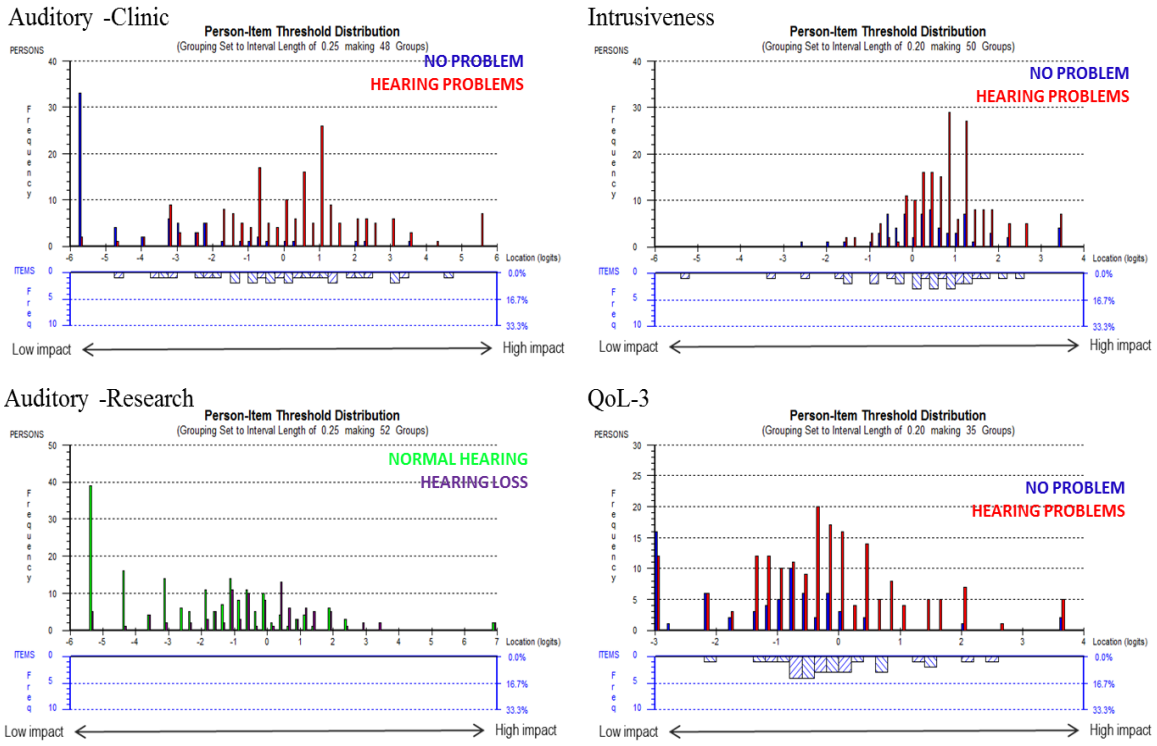

Figure 6.36. Person-item threshold distributions in Auditory, Intrusiveness and Quality of life subscales showing differences in targeting locations for hearing groups (no problem/hearing problems (clinical)) and hearing thresholds (Normal hearing/Hearing loss (research)).
The Auditory subscale shows targeting differences for both hearing measurements, whist Intrusiveness and QoL show targeting differences with perceived hearing loss alone (red/blue). Both show that participants with no problem with hearing were located lower on the logit scale.

A summary of the hearing DIF results for the items within each subscale are presented in Table 6.36. Significant differences in hearing were only observed with the clinical data. INTR1, AUD15, QOL20 and QOL21 (both versions of subscale) showed significant main effects of populations. Inspection of the item characteristics curves indicated that given equal levels of tinnitus impact in three of the four items, people with "no problem" with hearing in general scored lower across the continuum measured by the subscale (Figure 6.37). For example, in AUD15, people with hearing problems closely follow the expected characteristic curve, whilst people with no problems consistently report lower levels of tinnitus problems, despite having equivalent estimated person locations as the hearing problem group (Figure 6.37).

Table 6.36. ANOVA results for the differences in item functioning between hearing factor groups.

		Clinical				Research			
		Hearing*		Hearing X CInt*		Hearing*		Hearing X CInt*	
	Items	F	*p**	F	*p**	F	*p**	F	*p**
INTR*	INTR1	8.991	**0.027**	-0.338	1.000	2.455	1.000	-0.503	1.000
	INTR2	2.238	1.000	3.236	1.000	0.049	1.000	0.604	1.000
	INTR3	0.211	1.000	0.817	1.000	1.379	1.000	0.872	1.000
SOC*	SOC4	0.531	1.000	1.478	1.000	0.273	1.000	1.387	1.000
	SOC5	0.186	1.000	5.195	0.216	0.476	1.000	2.478	1.000
	SOC6	0.467	1.000	0.323	1.000	0.119	1.000	0.213	1.000
COG*	COG7	0.571	1.000	0.564	1.000	0.282	1.000	0.074	1.000
	COG8	0.035	1.000	0.432	1.000	0.029	1.000	0.323	1.000
	COG9	0.006	1.000	0.087	1.000	0.226	1.000	0.372	1.000
SLP*	SLP10	5.689	0.162	0.000	1.000	0.055	1.000	1.911	1.000
	SLP11	0.021	1.000	1.703	1.000	0.161	1.000	5.235	0.207
	SLP12	5.343	0.198	0.557	1.000	0.093	1.000	-0.012	1.000
AUD*	AUD13	7.652	**0.054**	1.012	1.000	5.677	0.162	2.927	0.792
	AUD14	3.263	1.000	-1.853	1.000	2.083	1.000	-1.388	1.000
	AUD15	9.255	**0.027**	-0.655	1.000	5.922	0.144	3.359	0.612
REL*	REL16	2.667	0.936	4.567	0.306	0.020	1.000	0.040	1.000
	REL17	0.539	1.000	0.299	1.000	0.337	1.000	0.316	1.000
	REL18	5.419	0.189	4.009	0.414	0.577	1.000	0.245	1.000
QoL-4§	QOL19	0.052	1.000	3.485	0.756	1.821	1.000	0.013	1.000
	QOL20	17.504	**>0.001**	7.956	0.060	0.020	1.000	0.072	1.000
	QOL21	10.277	**0.018**	-4.176	1.000	3.750	0.486	-0.487	1.000
	QOL22	6.928	0.108	4.356	0.456	3.567	0.720	0.093	1.000
QoL-3*	QOL19	1.355	1.000	1.454	1.000	0.529	1.000	-0.070	1.000
	QOL20	14.825	**>0.001**	5.547	0.171	0.969	1.000	0.400	1.000
	QOL21	17.821	**>0.001**	1.191	1.000	1.726	1.000	-0.533	1.000
EMO*	EMO23	0.089	1.000	0.062	1.000	1.727	1.000	0.363	1.000
	EMO24	3.116	0.711	0.699	1.000	1.748	1.000	0.092	1.000
	EMO25	2.243	1.000	1.081	1.000	0.000	1.000	0.082	1.000

* corrected for multiple comparisons. Bold = significant values. n = 540. Class intervals = 2.

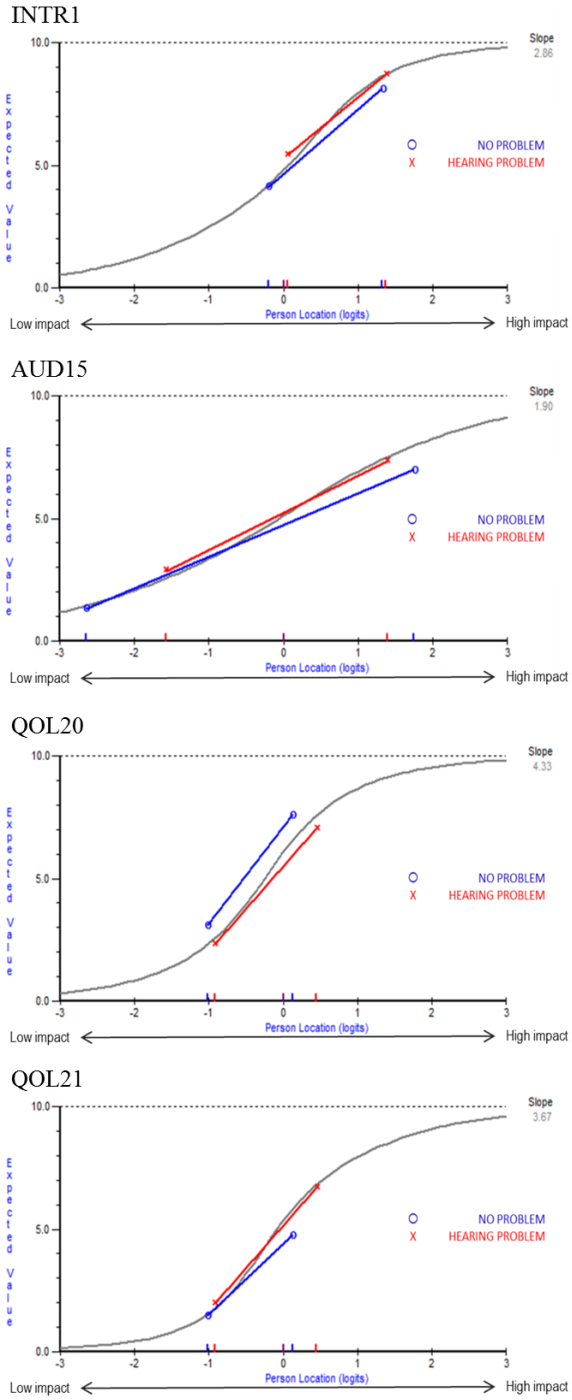

Figure 6.37. Item characteristic curves for the INTR1, AUD15, QOL20 and QOL21 showing differences in item functioning between persons with self-reported hearing problems and no problem hearing.
For INTR1, AUD15, QOL21 participants with hearing problems consistently scored higher than those without, given the same level of tinnitus impact, whilst QOL20 showed the opposite result.

The two items with DIF in the QoL subscale displayed opposing results. QOL21 displayed the same results as above, whilst for QOL20, the "no problem"

group consistently scored higher across the continuum (Fig. 6.37). The hearing effects were therefore cancelled out when the overall score for subscale was calculated.

In the six-factor structure (TFI-19/TFI-18), there were three testlets showing significant hearing effects in the clinical data; Cognition, QoL (4/3) and Emotional, and one testlet in the research data; Relaxation (Table 6.33). For the Cognition and QoL (4/3) testlets, people with no problems consistently reported lower levels of tinnitus (Fig.6.38), whilst for the Emotional and Relaxation testlets, given equal levels of tinnitus impact, people with hearing problems/hearing loss consistently scored lower on the continuum than normal hearing, no problem groups, which are closely aligned to the expected characteristic curve (Figure 6.38). In other words, people with hearing problems were less likely to report emotional difficulties or problems with relaxation, despite having equal levels of overall tinnitus impact as the normal hearing, no problem groups. No adjustments were made as the impact of different levels of hearing could not be fully evaluated and we cannot say with confidence that these differences violate the assumption of Rasch. Therefore the items and testlets remained unchanged and, with the exception of the Auditory subscale, were assumed to conform to the Rasch model.

6.3.5.3 Transforming the raw scores

Linear transformations were only conducted on seven of the eight subscales. Linear transformations were conducted on the TFI-18 (six-factor structure). The Auditory subscale consistently failed to conform to the Rasch model, and as per a priori criteria, the logit scores were not converted into understandable metric scores.

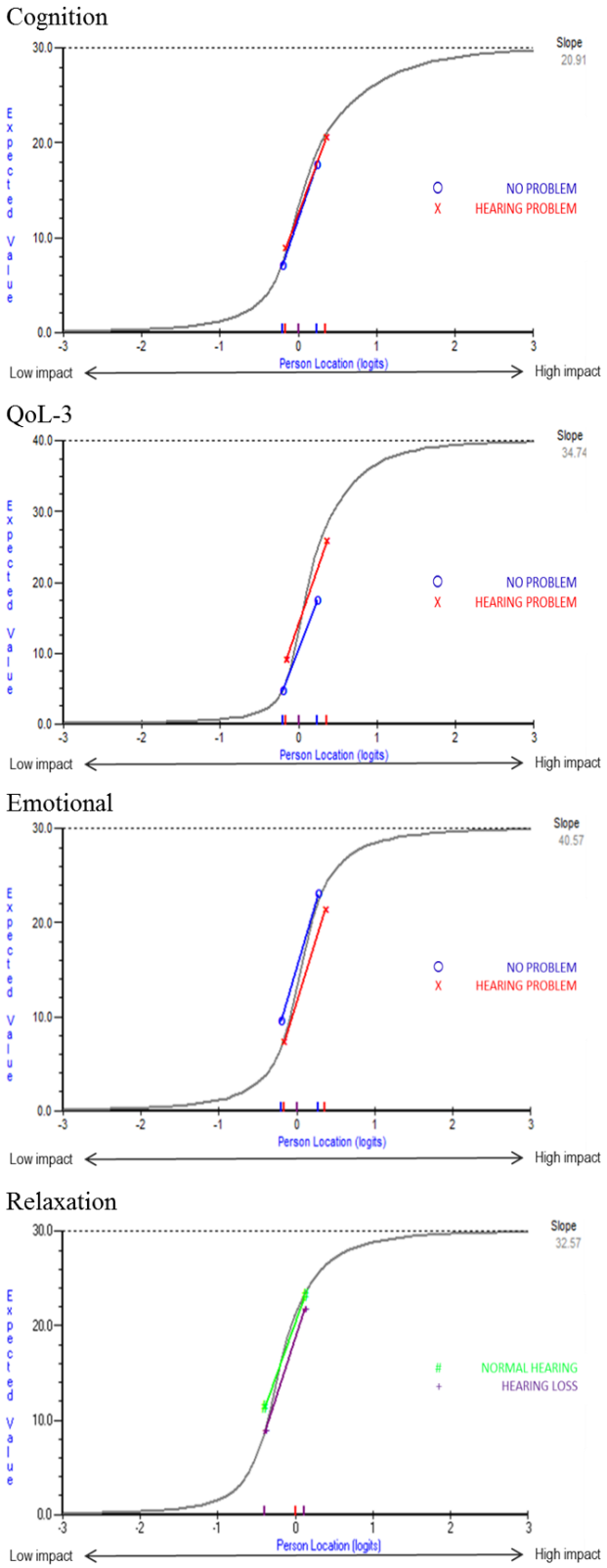

Figure 6.38. Item characteristic curves for the Cognition, QoL-3, Emotional and Relaxation testlets showing differences in item functioning between hearing groups.
For the Cognition and QoL-3 testlets, participants with hearing problems/hearing loss (red) consistently scored higher than those without (blue); whilst the Emotional and Relaxation testlets showed the opposite.

Although the Sleep subscale potentially showed some overlap in contents, this was not consistent across the different datasets, therefore this subscale was provided with transformation scores with the caveat that these scores should be reassessed.

Initially, the estimated linear interval measurement points were plotted against the raw scores. The sigmoidal curve (S-shape) indicated one point change in TFI-18 raw scores were not equivalent to equal changes in the interval locations across the scale (Figure 6.39). In other words, a ten point change in TFI raw score (clinical) from 90 to 100 equalled a change in interval value of 0.046, whilst a ten point change between TFI raw score at the higher end of the scale, i.e. from 150 to 160 implied a larger change in interval value of 0.217 (Figure 6.39).

This pattern in actual change can be seen throughout the subscales. For example, in the Relaxation subscale (Figure 6.40), the raw scores located at the asymptotes of the curve represented larger interval changes (a two point raw score change (26 to 28) equals 1.140 logit change) than those clustered in middle (i.e. 0.240 logit change for raw score change 12 to 14). The curve for the Cognition subscale was similar but slightly flatter indicating a more linear relationship between raw score change and interval change, especially in the centre of the scale, where one point change in raw score is fairly representative of the change seen in the interval scores (Figure 6.41). These plots provide the interval-level location values for each raw score which are presented in the transformation tables for each subscale (Tables 6.37 – 6.41) and six-factor-18 (Tables 6.42 – 6.45). Corresponding with the sigmoidal curve, the change in metric score was not linear to the change in raw scores. For every one-point raw score change the metric equivalent ranged from 0.08 to 7.67.

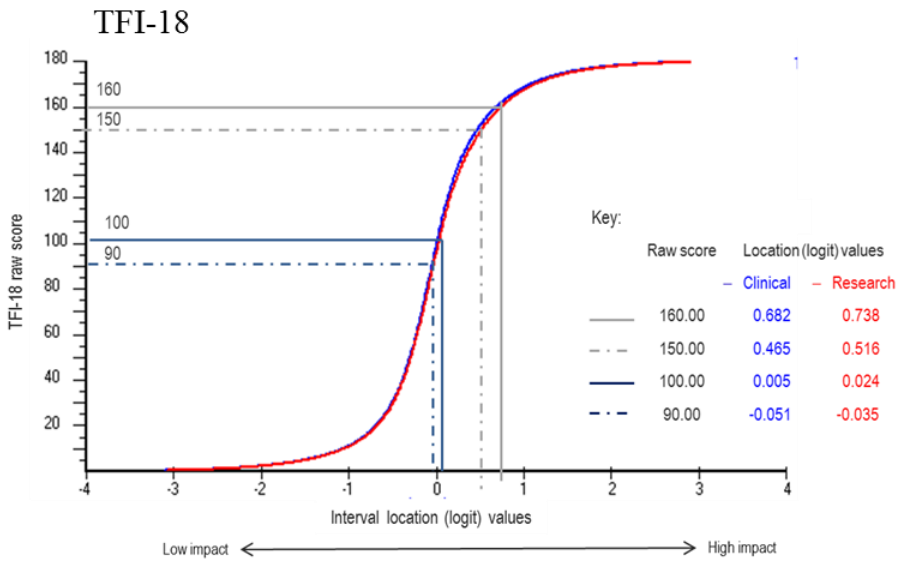

Figure 6.39. Plot of TFI-18 raw scores against interval location values (split for population differences).
The raw score changes for the clinical and research population do not equal an equivalent change in the interval logit values across the scale. For example, the change in interval logit value (-0.227) for a ten-point change in raw scores in the middle of the scale (blue lines) differs from the interval logit value (-0.056) for a ten-point change in raw scores at the higher end of the scale (grey lines).

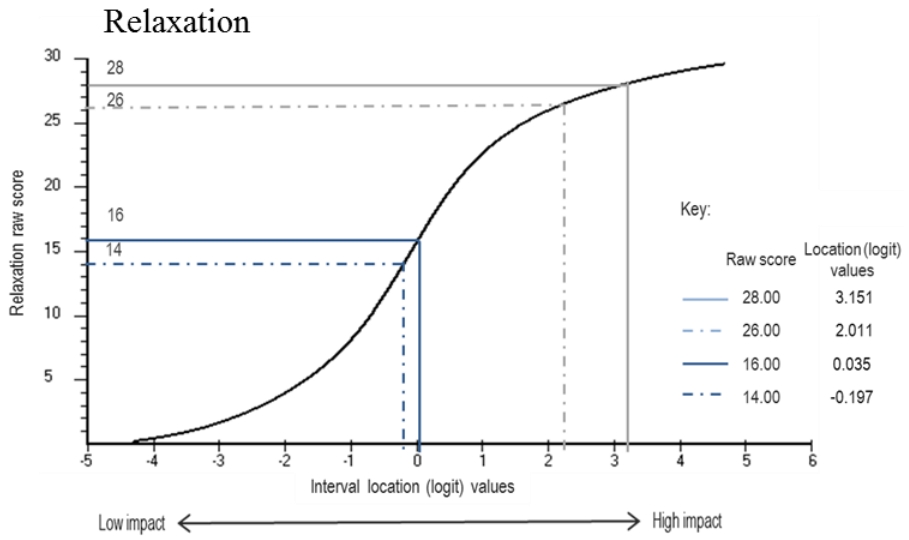

Figure 6.40. Plot of Relaxation subscale raw scores against interval location values.
The raw score changes for the Relaxation subscale do not equal an equivalent change in the interval logit values across the scale which are larger at the higher end of the scale. For example, a ten-point change in raw scores at the higher end of the scale equals -1.14 change in interval logit values (grey lines), whilst a ten-point change in raw scores in the middle of scale equals -0.232 change in interval logit values (blue lines).

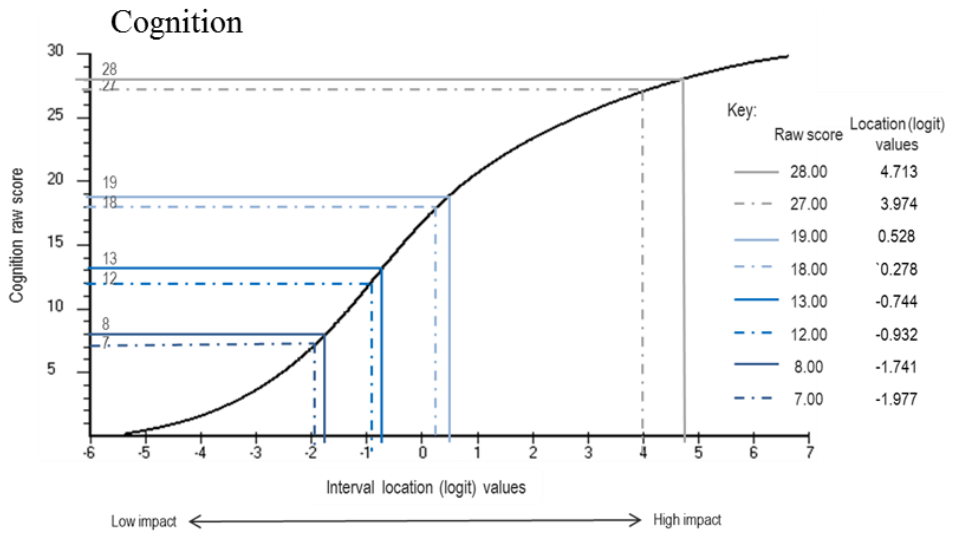

Figure 6.41. Plot of Cognition subscale raw scores against interval location values.
The raw score changes for the Cognition subscale are reasonably consistent with a change in do change in the interval logit values across the scale except for the extremes. For example, a ten-point change in raw scores at the higher end of the scale equals -0.739 change in interval logit values (grey lines), whilst a ten-point changes in raw scores lower in the scale equals -0.23 changes in interval logit values (blue lines).

For example, in the Sense of Control subscale (clinical), a one-point raw score change from 14 to 15 was equivalent to a difference in metric score of 0.33, whilst a one-point change from 28 to 29 equalled a difference in metric scores of 2.75 (Table 6.39).

In terms of the magnitude of change between one-point raw score changes, location mattered. For example, if a person experiencing extremely high impact of tinnitus had a reduction of three points (raw score) this would indicate a larger improvement than a three-point change for a person experiencing moderate tinnitus impact (located more centrally). For the TFI, these findings are particularly important.

Table 6.37. Transformation table for Intrusiveness subscale raw scores and corresponding converted metric scores (new) for use in clinical and research populations.

Intrusiveness							
Clinical				Research			
Raw	**Logit**	**New**	**Difference**	**Raw**	**Logit**	**New**	**Difference**
0	-5.132	**0**	–	**0**	-5.125	**0**	–
1	-3.805	**4.74**	4.74	**1**	-3.782	**4.54**	4.54
2	-2.927	**7.88**	3.14	**2**	-2.885	**7.58**	3.03
3	-2.345	**9.96**	2.08	**3**	-2.287	**9.60**	2.02
4	-1.914	**11.51**	1.54	**4**	-1.842	**11.11**	1.51
5	-1.575	**12.72**	1.21	**5**	-1.491	**12.29**	1.19
6	-1.3	**13.70**	0.98	**6**	-1.208	**13.25**	0.96
7	-1.071	**14.52**	0.82	**7**	-0.975	**14.04**	0.79
8	-0.878	**15.21**	0.69	**8**	-0.781	**14.70**	0.66
9	-0.712	**15.80**	0.59	**9**	-0.616	**15.25**	0.56
10	-0.568	**16.32**	0.51	**10**	-0.472	**15.74**	0.49
11	-0.44	**16.78**	0.46	**11**	-0.345	**16.17**	0.43
12	-0.323	**17.19**	0.42	**12**	-0.229	**16.56**	0.39
13	-0.215	**17.58**	0.39	**13**	-0.121	**16.93**	0.37
14	-0.112	**17.95**	0.37	**14**	-0.019	**17.27**	0.35
15	-0.012	**18.31**	0.36	**15**	0.079	**17.60**	0.33
16	0.086	**18.66**	0.35	**16**	0.174	**17.93**	0.32
17	0.184	**19.01**	0.35	**17**	0.264	**18.23**	0.30
18	0.284	**19.36**	0.36	**18**	0.352	**18.53**	0.30
19	0.385	**19.72**	0.36	**19**	0.438	**18.82**	0.29
20	0.491	**20.10**	0.38	**20**	0.523	**19.11**	0.29
21	0.6	**20.49**	0.39	**21**	0.612	**19.41**	0.30
22	0.716	**20.91**	0.41	**22**	0.708	**19.73**	0.32
23	0.841	**21.36**	0.45	**23**	0.813	**20.09**	0.36
24	0.978	**21.84**	0.49	**24**	0.941	**20.52**	0.43
25	1.136	**22.41**	0.56	**25**	1.103	**21.07**	0.55
26	1.326	**23.09**	0.68	**26**	1.329	**21.83**	0.76
27	1.569	**23.96**	0.87	**27**	1.651	**22.92**	1.09
28	1.905	**25.16**	1.20	**28**	2.105	**24.46**	1.54
29	2.432	**27.04**	1.88	**29**	2.769	**26.71**	2.25
30	3.259	**30**	2.96	**30**	3.743	**30**	3.29

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.38. Transformation table for Sense of control subscale raw scores and corresponding converted metric scores (new) for use in clinical and research populations.

Sense of control							
Clinical				Research			
Raw	**Logit**	**New**	**Difference**	**Raw**	**Logit**	**New**	**Difference**
0	-3.34	**0**	–	**0**	-3.233	**0.00**	–
1	-2.571	**3.59**	3.59	**1**	-2.464	**2.66**	2.66
2	-2.056	**5.99**	2.40	**2**	-1.95	**4.44**	1.78
3	-1.711	**7.61**	1.61	**3**	-1.607	**5.62**	1.19
4	-1.449	**8.83**	1.22	**4**	-1.348	**6.52**	0.90
5	-1.237	**9.82**	0.99	**5**	-1.14	**7.24**	0.72
6	-1.058	**10.65**	0.84	**6**	-0.966	**7.84**	0.60
7	-0.903	**11.38**	0.72	**7**	-0.817	**8.35**	0.52
8	-0.766	**12.02**	0.64	**8**	-0.687	**8.80**	0.45
9	-0.644	**12.59**	0.57	**9**	-0.573	**9.20**	0.39
10	-0.535	**13.10**	0.51	**10**	-0.472	**9.54**	0.35
11	-0.437	**13.55**	0.46	**11**	-0.381	**9.86**	0.31
12	-0.348	**13.97**	0.42	**12**	-0.299	**10.14**	0.28
13	-0.267	**14.35**	0.38	**13**	-0.223	**10.41**	0.26
14	-0.191	**14.70**	0.35	**14**	-0.152	**10.65**	0.25
15	-0.12	**15.03**	0.33	**15**	-0.084	**10.89**	0.24
16	-0.052	**15.35**	0.32	**16**	-0.019	**11.11**	0.22
17	0.015	**15.66**	0.31	**17**	0.046	**11.34**	0.22
18	0.083	**15.98**	0.32	**18**	0.113	**11.57**	0.23
19	0.152	**16.30**	0.32	**19**	0.182	**11.81**	0.24
20	0.224	**16.64**	0.34	**20**	0.254	**12.05**	0.25
21	0.302	**17.00**	0.36	**21**	0.333	**12.33**	0.27
22	0.387	**17.40**	0.40	**22**	0.422	**12.64**	0.31
23	0.486	**17.86**	0.46	**23**	0.525	**12.99**	0.36
24	0.604	**18.41**	0.55	**24**	0.653	**13.43**	0.44
25	0.749	**19.09**	0.68	**25**	0.818	**14.00**	0.57
26	0.938	**19.97**	0.88	**26**	1.05	**14.81**	0.80
27	1.196	**21.18**	1.20	**27**	1.411	**16.05**	1.25
28	1.569	**22.92**	1.74	**28**	2.043	**18.24**	2.18
29	2.158	**25.67**	2.75	**29**	3.227	**22.33**	4.09
30	3.086	**30**	4.33	**30**	5.445	**30**	7.67

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.39. Transformation table for Cognition and Sleep subscales raw scores and corresponding converted metric scores (new).

Cognition				Sleep			
Raw	**Logit**	**New**	**Difference**	**Raw**	**Logit**	**New**	**Difference**
0	-5.368	**0**	–	**0**	-3.881	**0**	–
1	-4.436	**2.31**	2.31	**1**	-3.003	**3.32**	3.32
2	-3.757	**3.99**	1.68	**2**	-2.388	**5.64**	2.32
3	-3.262	**5.22**	1.23	**3**	-1.958	**7.27**	1.62
4	-2.864	**6.21**	0.99	**4**	-1.625	**8.53**	1.26
5	-2.529	**7.04**	0.83	**5**	-1.358	**9.53**	1.01
6	-2.237	**7.76**	0.72	**6**	-1.138	**10.37**	0.83
7	-1.977	**8.41**	0.64	**7**	-0.956	**11.05**	0.69
8	-1.741	**8.99**	0.59	**8**	-0.801	**11.64**	0.59
9	-1.523	**9.53**	0.54	**9**	-0.666	**12.15**	0.51
10	-1.318	**10.04**	0.51	**10**	-0.545	**12.61**	0.46
11	-1.122	**10.53**	0.49	**11**	-0.435	**13.02**	0.42
12	-0.932	**11.00**	0.47	**12**	-0.334	**13.40**	0.38
13	-0.744	**11.47**	0.47	**13**	-0.236	**13.77**	0.37
14	-0.555	**11.93**	0.47	**14**	-0.141	**14.13**	0.36
15	-0.363	**12.41**	0.48	**15**	-0.047	**14.49**	0.36
16	-0.162	**12.91**	0.50	**16**	0.048	**14.85**	0.36
17	0.05	**13.43**	0.53	**17**	0.144	**15.21**	0.36
18	0.278	**14.00**	0.57	**18**	0.246	**15.60**	0.39
19	0.528	**14.62**	0.62	**19**	0.354	**16.00**	0.41
20	0.804	**15.30**	0.68	**20**	0.472	**16.45**	0.45
21	1.112	**16.07**	0.76	**21**	0.604	**16.95**	0.50
22	1.459	**16.93**	0.86	**22**	0.755	**17.52**	0.57
23	1.851	**17.90**	0.97	**23**	0.932	**18.19**	0.67
24	2.291	**18.99**	1.09	**24**	1.141	**18.98**	0.79
25	2.786	**20.22**	1.23	**25**	1.392	**19.93**	0.95
26	3.342	**21.60**	1.38	**26**	1.694	**21.07**	1.14
27	3.974	**23.16**	1.57	**27**	2.061	**22.45**	1.39
28	4.713	**25.00**	1.83	**28**	2.523	**24.20**	1.75
29	5.62	**27.25**	2.25	**29**	3.166	**26.63**	2.43
30	6.731	**30**	2.75	**30**	4.058	**30**	3.37

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.40. Transformation table for Relaxation and Emotional subscales raw scores and corresponding converted metric scores (new).

Relaxation				Emotional			
Raw	**Logit**	**New**	**Difference**	**Raw**	**Logit**	**New**	**Difference**
0	-4.315	**0**	–	**0**	-3.819	**0**	
1	-3.44	**2.77**	2.77	**1**	-2.879	**3.68**	3.68
2	-2.809	**4.77**	2.00	**2**	-2.239	**6.18**	2.50
3	-2.355	**6.21**	1.44	**3**	-1.803	**7.89**	1.71
4	-1.992	**7.36**	1.15	**4**	-1.483	**9.14**	1.25
5	-1.689	**8.32**	0.96	**5**	-1.238	**10.10**	0.96
6	-1.433	**9.13**	0.81	**6**	-1.045	**10.85**	0.76
7	-1.213	**9.83**	0.70	**7**	-0.887	**11.47**	0.62
8	-1.022	**10.44**	0.61	**8**	-0.752	**12.00**	0.53
9	-0.854	**10.97**	0.53	**9**	-0.634	**12.46**	0.46
10	-0.703	**11.45**	0.48	**10**	-0.528	**12.88**	0.41
11	-0.565	**11.88**	0.44	**11**	-0.43	**13.26**	0.38
12	-0.437	**12.29**	0.41	**12**	-0.338	**13.62**	0.36
13	-0.315	**12.68**	0.39	**13**	-0.248	**13.97**	0.35
14	-0.197	**13.05**	0.37	**14**	-0.16	**14.32**	0.34
15	-0.081	**13.42**	0.37	**15**	-0.071	**14.66**	0.35
16	0.035	**13.79**	0.37	**16**	0.018	**15.01**	0.35
17	0.152	**14.16**	0.37	**17**	0.111	**15.38**	0.36
18	0.274	**14.54**	0.39	**18**	0.21	**15.76**	0.39
19	0.403	**14.95**	0.41	**19**	0.317	**16.18**	0.42
20	0.542	**15.39**	0.44	**20**	0.436	**16.65**	0.47
21	0.696	**15.88**	0.49	**21**	0.569	**17.17**	0.52
22	0.871	**16.44**	0.55	**22**	0.721	**17.76**	0.59
23	1.075	**17.08**	0.65	**23**	0.898	**18.45**	0.69
24	1.321	**17.86**	0.78	**24**	1.103	**19.26**	0.80
25	1.626	**18.83**	0.97	**25**	1.344	**20.20**	0.94
26	2.011	**20.05**	1.22	**26**	1.626	**21.30**	1.10
27	2.506	**21.62**	1.57	**27**	1.966	**22.63**	1.33
28	3.151	**23.66**	2.04	**28**	2.395	**24.31**	1.68
29	4.009	**26.38**	2.72	**29**	2.999	**26.67**	2.36
30	5.151	**30**	3.62	**30**	3.849	**30**	3.33

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.41. Transformation table for QoL subscale raw scores and corresponding converted metric scores (new) for use in clinical and research populations

QoL-3							
Clinical				Research			
Raw	**Logit**	**New**	**Difference**	**Raw**	**Logit**	**New**	**Difference**
0	-2.956	**0.00**	–	**0**	-2.935	**0**	–
1	-2.231	**3.34**	3.34	**1**	-2.202	**3.32**	3.32
2	-1.782	**5.41**	2.07	**2**	-1.745	**5.38**	2.07
3	-1.503	**6.70**	1.29	**3**	-1.46	**6.67**	1.29
4	-1.305	**7.61**	0.91	**4**	-1.258	**7.59**	0.91
5	-1.153	**8.31**	0.70	**5**	-1.1	**8.30**	0.71
6	-1.026	**8.90**	0.59	**6**	-0.971	**8.89**	0.58
7	-0.918	**9.39**	0.50	**7**	-0.859	**9.39**	0.51
8	-0.822	**9.84**	0.44	**8**	-0.759	**9.85**	0.45
9	-0.734	**10.24**	0.41	**9**	-0.668	**10.26**	0.41
10	-0.652	**10.62**	0.38	**10**	-0.583	**10.64**	0.38
11	-0.573	**10.98**	0.36	**11**	-0.502	**11.01**	0.37
12	-0.496	**11.34**	0.35	**12**	-0.422	**11.37**	0.36
13	-0.42	**11.69**	0.35	**13**	-0.344	**11.72**	0.35
14	-0.344	**12.04**	0.35	**14**	-0.266	**12.08**	0.35
15	-0.265	**12.40**	0.36	**15**	-0.187	**12.43**	0.36
16	-0.183	**12.78**	0.38	**16**	-0.106	**12.80**	0.37
17	-0.097	**13.18**	0.40	**17**	-0.02	**13.19**	0.39
18	-0.005	**13.60**	0.42	**18**	0.071	**13.60**	0.41
19	0.096	**14.07**	0.47	**19**	0.17	**14.05**	0.45
20	0.209	**14.59**	0.52	**20**	0.279	**14.54**	0.49
21	0.335	**15.17**	0.58	**21**	0.402	**15.10**	0.56
22	0.479	**15.83**	0.66	**22**	0.543	**15.74**	0.64
23	0.645	**16.60**	0.77	**23**	0.706	**16.48**	0.74
24	0.838	**17.49**	0.89	**24**	0.899	**17.35**	0.87
25	1.064	**18.53**	1.04	**25**	1.126	**18.38**	1.03
26	1.333	**19.77**	1.24	**26**	1.4	**19.62**	1.24
27	1.662	**21.28**	1.52	**27**	1.735	**21.13**	1.52
28	2.083	**23.22**	1.94	**28**	2.17	**23.10**	1.97
29	2.687	**26.01**	2.78	**29**	2.795	**25.93**	2.83
30	3.553	**30.00**	3.99	**30**	3.695	**30**	4.07

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.42. Transformation table for the TFI-18 raw scores (range 0 – 99) and corresponding converted metric scores (new) for use in clinics.

Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff
0	-3.098	**0**		**20**	-0.692	**38.78**	0.42	**40**	-0.388	**43.68**	0.16	**60**	-0.228	**46.26**	0.11	**80**	-0.107	**48.21**	0.10
1	-2.467	**10.17**	10.17	**21**	-0.669	**39.15**	0.37	**41**	-0.378	**43.84**	0.16	**61**	-0.222	**46.36**	0.10	**81**	-0.101	**48.31**	0.10
2	-2.071	**16.55**	6.38	**22**	-0.647	**39.51**	0.35	**42**	-0.368	**44.00**	0.16	**62**	-0.215	**46.47**	0.11	**82**	-0.096	**48.39**	0.08
3	-1.821	**20.58**	4.03	**23**	-0.626	**39.85**	0.34	**43**	-0.359	**44.15**	0.15	**63**	-0.209	**46.57**	0.10	**83**	-0.090	**48.48**	0.10
4	-1.640	**23.50**	2.92	**24**	-0.606	**40.17**	0.32	**44**	-0.350	**44.29**	0.15	**64**	-0.202	**46.68**	0.11	**84**	-0.084	**48.58**	0.10
5	-1.498	**25.79**	2.29	**25**	-0.587	**40.47**	0.31	**45**	-0.341	**44.44**	0.15	**65**	-0.196	**46.78**	0.10	**85**	-0.079	**48.66**	0.08
6	-1.383	**27.64**	1.85	**26**	-0.570	**40.75**	0.27	**46**	-0.333	**44.57**	0.13	**66**	-0.190	**46.87**	0.10	**86**	-0.073	**48.76**	0.10
7	-1.286	**29.21**	1.56	**27**	-0.553	**41.02**	0.27	**47**	-0.324	**44.71**	0.15	**67**	-0.184	**46.97**	0.10	**87**	-0.068	**48.84**	0.08
8	-1.203	**30.54**	1.34	**28**	-0.537	**41.28**	0.26	**48**	-0.316	**44.84**	0.13	**68**	-0.177	**47.08**	0.11	**88**	-0.062	**48.94**	0.10
9	-1.131	**31.71**	1.16	**29**	-0.521	**41.54**	0.26	**49**	-0.308	**44.97**	0.13	**69**	-0.171	**47.18**	0.10	**89**	-0.056	**49.03**	0.10
10	-1.068	**32.72**	1.02	**30**	-0.506	**41.78**	0.24	**50**	-0.300	**45.10**	0.13	**70**	-0.165	**47.28**	0.10	**90**	-0.051	**49.11**	0.08
11	-1.012	**33.62**	0.90	**31**	-0.492	**42.01**	0.23	**51**	-0.292	**45.23**	0.13	**71**	-0.159	**47.37**	0.10	**91**	-0.045	**49.21**	0.10
12	-0.962	**34.43**	0.81	**32**	-0.479	**42.21**	0.21	**52**	-0.285	**45.34**	0.11	**72**	-0.153	**47.47**	0.10	**92**	-0.039	**49.31**	0.10
13	-0.917	**35.15**	0.73	**33**	-0.466	**42.42**	0.21	**53**	-0.277	**45.47**	0.13	**73**	-0.147	**47.57**	0.10	**93**	-0.034	**49.39**	0.08
14	-0.877	**35.80**	0.64	**34**	-0.454	**42.62**	0.19	**54**	-0.270	**45.58**	0.11	**74**	-0.141	**47.66**	0.10	**94**	-0.029	**49.47**	0.08
15	-0.839	**36.41**	0.61	**35**	-0.442	**42.81**	0.19	**55**	-0.263	**45.70**	0.11	**75**	-0.136	**47.74**	0.08	**95**	-0.023	**49.56**	0.10
16	-0.805	**36.96**	0.55	**36**	-0.430	**43.00**	0.19	**56**	-0.256	**45.81**	0.11	**76**	-0.130	**47.84**	0.10	**96**	-0.018	**49.65**	0.08
17	-0.774	**37.46**	0.50	**37**	-0.419	**43.18**	0.18	**57**	-0.249	**45.92**	0.11	**77**	-0.124	**47.94**	0.10	**97**	-0.012	**49.74**	0.10
18	-0.745	**37.93**	0.47	**38**	-0.408	**43.36**	0.18	**58**	-0.242	**46.03**	0.11	**78**	-0.118	**48.03**	0.10	**98**	-0.006	**49.84**	0.10
19	-0.718	**38.36**	0.44	**39**	-0.398	**43.52**	0.16	**59**	-0.235	**46.15**	0.11	**79**	-0.113	**48.11**	0.08	**99**	-0.001	**49.92**	0.08

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.43. Transformation table for the TFI-18 raw scores (range 100 – 180) and corresponding converted metric scores (new) for use in clinics.

Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff
100	0.005	**50.02**	0.10	**117**	0.110	**51.71**	0.11	**134**	0.251	**53.98**	0.16	**151**	0.483	**57.72**	0.29	**168**	0.965	**65.49**	0.73
101	0.011	**50.11**	0.10	**118**	0.117	**51.82**	0.11	**135**	0.261	**54.14**	0.16	**152**	0.502	**58.03**	0.31	**169**	1.014	**66.28**	0.79
102	0.016	**50.19**	0.08	**119**	0.124	**51.93**	0.11	**136**	0.272	**54.32**	0.18	**153**	0.521	**58.33**	0.31	**170**	1.069	**67.17**	0.89
103	0.022	**50.29**	0.10	**120**	0.131	**52.05**	0.11	**137**	0.283	**54.50**	0.18	**154**	0.541	**58.66**	0.32	**171**	1.130	**68.15**	0.98
104	0.028	**50.39**	0.10	**121**	0.138	**52.16**	0.11	**138**	0.294	**54.67**	0.18	**155**	0.562	**58.99**	0.34	**172**	1.198	**69.25**	1.10
105	0.033	**50.47**	0.08	**122**	0.145	**52.27**	0.11	**139**	0.306	**54.87**	0.19	**156**	0.583	**59.33**	0.34	**173**	1.277	**70.52**	1.27
106	0.040	**50.58**	0.11	**123**	0.153	**52.40**	0.13	**140**	0.318	**55.06**	0.19	**157**	0.606	**59.70**	0.37	**174**	1.368	**71.99**	1.47
107	0.046	**50.68**	0.10	**124**	0.161	**52.53**	0.13	**141**	0.331	**55.27**	0.21	**158**	0.630	**60.09**	0.39	**175**	1.478	**73.76**	1.77
108	0.052	**50.77**	0.10	**125**	0.169	**52.66**	0.13	**142**	0.344	**55.48**	0.21	**159**	0.655	**60.49**	0.40	**176**	1.613	**75.93**	2.18
109	0.058	**50.87**	0.10	**126**	0.177	**52.79**	0.13	**143**	0.357	**55.69**	0.21	**160**	0.682	**60.93**	0.44	**177**	1.790	**78.79**	2.85
110	0.064	**50.97**	0.10	**127**	0.185	**52.92**	0.13	**144**	0.371	**55.92**	0.23	**161**	0.710	**61.38**	0.45	**178**	2.039	**82.80**	4.01
111	0.070	**51.06**	0.10	**128**	0.194	**53.06**	0.15	**145**	0.386	**56.16**	0.24	**162**	0.739	**61.85**	0.47	**179**	2.445	**89.35**	6.54
112	0.077	**51.18**	0.11	**129**	0.203	**53.21**	0.15	**146**	0.401	**56.40**	0.24	**163**	0.771	**62.36**	0.52	**180**	3.106	**100**	10.65
113	0.083	**51.27**	0.10	**130**	0.212	**53.35**	0.15	**147**	0.416	**56.64**	0.24	**164**	0.804	**62.89**	0.53				
114	0.089	**51.37**	0.10	**131**	0.221	**53.50**	0.15	**148**	0.432	**56.90**	0.26	**165**	0.840	**63.48**	0.58				
115	0.096	**51.48**	0.11	**132**	0.231	**53.66**	0.16	**149**	0.448	**57.16**	0.26	**166**	0.878	**64.09**	0.61				
116	0.103	**51.60**	0.11	**133**	0.241	**53.82**	0.16	**150**	0.465	**57.43**	0.27	**167**	0.920	**64.76**	0.68				

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.44. Transformation table for the TFI-18 raw scores (range 0 – 99) and corresponding converted metric scores (new) for use in research

Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff
0	-3.075	**0**		**20**	-0.679	**38.39**	0.40	**40**	-0.376	**43.24**	0.16	**60**	-0.217	**45.79**	0.11	**80**	-0.093	**47.77**	0.10
1	-2.448	**10.04**	10.04	**21**	-0.655	**38.77**	0.38	**41**	-0.367	**43.38**	0.14	**61**	-0.211	**45.88**	0.10	**81**	-0.087	**47.87**	0.10
2	-2.053	**16.37**	6.33	**22**	-0.633	**39.12**	0.35	**42**	-0.357	**43.54**	0.16	**62**	-0.204	**45.99**	0.11	**82**	-0.082	**47.95**	0.08
3	-1.804	**20.36**	3.99	**23**	-0.612	**39.46**	0.34	**43**	-0.348	**43.69**	0.14	**63**	-0.197	**46.11**	0.11	**83**	-0.076	**48.05**	0.10
4	-1.623	**23.26**	2.90	**24**	-0.593	**39.76**	0.30	**44**	-0.339	**43.83**	0.14	**64**	-0.191	**46.20**	0.10	**84**	-0.070	**48.14**	0.10
5	-1.482	**25.52**	2.26	**25**	-0.574	**40.07**	0.30	**45**	-0.330	**43.98**	0.14	**65**	-0.185	**46.30**	0.10	**85**	-0.064	**48.24**	0.10
6	-1.367	**27.36**	1.84	**26**	-0.556	**40.36**	0.29	**46**	-0.322	**44.10**	0.13	**66**	-0.178	**46.41**	0.11	**86**	-0.058	**48.33**	0.10
7	-1.270	**28.92**	1.55	**27**	-0.540	**40.61**	0.26	**47**	-0.313	**44.25**	0.14	**67**	-0.172	**46.51**	0.10	**87**	-0.052	**48.43**	0.10
8	-1.187	**30.25**	1.33	**28**	-0.524	**40.87**	0.26	**48**	-0.305	**44.38**	0.13	**68**	-0.165	**46.62**	0.11	**88**	-0.046	**48.53**	0.10
9	-1.115	**31.40**	1.15	**29**	-0.509	**41.11**	0.24	**49**	-0.297	**44.50**	0.13	**69**	-0.159	**46.72**	0.10	**89**	-0.040	**48.62**	0.10
10	-1.052	**32.41**	1.01	**30**	-0.494	**41.35**	0.24	**50**	-0.289	**44.63**	0.13	**70**	-0.153	**46.81**	0.10	**90**	-0.035	**48.70**	0.08
11	-0.996	**33.31**	0.90	**31**	-0.480	**41.57**	0.22	**51**	-0.281	**44.76**	0.13	**71**	-0.147	**46.91**	0.10	**91**	-0.029	**48.80**	0.10
12	-0.947	**34.09**	0.79	**32**	-0.467	**41.78**	0.21	**52**	-0.274	**44.87**	0.11	**72**	-0.141	**47.00**	0.10	**92**	-0.023	**48.89**	0.10
13	-0.902	**34.81**	0.72	**33**	-0.454	**41.99**	0.21	**53**	-0.267	**44.99**	0.11	**73**	-0.135	**47.10**	0.10	**93**	-0.017	**48.99**	0.10
14	-0.861	**35.47**	0.66	**34**	-0.442	**42.18**	0.19	**54**	-0.259	**45.11**	0.13	**74**	-0.129	**47.20**	0.10	**94**	-0.011	**49.09**	0.10
15	-0.824	**36.06**	0.59	**35**	-0.430	**42.37**	0.19	**55**	-0.252	**45.23**	0.11	**75**	-0.123	**47.29**	0.10	**95**	-0.005	**49.18**	0.10
16	-0.791	**36.59**	0.53	**36**	-0.418	**42.57**	0.19	**56**	-0.245	**45.34**	0.11	**76**	-0.117	**47.39**	0.10	**96**	0.000	**49.26**	0.08
17	-0.759	**37.10**	0.51	**37**	-0.407	**42.74**	0.18	**57**	-0.238	**45.45**	0.11	**77**	-0.111	**47.48**	0.10	**97**	0.006	**49.36**	0.10
18	-0.730	**37.57**	0.46	**38**	-0.397	**42.90**	0.16	**58**	-0.231	**45.56**	0.11	**78**	-0.105	**47.58**	0.10	**98**	0.012	**49.46**	0.10
19	-0.704	**37.98**	0.42	**39**	-0.386	**43.08**	0.18	**59**	-0.224	**45.67**	0.11	**79**	-0.099	**47.68**	0.10	**99**	0.018	**49.55**	0.10

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

Table 6.45. Transformation table for the TFI-18 raw scores (range 100 – 180) and corresponding converted metric scores (new) for use in research.

Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff	Raw	Logit	New	Diff
100	0.024	**49.65**	0.10	**117**	0.137	**51.46**	0.13	**134**	0.289	**53.89**	0.18	**151**	0.534	**57.82**	0.29	**168**	1.024	**65.67**	0.74
101	0.030	**49.74**	0.10	**118**	0.144	**51.57**	0.11	**135**	0.300	**54.07**	0.18	**152**	0.553	**58.12**	0.30	**169**	1.073	**66.45**	0.79
102	0.036	**49.84**	0.10	**119**	0.152	**51.70**	0.13	**136**	0.312	**54.26**	0.19	**153**	0.573	**58.44**	0.32	**170**	1.128	**67.33**	0.88
103	0.043	**49.95**	0.11	**120**	0.159	**51.81**	0.11	**137**	0.324	**54.45**	0.19	**154**	0.593	**58.76**	0.32	**171**	1.189	**68.31**	0.98
104	0.049	**50.05**	0.10	**121**	0.167	**51.94**	0.13	**138**	0.336	**54.65**	0.19	**155**	0.615	**59.12**	0.35	**172**	1.257	**69.40**	1.09
105	0.055	**50.14**	0.10	**122**	0.175	**52.07**	0.13	**139**	0.348	**54.84**	0.19	**156**	0.637	**59.47**	0.35	**173**	1.335	**70.65**	1.25
106	0.061	**50.24**	0.10	**123**	0.184	**52.21**	0.14	**140**	0.361	**55.05**	0.21	**157**	0.661	**59.85**	0.38	**174**	1.427	**72.12**	1.47
107	0.068	**50.35**	0.11	**124**	0.192	**52.34**	0.13	**141**	0.375	**55.27**	0.22	**158**	0.685	**60.24**	0.38	**175**	1.537	**73.89**	1.76
108	0.074	**50.45**	0.10	**125**	0.201	**52.48**	0.14	**142**	0.388	**55.48**	0.21	**159**	0.711	**60.65**	0.42	**176**	1.672	**76.05**	2.16
109	0.081	**50.56**	0.11	**126**	0.210	**52.63**	0.14	**143**	0.403	**55.72**	0.24	**160**	0.738	**61.09**	0.43	**177**	1.850	**78.90**	2.85
110	0.087	**50.66**	0.10	**127**	0.219	**52.77**	0.14	**144**	0.417	**55.94**	0.22	**161**	0.766	**61.53**	0.45	**178**	2.100	**82.91**	4.01
111	0.094	**50.77**	0.11	**128**	0.228	**52.92**	0.14	**145**	0.432	**56.18**	0.24	**162**	0.796	**62.02**	0.48	**179**	2.507	**89.43**	6.52
112	0.101	**50.88**	0.11	**129**	0.238	**53.08**	0.16	**146**	0.448	**56.44**	0.26	**163**	0.828	**62.53**	0.51	**180**	3.167	**100**	10.57
113	0.108	**50.99**	0.11	**130**	0.247	**53.22**	0.14	**147**	0.464	**56.70**	0.26	**164**	0.862	**63.07**	0.54				
114	0.115	**51.11**	0.11	**131**	0.257	**53.38**	0.16	**148**	0.481	**56.97**	0.27	**165**	0.898	**63.65**	0.58				
115	0.122	**51.22**	0.11	**132**	0.268	**53.56**	0.18	**149**	0.498	**57.24**	0.27	**166**	0.937	**64.27**	0.62				
116	0.129	**51.33**	0.11	**133**	0.278	**53.72**	0.16	**150**	0.516	**57.53**	0.29	**167**	0.978	**64.93**	0.66				

Also presented are the interval measures the raw scores imply (logits) and the differences between converted metric score per raw score change (difference).

For comparative purposes, the TFI raw scores (baseline only) were transformed using the tables provided (Table 6.46). Mean scores (± SD) were calculated for the TFI-18 transformed scores, raw scores and total scores (Meikle et al. (2012) calculation) and the original distribution of the TFI-25 (Chapters 4 and 5) (Table 6.47). There were similarities in the mean and median scores between TFI-18 transformed and total and TFI-25 total in both the clinical and research population. But the SD around the mean scores in both populations tells a different story. The SD for TFI-18 transformed scores was considerably smaller (<8 points) than that seen in the TFI-18 and TFI-25 total scores (> 20 points) suggesting that range in scores was somewhat smaller than expected based on the original total scores.

In fact the difference in the distribution of the scores was clearly apparent from Figure 6.42 (clinical population) and Figure 6.43 (research population). Whilst the TFI-18 and TFI-25 total score distributions were flatter and range the whole scale, the distribution for TFI-18 transformed scores peaked in the centre of the scale. The raw scores located in the middle of the curve were located much closer together on the interval scale, therefore, it is perhaps unsurprising that the majority of transformed scores were now located centrally in the distribution.

In terms of subscales, the mean and median TFI-18 transformed scores differed from the total scores. The main reason was that the new transformed score were based on a 0 to 30 range, whereas the total scores calculated using instructions from Meikle et al. (2012), were based on a 0 to 100 range. The SD of the transformed scores indicated a reasonably wide distribution for the subscales that covered the full range in the scale. Accordingly, the new transformed scores suggested that the majority of the participants were experiencing moderate tinnitus impact.

Table 6.46. Descriptive statistics for the TFI-18 and subscales transformed scores, raw and total scores and the TFI-25 original raw and total scores.

	Person factor	n (missing)	Mean (SD)			Median			Range					
			Transform	Raw	Total	Transform	Raw	Total	Transform		Raw		Total	
TFI- 25	Clinical	255	—	158.88 (45.31)	52.72 (21.68)	—	167.0	52.0	—	—	62	216	7.6	100
	Research	247 (38)	—	101.63 (50.22)	40.64 (20.09)	—	96.0	38.40	—	—	10	232	4	93
	Total	540	—	115.49 (54.03)	46.7 (21.33)	—	112.0	45.20	—	—	7	250	7.6	100
TFI-18	Clinic	242 (13)	51.07 (7.45)	99.07 (39.58)	55.04 (21.99)	49.97	99.5	55.28	35	100	13	180	7	100
	Res	283 (2)	47.00 (5.28)	76.20 (38.06)	42.33 (21.14)	46.81	70	38.89	29	66	7	168	4	100
	Total	525 (15)		86.74 (40.37)	48.19 (22.43)		85	47.22			7	180	4	100
INTR	Clinic	251 (4)	20.06 (3.58)	18.69 (6.59)	62.28 (21.96)	19.72	19.00	63.33	10	30	0	30	10	100
	Res	283(2)	17.68 (2.75)	15.96 (6.39)	53.19 (21.31)	17.93	16.00	53.33	8	25	0	28	7	94
	Total	535 (5)		17.24 (6.62)	57.46 (22.07)		17.00	56.67			2	30	7	100
SOC	Clinic	251 (4)	17.18 (4.30)	19.36 (6.50)	64.54 (21.66)	16.64	20.00	66.67	0	30	0	30	0	100
	Res	285 (0)	11.13 (3.11)	16.07 (6.93)	53.56 (23.10)	11.34	17.00	56.67	0	30	0	30	0	100
	Total	536 (4)		17.61 (6.92)	58.70 (23.08)		18.00	60.00			0	30	0	100
QoL-3	Clinic	251 (4)	11.20 (6.61)	12.31 (9.12)	41.05 (30.39)	11.34	12.00	40.00	0	30	0	30	0	100
	Res	284 (1)	8.51 (5.67)	8.38 (7.98)	27.93 (26.58)	8.89	6.00	20.00	0	30	0	30	0	100
	Total	535 (5)		10.23 (8.75)	34.09 (29.15)		8.00	26.67			0	30	0	100
COG	Total	540 (0)	10.94 (6.07)	12.25 (8.23)	40.84 (27.45)	11.00	12.00	40.00	0	30	0	30	0	100
SLP	Total	538 (2)	13.84 (7.77)	14.18 (9.89)	47.26 (32.96)	14.13	14.00	46.67	0	30	0	30	0	100
REL	Total	539 (2)	15.49 (6.46)	17.79 (8.59)	59.29 (28.63)	15.39	20.00	66.67	0	30	0	30	0	100
EMO	Total	540(0)	12.67 (6.54)	11.71 (8.93)	39.05 (29.78)	12.88	10.00	33.33	0	30	0	30	0	100

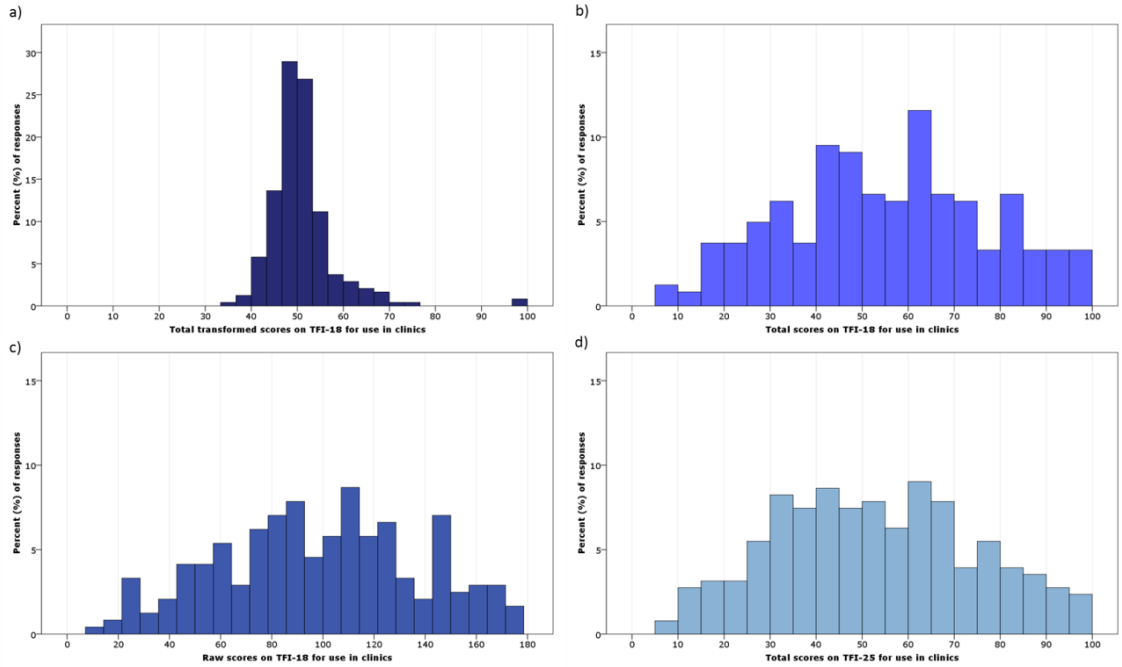

Figure 6.42. Distribution plots for TFI-18 transformed scores, raw and total scores and TFI-25 total scores in a clinical population

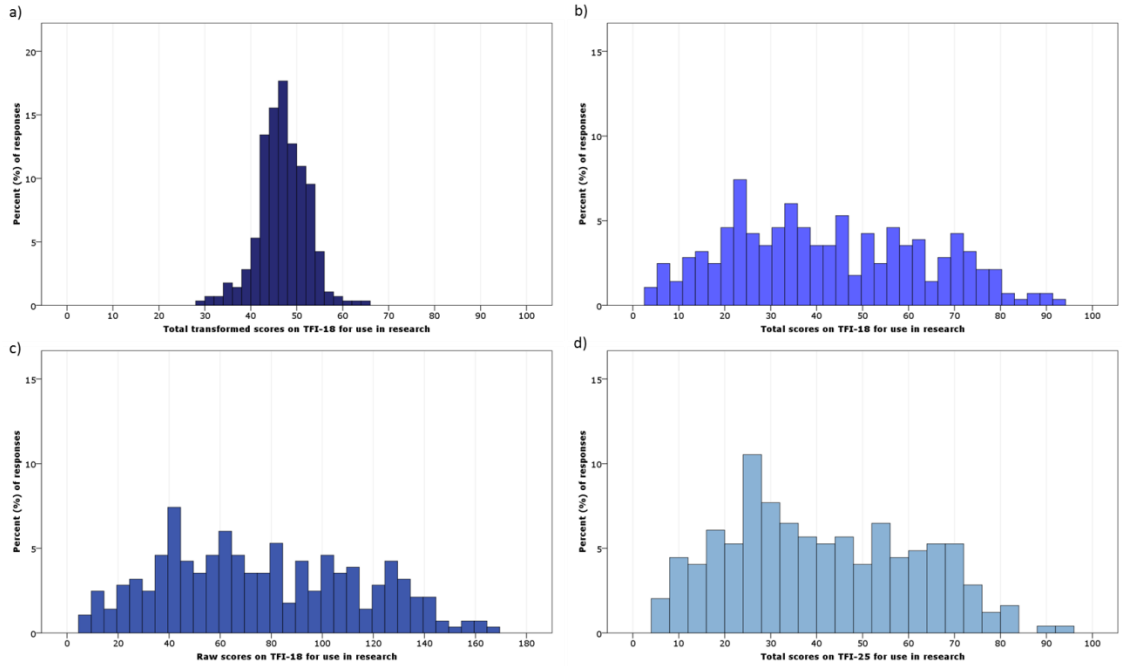

Figure 6.43. Distribution plots for TFI-18 transformed scores, raw and total scores and TFI-25 total scores in a research population

6.4. SUMMARY

This chapter reports the first application of Rasch analysis to a multi-item tinnitus questionnaire.

Targeting

Findings revealed differences between clinical and research populations in terms of their pattern of responses. In general, the clinical population reported higher levels of tinnitus impact than the research population. All the subscales, with the exception of the Auditory subscale, were reasonably well-targeted to the population. Category response thresholds were ordered for all items except for item 4 in the Sense of Control subscale. My conclusion is that the TFI is appropriately targeted to both populations, as TFI-18, and not the original 25-item TFI.

The TFI structure

The Auditory subscale was unrelated to the construct measured by the other subscales and is not targeted for the intended population. Although the hearing loss DIF analysis was under-powered, the results were interesting because the responses of those people without hearing loss did not conform to the model expectation. Therefore, individuals are answering the items in relation to their hearing rather than their tinnitus. The Sleep subscale was unrelated to the construct measured by the other subscales. My conclusion is that the six-factor structure in general conforms to the Rasch model expectations. I recommend that the Auditory and Sleep subscales should not be retained in the second-order construct. This would allow for the maximum information, measurement precision, and accuracy.

Six of the subscales in their original form were confirmed as standalone subscales that appropriately target the population of interest and reliably measure the

underlying construct they purport to measure, although the Sense of control thresholds had to remain disordered. The QoL subscale had one item (QOL22) that appeared to be the consistent source of the deviations in the model. The results showed a clear trend in which removing QOL22 improved the precision and accuracy of the scale and reduced the differences in item functioning, in age groups in particular. Based on this analysis, it is recommended that QOL22 is removed from the TFI to create QoL-3. Items on the Sleep subscale showed some evidence of overlapping content indicative of item redundancy. My recommendation for the Sleep subscale is to not remove the redundant item because a two-item subscale would not be reliable, but to re-evaluate the wording of these problematic items.

Transformation

The global TFI-18 scores and all subscales, except for the Auditory subscale, were provided with transformation scores. My recommendation is that clinicians and researchers remain mindful of the impact of the transformations in terms the distribution of the scores across the population. Transformation to an interval scale reduces the SD of the overall distribution which in turn reduces the responsiveness of the TFI to small changes.

CHAPTER 7. GENERAL DISCUSSION

The TFI was developed through an international effort over eight years to be used as both a tinnitus severity diagnostic tool and to be a sensitive measure of treatment-related change, addressing eight separate subscales of tinnitus-related functional impact. Although some psychometric evaluations were conducted on earlier versions of the TFI (exploratory factor analysis, convergent validity, and test-retest reliability), the final 25 item version was never subjected to any formal validation. This PhD project was the first formal validation of the TFI, providing empirical evidence on the validity and reliability of the TFI and establishes whether the TFI is a measure of tinnitus appropriate for use in clinical practice and research in the UK. I evaluated the TFI using approaches from two different measurement theories: classical test theory and modern test theory. Classical psychometrics properties of validity, reliability, responsiveness and interpretability of the TFI were examined using data from clinical and research populations, separately. Rasch measurement theory was applied to the data from both populations.

My specific objectives were to evaluate whether the TFI is

(i) valid and covers a broad range of problems and symptoms associated with tinnitus-related distress, in particular to verify the eight-factor TFI structure proposed by Meikle et al. (2012).

(ii) a reliable measure of the functional impact of tinnitus that distinguishes between individuals.

(iii) responsive, able to measure small changes over time above measurement error

(iv) interpretable, such that scores and the change in scores are clinically meaningful.

7.1. KEY FINDINGS

7.1.1. **The proposed structure to the TFI was not confirmed**.

The TFI is a composite measure designed to be comprehensive in covering all of the symptoms and impacts that were deemed important by expert clinicians. However, the eight-factor structure proposed by Meikle et al. (2012) was not confirmed at any point over the three studies here. Only four (Sense of control, Cognition, Relaxation, Emotional) of the eight subscales in their original form (including the items associated with the factors) were reliably associated with each other and the underlying construct. A number of discrepancies were observed for the remaining four factors. The most important of these was the evidence that the Auditory factor was not measuring the functional impact of tinnitus and as a consequence was not targeting the intended population. It was identified as the source of large deviations in the model expectation and fit across all three studies. Evidence showed the Auditory factor to be unrelated to the underlying construct. Although, Meikle et al. (2012) suggested that the Auditory subscale might be measuring a unique construct of tinnitus impact, I expected that it should still be measuring some aspect of tinnitus that relates to aspects of tinnitus measured by the other factors, especially considering the TFI was designed to be provide an overall measure of tinnitus impact, but this was not the case. Rasch analysis clearly showed that the Auditory factor is not appropriate for measuring tinnitus impact. The Auditory factor seems linked to hearing related problems per se rather than tinnitus-related hearing problems and therefore was not targeting the intended population or concept. As such it has poor content validity and is not appropriate for use in any form in the UK. Tinnitus is often co-morbid with hearing loss (Hoare et al., 2014) and as a consequence it is particularly hard to reliably capture hearing problems in relation to

tinnitus. People tend to attribute their hearing difficulties to tinnitus such that it is difficult to disentangle whether the degree of hearing loss is the source of their tinnitus problems or whether the severity of their tinnitus is the source of their hearing loss (Ratnayake et al., 2009). Given the nature of questions in tinnitus questionnaires, they can be susceptible to measuring hearing difficulties. Questions asked about communication or concentration problems for example, can elicit responses that are predisposed towards hearing difficulties rather than tinnitus (Kuk et al., 1990; Newman et al., 1998; Ratnayake et al., 2009).

There are numerous reasons for wanting to disambiguate handicap related to hearing from that related to tinnitus. The Tinnitus and Hearing Survey (THS; (Henry et al., 2014)) was developed with this specific purpose in mind and includes two subscales. The first asks about tinnitus problems that are unrelated to hearing difficulties and the second asking about "commonly experienced hearing problems that would not be confounded by tinnitus complaints" (p.68). The scale is designed to be used as an initial screening to identify the extent of hearing and tinnitus complaints before making clinical decisions. Interestingly, the item content in the hearing subscale is similar to that of the Auditory subscale items in the TFI. For example, the THS item 4 asks "*I couldn't understand what was being said in group conversations*" whilst the TFI Auditory subscale item 15 asks "how much has your tinnitus interfered with *your ability to follow conservations in a group or meeting*?". This is perhaps because the THS was developed by the same researchers who developed the TFI.

From classic psychometrics it was apparent the Sleep factor did not contribute as much to the underlying construct as the other six factors but it was above the critical loading levels and maintained within these studies. However,

within the confines of Rasch, this deviation from the rest of the factors was far more apparent. The Sleep factor undermined the TFI structure and was not measuring the overall construct. Therefore, from a psychometric perspective the Sleep factor should be removed from the second-order construct and from the overall score. However, sleep is a significant problem domain for many people with tinnitus (Andersson et al., 2005; Crönlein et al., 2007; Miguel et al., 2014). Clinically, it is something that is important to measure. Having said this, the Sleep subscale alone was found to be reasonably reliable, with high measurement precision, targeted to the intended population and was able to reliably differentiate different levels of tinnitus impact, so should be informative to clinicians and researchers. Given evidence of overlap in item content however, the actual scores should be interpreted with caution and the item content should re-evaluated.

At first, the QoL factor as a whole appeared to be reliably associated with the underlying construct. But one item was repeatedly identified as a source of the deviations from model fit. QOL22 was not measuring the same underlying construct as the other items on the QoL factor. Evidence from all three studies indicate that the content of the question is not clear, it was associated with aspects of cognition and hearing, and given the floor effects observed, it is not relevant to the target population. Maintaining this item within the questionnaire would reduce reliability and as shown in the Rasch analysis (Chapter 6), reduce measurement precision and accuracy of the QoL subscale and the overall score. Therefore, from a psychometrics perspective, the QoL factor should be reduced to a 3-item scale within the TFI structure.

The Intrusiveness factor is more complex. Rasch analysis indicated that it was measuring a slightly different aspect of tinnitus impact than the other factors.

Inspection of the item content revealed that the intrusiveness subscale clearly measures pure tinnitus associated with the perceptual characteristics of tinnitus, such as the perceived intensity or magnitude of tinnitus. For example, INTR1 asks "*How strong or loud was your tinnitus?".* The descriptors of the items are different to those used for the other items in the TFI. They focus on awareness and annoyance with tinnitus, rather than any type of reaction to or consequence of tinnitus. Consistent with this, two of the Intrusiveness items were utilised in the "new" 3-item Tinnitus Magnitude Index (TMI) (Schmidt et al., 2014). Therefore, the Intrusiveness factor may be construed as a different construct from that being measured by the other factors. However, classic psychometrics contradict this. The Intrusiveness factor showed acceptable fit with the underlying construct in both studies using classic psychometrics, and had strong correlations with all the other factors (with the exception of the Auditory factor). Schmidt et al. (2014) claimed that discriminant validity was demonstrated with the TFI subscales. But according to my guidelines (Table 2.2), the correlations between the subscales and the TMI (>0.6) indicated inadequate discriminant validity. My data showed similar patterns of findings. Intrusiveness (magnitude) of tinnitus would seem to contribute less to the overall construct than the other factors, but is still an important domain to include as it improves measurement precision and is integral for identifying the first signs of tinnitus impacting on daily lives. The Intrusiveness item thresholds provide the first point of measurement of tinnitus impact above the five other subscales and without these, patients with milder levels of tinnitus impact would be misclassified as not having a problem. The content of the Intrusiveness subscale therefore may capture the aspects of the early natural history of tinnitus before it to impacting on emotional well-being and cognition. The Intrusiveness factor improved the measurement

precision and maximised information retained in the second-order six-factor model and therefore, although, perhaps measuring a slightly different construct, it was maintained within the structure.

Finally, within Rasch analysis, seven of the eight subscales were found to be reasonably well-targeted, reliable scales that could be used separately to provide measures of individual constructs of tinnitus impact. Specifically, this analysis suggests the Intrusiveness, Sense of Control, Cognition, Relaxation, QoL-3 and Emotional subscales can be reliably used for initial assessments in research and clinical practice.

Using evidence from classic psychometrics and Rasch analysis, a six-factor TFI structure with 18-items was identified as providing the best explanation for the data and a stable underlying construct. Further research is warranted to modify the original version of the TFI to exclude all the Auditory items and QOL22 item from the questionnaire format and then re-evaluate some of the key properties in a new UK population. In particular, minimal change should be re-evaluated.

The TFI has been publically available since 2012, and is already widely use within UK audiology (Hoare et al., 2015), and as a baseline assessment and outcome measure in numerous research studies (Shekhawat et al., 2014; Michiels et al., 2014; Wilson et al., 2015; Krings et al., 2015; Henry et al., 2015). Until the modifications described above can be implemented across the UK, clinics are likely to continue to use the TFI in its current form; the implications of doing this needs to be considered.

In clinical practice guidelines, the overall scores on tinnitus questionnaires are recommended to quantify the severity of tinnitus and inform the clinical pathway. For example, high scores determine who needs referrals to mental health

professionals (Tunkel et al., 2014). If the TFI global score is calculated, clinicians should be mindful that the Auditory and Sleep subscales undermine and dilute the overall score, of the functional impact of tinnitus. Otherwise, on a practical level, the internal structure of the TFI is appropriate to be used as a guide for clinicians to think about the clinical approach and to encourage patients to talk about the range of problems they are experiencing. The subscales scores can be particularly useful for identifying specific problems with tinnitus and for understanding the degree of problems faced by the patient.

Within research, there is greater flexibility to use different elements of questionnaires that are more applicable to the intervention being evaluating. Therefore, UK researchers can use the TFI for baseline assessment to assess the different problems being reported and can use the most appropriate subscale to evaluate the intervention. The global TFI score can be calculated using the 18 items (six-factors) alone. However, given that the minimal important change score is not be applicable to this structure, it is viable to use the 25-item questionnaire until a minimal change score is identified for TFI-18, but researchers should plan to do a secondary analysis using 6-factor to overcome confounding results.

7.1.2. The TFI reliably distinguishes individual participants

Despite a majority of effort during TFI development being focussed on optimising the evaluative properties of the questionnaire, its discriminative properties were not compromised. It has excellent ability to distinguish individual differences in the degree of tinnitus impact across both populations using classic psychometrics. Within Rasch, the reliability estimates and person item distributions indicated that the reduced 6 factor TFI-18 structure was well targeted for the population and able to discriminate according to the functional status of the sample. Similarly the TFI

subscales, with the exception of the Auditory subscale, were also able to distinguish between individuals. Therefore, the TFI can be used as a diagnostic tool.

Although Rasch indicated that all the subscales, except the auditory subscale, were reasonably well-targeted, there were problems with measurement points in the extremes. However, for the most part, the skew was towards the lower end of the scale. Therefore distinguishing between patients with milder levels of tinnitus would be less accurate. But this comes down to a matter of practicality and priority. The measurement points in each subscale were in general located centrally and higher on the scale, providing the majority of information about these higher levels of tinnitus impact. Practically speaking, these higher levels are more important for clinicians to identify and distinguish patients and make informed decision about the management plan and care-pathway. For example, patients with high scores indicating severe emotional impact of tinnitus would complete additional questionnaires measuring specific psychological problems.

7.1.3. The TFI is responsive to change but suffers from issues of variability and becomes less responsive over time

The ability of a questionnaire to detect changes is the single most important factor for clinical trials and clinical audits in the assessment of outcome. Primary outcomes provide the means to determine what interventions are effective and hence to influence therapeutic management strategies. Maximising responsiveness to change was a key element in the development of the TFI. Items were specifically chosen because they describe attributes that were likely to undergo changes following intervention (Meikle et al., 2012). The TFI was shown to have the ability to detect changes in scores above measurement error in both populations, but was perhaps not as responsive as the authors intended.

Substantial floor effects were observed in both populations indicating that for some items, and most of the subscales, the ability to detect improvements was somewhat limited, with the QoL subscale being least responsive. Having said this, the TFI was responsive and reliably detected changes above measurement error, but the ability to detect change and improvements did not extend past 6 months. At 9 months, the difference scores and magnitude of change for improvements was notably lower than previous months, and as a consequence participants experiencing improvements were harder to discriminate from those who remained unchanged. The magnitude of change is dependent on the time-frame in which the scores are compared. When assessing change previous studies using Patient-Reported Outcomes (PROs) for different health conditions have used shorter time frames (baseline to 3 months) (Crosby et al., 2004; Yost et al., 2011) or combined time-points (Jaeschke et al., 1989). Cella et al. (2002) measured change on the Functional Assessment of Cancer Therapy Amelia and Fatigue Scales at 6 and 9 months and although unobserved by the authors, the change scores at 9 months were smaller than those at 6 months for the improved group. Therefore, when identifying important changes consideration should be placed on the time-intervals which would vary depending on the condition being measured.

Large variation in changes in scores for participants who experienced no change in their tinnitus or were expected to be stable (short time interval between test-retest in research) indicated that a "true change" in TFI needs to be large. Small changes including the MIC could be hidden by this noise in the measurement. The TFI subscales were particularly vulnerable to this effect, especially over the long periods of time. The LoA estimates from the two classic psychometric studies indicated that the variability in scores was similar across both populations (LoA

Clinic = 20.2; research = 22.1). In fact, similar estimates for LoA were identified in both the THI (19.5) (Newman et al., 1998) and THQ (20.0) (Newman et al., 1995) indicating that this is the typical variability that would be expected in a tinnitus population.

A fundamental problem in measuring tinnitus is the natural history of the condition over days, weeks and months. Tinnitus patients adjust their perception of their tinnitus, through natural coping mechanisms or re-evaluation of internal standards of health-status. Hesser et al. (2011) observed that waiting-list control groups showed significant improvements over 12 weeks. Furthermore, within this thesis, small effects, considerably larger than expected, were reported for the TFI global and subscale scores in 'no change' groups at all time points. This further highlights the need to identify the degree of error and variability within a measure (measurement precision). Researchers, in particular, need to be aware of this natural variability when making judgements on the significance of treatment effects and when claiming that a scale is responsive to change. In a recent article, James Henry, one of the leading authors for the TFI development, claimed that the large effect sizes (ES) observed between the waiting-list and immediate care groups lend credibility to claims that the TFI can detect treatment-related changes (Henry et al. 2015). The magnitude of ES is associated with the intervention and the responsiveness of the questionnaire, not measurement precision (Mokkink et al., 2012). Therefore, although the ES observed for the TFI here do show changes, it would be near impossible to identify whether the magnitude of effect is a reflection of the intervention and the responsiveness of the TFI or the noise in the measurement.

Transformation scores would also indicate that the TFI was limited in detecting small changes. The distribution of the interval scores in Rasch indicated that large changes in TFI raw scores were required for the interval scores to show any noticeable changes, especially for participants with moderate levels of tinnitus which were all tightly grouped centrally on the scale. On the other hand, for scores in the extremes only small changes in raw scores were required to observe a large change in interval scores. This highlights that baseline scores matter when it comes to maximising the responsiveness. This is the first tinnitus questionnaire to be subjected to Rasch and it provides interesting results in that the short range seen in these transformation scores could potentially explain why we do not always see large treatment effects in clinical trials for tinnitus using traditional questionnaire raw ordinal scores. If transformed scores are used, then it would provide critical information on the magnitude of the changes that are observed and in turn of the efficacy of the intervention. Therefore, from a clinical perspective, these transformation scores could improve interpretations of intervention, but would not provide information for clinical decision making and would be rather time consuming to use.

7.1.4. Minimal important change on the TFI is above measurement error but is affected by baseline.

Identification of a minimal change that is clinically meaningful is fundamental in health research and clinical trials. MIC scores are used to estimate the required sample sizes for research and as a way of monitoring individual patient progress in clinical practice. However, interpreting changes in scores is difficult as there are multiple properties of measurement relevant to the “true” change and not error within the measurement. MIC estimates were identified for the clinical population using an

integrated approach of anchor-based ratings of perceived change and ROC optimal values and the statistical properties of the scores. A change in global TFI scores of at least -18 points was identified as an important (improved) change above measurement error and the majority of variability. To account for all variability, a change in global TFI score of at least 23 points is required for individual assessment. The magnitude of this change is considerably larger than the 13-point difference proposed by Meikle et al. (2012) as a clinically meaningful change. This discrepancy was larger than expected even though Meikle et al. (2012) also used an anchor-based rating of change for moderate-to-much improved to identify their MIC. However, Meikle et al. (2012) did not report estimates of measurement error which can lead to unrealistically low cut-offs that sit within the measurement error (Crosby et al., 2004; de Vet et al., 2006b). Therefore, the estimates provide here are more reliable and we can be confident that the change identified is a realistic reflection of true change in score.

The MIC values identified using the magnitude of change scores between ratings of change groups were dependent on the baseline questionnaire score. Participants with higher baseline values required larger changes on the TFI than participants with lower baseline values to consider it an important change. Logically, participants with high baseline scores have more opportunity to register greater improvements than the other baseline scores. The MIC estimate of -18 is larger than the estimate associated with the big to very problems at baseline and therefore would account for these and be above the measurement error. However, the SDC and SEM were calculated without consideration to baseline. Crosby et al (2002) have shown that the SEM can be dependent on baseline values. Therefore the error for the baseline grades may differ from that identified here and the different MIC estimates

for baseline could possibly be used. The Edward-Nunnally method adjusts for regression to the mean at baseline and classifies the change scores based on the confidence intervals around the actual pre-test scores, but unfortunately was not possible here as it requires a large sample sizes and normative means (Edwards et al., 1978; Speer, 1992; Crosby et al., 2004). Future studies should examine the measurement error associated with the different baseline values.

Although the sample was reasonably representative of the clinical population and consistent over the different time points, it is yet to be determined whether the same MIC estimates would be identified in a different population. Due the retrospective nature of the validation study conducted in the research population, I was unable to include global ratings of perceived change and therefore minimal important change could not be evaluated in this population, but a SDC score was identified which was comparable to the score identified in the clinical population. The MIC therefore should be re-examined, especially in the modified version of the TFI. Given that MIC values are not considered constant and are essential for effective assessment of interventions in clinical trials it is recommended that researchers incorporate a global ratings of perceived change question into all clinical trials as this would provide addition evidentially support for the MIC and can be used to identify the degree of variability in participants who perceived no change in their tinnitus.

7.1.5. The TFI can be used to grade tinnitus severity

A new grading system was developed here to quantify the TFI scores into distinct grades of tinnitus impact. Although, these grades were identified using the statistical properties of the scores, patient experience of the degree to which their tinnitus was a problem and the current "gold standard" THI grading system were used to identify

the grades of impact. The TFI is the first tinnitus questionnaire to include patient experience into the grading system; the THI grading system for instance was based on statistical properties and clinician experience (Newman et al., 1998; McCombe et al., 2001). The inclusion of patient experience means that there is more confidence in the reliability of these grades to reflect patient experiences. Rather than providing vague descriptors to the different grades of tinnitus impact that would be solely based on the research team's experience, the response category descriptors used in the global rating of perceived problem question were adopted to provide qualitative meaning to the grades. Future research should focus on identifying what these descriptors mean to patients and professionals, for example identifying whether "small problem" means the same as mild tinnitus as used for the THI. A qualitative approach should be used to add specific clinical meaning to the grades including detailed explanations of what each grade means for clinical practice, similar to the descriptions provided with the THI grading system.

7.2. LIMITATIONS AND STRENGTHS

The inclusion of the global rating of perceived change compared to 3 months ago did not necessarily contribute to understanding change in score, except possibly when assessing participants whose tinnitus remained unchanged and potentially over complicated matters. Participants seemed unable to conceptualise changes across the time frame for 3 months. Without a prominent memory to recall, such as the first clinical appointment, the global rating of change every 3 months becomes more susceptible to recall bias.

The wording of the global rating of perceived change question could also have affected discrimination. Since the main aim of the thesis was to identify

whether the global TFI detects changes over time, the global change question was aimed at identifying change in overall tinnitus rather than the specific aspects of tinnitus measured by the subscales. An improvement (or worsening) in overall tinnitus does not necessarily imply that all the subscales will change in the same direction, and therefore there may be a small amount of unaccounted for variability that could be attributed to mis-classification. Future studies interested in identifying changes within the subscales should use global ratings of change that reflect the subscale being measured.

A particular strength of the clinical study was that the dropout was less than expected. Despite small dropout, the sample sizes within the global ratings of change groups in particular, the numbers of participants within the "worse" groups were insufficient to provide meaningful conclusions. As such, I was unable to provide a minimal important change score for worsening. Similarly, to identify an optimal cut-off value in relation to baseline values, the sample size within the improved group would have to be much larger.

7.3. CONCLUSIONS AND FUTURE DIRECTIONS

As hearing research continues to develop a foothold in the field of measurement, it is important to conduct these in-depth evaluations of current questionnaires to confirm that the response structure is working as expected. The TFI was found to be a reasonably comprehensive measure of tinnitus impact that was reliable as a diagnostic tool, and could detect changes in tinnitus impact, although it was slightly less responsive than originally proposed, the measurement error was comparable to the other tinnitus questionnaires currently available. However, it was not as comprehensive as the author envisioned. Three of the original factors

measured alternative constructs, two of which were unrelated to overall construct. A suggested future iteration of the TFI is to modify it excluding all three Auditory items and QOL22 item and evaluating some of the key properties in a new population, in terms of the validity of the 6-factor structure following the suggested change to the format, minimal change, and a grading system.

This highlights a fundamental problem with multidimensional composite questionnaires which is the implicit assumption that all the different symptom domains, associated with tinnitus, in this case, can be combined to measure an overarching construct of the functional impact of tinnitus. Essentially, these symptom domains can be measuring very different concepts of tinnitus. Although they are related to each other in content (they are all measuring an element of tinnitus), it does not mean that combined they would equally contribute to the same measurement of tinnitus.

One approach in other disciplines is to develop one instrument per domain. For tinnitus this would mean eight separate questionnaires using the TFI domains as a guiding principle, for example. Whether these are the 'right' domains is still an empirical question. Perhaps what we need to do is move away from the concept of overall tinnitus distress, and begin to focus on the symptoms of tinnitus in which the intervention is intended to alleviate rather than 'noise' which might obscure a treatment effect (Kirshner & Guyatt, 1985). One possible action is to remove any domains considered irrelevant to an individual at pre-intervention so they are not measured at post-treatment assessment (Tyler et al., 2014). For instance one of the seven validated TFI subscales could be used for follow-up assessment if appropriate for the intervention. The key for effective assessments is to allow patients to express exactly what problems they are having as a result of their tinnitus and measure the

problem with the appropriate measurement tool that is more domain-specific as described above.

In fact, a question that has repeatedly occurred to me is whether we are just asking too much of questionnaires? In a condition as complex as tinnitus is it too much to ask for a questionnaire to have both diagnostic properties that can deal with the multiple symptoms that impact on everyday lives as well as contend with the possible changes that occur over time. That said, the domains themselves or the importance of them are not even well understood. The problem domains identified by Tyler and Baker in 1983 and Sanchez and Stephens (1997;2000) are the foundation for the majority of knowledge on problems domains. Although informative at the time, these were small studies that provided a limited amount of knowledge about the importance of these domains and the degree of problems experienced at that time; they may not be as relevant for today's tinnitus patients. Furthermore, given that in a recent study where we examined patient responses to a question asking why tinnitus is a problem, 20 domains were identified as problems (including some not previously described, i.e. fear, awareness and loss of peace) experienced by patient due to their tinnitus (Watts et al., in prep), future work on measurement of tinnitus should first consider what really are the important problems that should be measured either for diagnosis or outcome.

The COST Action TINNET Outcome Measure Working Group 5, of which I am a member of the management committee, (http://tinnet.tinnitusresearch.net/) is looking to establish a standard core domain set that should routinely be used in clinical trials of tinnitus (http://www.comet-initiative.org/studies/details/703). The first stage involved two systematic reviews to create a list of problem domains that are currently reported (Hall et al., 2015a). The first review identified current-reported

outcome domains in tinnitus clinical trials for tinnitus interventions (Hall et al., 2015b, Hall et al., in review.). The second review aims to identify the problems that have been reported by patients and their significant others (Haider et al., in press). For this review, we are examining and extracting information from articles that detail patient-reported complaints/problems (i.e. Tyler & Baker, 1983; Sanchez & Stephens, 1997, 2000; Watts et al., in prep)) using narrative synthesis. This involves tabulating and summarising the text-based data and descriptive labels used for the patient-reported problems, clustering the related problems together (concepts) to form a list (Haider et al., in press). Alongside this, I am managing a project identifying problems domain from tinnitus questionnaire items. For this, we have identified tinnitus questionnaire items that are currently available and using a thematic analysis approach, we are grouping the items together solely based on the item wording, ignoring any existing subscales/domains previously reported by the developers. The intention from this is to ensure a comprehensive list of problem domains has been identified. This list will inform a modified international Delphi survey that I will be managing and overseeing, the aim of which is to establish consensus across key stakeholders in tinnitus (i.e. patients, clinicians, and researchers) on the important domains to measure in clinical effectiveness trials. Once we have identified "what" should be measured (key domains), future work will establish "how" to measure these domains and identify appropriate measurement tools (for roadmap see Hall et al., 2015a).

CHAPTER 8. APPENDICES

8.1. APPENDIX A

ARTICLE IN PRESS

Hearing Research xxx (2015) 1–16

Contents lists available at ScienceDirect

Hearing Research

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

Research paper

Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population

Kathryn Fackrell [a, b, *], Deborah A. Hall [a, b], Johanna G. Barry [c, d], Derek J. Hoare [a, b]

[a] NIHR Nottingham Hearing Biomedical Research Unit, Nottingham, NG1 5DU, UK
[b] Otology and Hearing Group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, NG7 2RD, UK
[c] MRC Institute of Hearing Research, University Park, Nottingham, NG7 2RD, UK
[d] Nottingham University Hospitals NHS Trust, Nottingham, NG5 1PB, UK

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form
17 September 2015
Accepted 22 September 2015
Available online xxx

Keywords:
Outcome instruments
Reproducibility
Reliability
Confirmatory Factor Analysis
Convergent validity
Discriminant validity
Responsiveness

ABSTRACT

Objectives: Questionnaires are essential for measuring tinnitus severity and intervention-related change but there is no standard instrument used routinely in research settings. Most tinnitus questionnaires are optimised for measuring severity but not change. However, the Tinnitus Functional Index (TFI) claims to be optimised for both. It has not however been fully validated for research purposes. Here we evaluate the relevant psychometric properties of the TFI, specifically the questionnaire factor structure, reproducibility, validity and responsiveness guided by quality criteria for the measurement properties of health-related questionnaires.

Methods: The study involved a retrospective analysis of data collected for 294 members of the general public who participated in a randomised controlled trial of a novel tinnitus device (ClinicalTrials.gov Identifier: NCT01541969). Participants completed up to eight commonly used assessment questionnaires including the TFI, Tinnitus Handicap Inventory (THI), Tinnitus Handicap Questionnaire (THQ), a Visual Analogue Scale of loudness (VAS-Loudness), Percentage Annoyance question, the Beck's Depression Inventory (BDI), Beck's Anxiety Inventory (BAI), and the World Health Organisation Quality of Life-Bref (WHOQOL-BREF). A series of analyses assessed the study objectives. Forty four participants completed the TFI at a second visit (within 7–21 days and before receiving any intervention) providing data for reproducibility assessments.

Results: The 8-factor structure was not fully confirmed for this general (non-clinical) population. Whilst it was acceptable standalone subscale, the 'auditory' factor showed poor loading with the higher order factor 'functional impact of tinnitus'. Reproducibility assessments for the overall TFI indicate high internal consistency ($\alpha = 0.80$) and extremely high reliability (ICC: 0.91), whilst agreement was borderline acceptable (93%). Construct validity was demonstrated by high correlations between scores on the TFI and THI ($r = 0.82$) and THQ ($r = 0.82$), moderate correlations with VAS-L ($r = 0.46$), PR-A ($r = 0.58$), BDI ($r = 0.57$), BAI ($r = 0.39$) and WHOQOL ($r = -0.48$). Floor effects were observed for more than 50% of the items. A smallest detectable change score of 22.4 is proposed for the TFI global score.

Conclusion: Even though the proposed 8-factor structure was not fully confirmed for this population, the TFI appears to cover multiple symptom domains, and to measure the construct of tinnitus with an excellent reliability in distinguishing between patients. While the TFI may discriminate those whose tinnitus is not a problem, floor effects in many items means it is less appropriate as a measure of change in this subgroup. Further investigation is needed to determine whether these effects are relevant in other populations.

1. Introduction

The experience of tinnitus can involve much more than the 'phantom' sensation of sound, it can also impact on daily functioning, causing insomnia, difficulties in listening and

* Corresponding author. NIHR Nottingham Hearing Biomedical Research Unit, Ropewalk House, 113 The Ropewalk, Nottingham, NG1 5DU, UK.
E-mail address: msxklf@nottingham.ac.uk (K. Fackrell).

http://dx.doi.org/10.1016/j.heares.2015.09.009

concentrating, impaired symptom-specific quality of life, and poor psychological well-being (Tyler and Baker, 1983; Robinson et al., 2003; Stevens et al., 2007; Langguth et al., 2011; Nondahl et al., 2011; Pierce et al., 2012). But quantifying the severity of this impact, or how this severity changes as a result of time or intervention, is difficult. Psychoacoustic estimates of tinnitus loudness may partially explain some of the variance attributed to the functional impact or perceived annoyance/intrusiveness of tinnitus (Dauman and Tyler, 1992; Andersson, 2003). But ratings of loudness, annoyance or awareness of tinnitus made using a Visual Analogue Scale (VAS), recommended by some as standalone measures of tinnitus severity, do not correlate strongly with either psychoacoustic or multi-item questionnaire measures of tinnitus (Adamchic et al., 2012). Given that tinnitus is a multi-dimensional symptom, researchers typically rely on multi-attribute self-report questionnaires to quantify tinnitus severity and to assess intervention-related change over time.

Numerous questionnaire measures of tinnitus have been developed to date (for reviews see Fackrell et al., 2014; Meikle et al., 2008; Newman and Sandridge, 2004), and recommended for clinical use (Department of Health, 2009; Langguth et al., 2007; Tunkel et al., 2014). For tinnitus research, the international standards proposed by Landgrebe et al. (2012) calls for the routine use of the Tinnitus Handicap Inventory (THI; Newman et al., 1996), and that researchers define a validated tinnitus questionnaire as at least one of the primary outcome measures. Questionnaires are widely used in tinnitus research to either characterise the participant population (e.g. to aid comparison across different studies; Boyen et al., 2013; Melcher et al., 2013), to measure the effects of experimental intervention (e.g. Hoare et al., 2014a; Song et al., 2013), or to explore correlations between self-reported tinnitus severity and biological observations (e.g. Song et al., 2013; Szczepek et al., 2014). The approaches taken to validate tinnitus questionnaires to date have sometimes limited their utility (Meikle et al., 2008; Fackrell et al., 2014). For example, although the interpretability of the Tinnitus Handicap Questionnaire (THQ; Kuk et al., 1990) has been examined this has not led to defined categories of severity (Newman et al., 1995). The THI was developed specifically as a diagnostic tool with defined categories of severity (Newman et al., 1996; McCombe et al., 2001), and has been criticised for lacking sensitivity to change (Meikle et al., 2007). The Tinnitus Functional Index (TFI; Meikle et al., 2012) was developed to provide (i) comprehensive coverage of the broad range of symptoms associated with tinnitus severity, (ii) reliable measurement of tinnitus severity that distinguishes between individuals from those whose tinnitus is 'not a problem' to those whose tinnitus is a 'very big problem', and (iii) responsive measurement of change in tinnitus severity. It may therefore have a number of applications in research studies. The questionnaire underwent a systematic process of development to distil an initial item pool of 175 items through two prototypes (prototype 1 had 43 items, prototype 2 had 30 items) to arrive at a final questionnaire containing 25 items each mapping onto one of eight functional subscales (see Meikle et al., 2012 for details). The subscales (factors) were defined through Exploratory Factor Analysis and named as (i) Intrusiveness (items 1–3), (ii) Sense of control (items 4–6), (iii) Cognition (items 7–9), (iv) Sleep (items 10–12), (v) Auditory (items 13–15, (vi) Relaxation (items 16–18), (vii) Quality of life (items 19–22), and (viii) Emotional distress (items 23–25). The development pathway included a process of exploratory factor analysis, assessment of content validity, test-retest reliability, internal consistency, and convergent and discriminant validity. Development of the TFI used data collected from clinics in the USA, primarily specialist tinnitus clinics (42% of participants) and Veterans' Affairs (VA) hospitals (58% of participants). Those recruited from the VA sites tended to be male and experienced a range of co-morbidities, such as Post-Traumatic Stress Disorder (PTSD). Validation of the TFI is understood therefore relative to this mixed clinical population. It cannot be assumed that the questionnaire will show the same properties when administered to a different population. In fact the final 25-item version of the TFI has never been directly subjected to formal psychometric evaluation. The only assessment of validity and reliability was based on analysis of a subset of data collected for the 30-item prototype 2 of the questionnaire, and confirmatory factor analysis was not conducted (Meikle et al., 2012).

Here we examine the properties of the TFI for a general sample of UK adults experiencing tinnitus who presented themselves to take part in a clinical trial guided by quality criteria for the measurement properties of health-related questionnaires (Terwee et al., 2007; see also Fackrell et al., 2014). Specifically, the psychometric validation reported here focuses on assessing (a) the reliability of the 8-factor TFI structure reported by Meikle et al. (2012), i.e. verifying item identification with each factor and the underlying construct using Confirmatory Factor Analysis, and (b) the ability of the TFI to reliably measure tinnitus severity, distinguishing between individual differences in tinnitus-related distress, and responsively measure change in tinnitus severity.

2. Materials and methods

2.1. Participants and procedure

This was a retrospective analysis of data collected during a two-centre clinical trial conducted at the National Institute for Health Research Nottingham Hearing Biomedical Research Unit (BRU) and the University College London Ear Institute (RESET2, ClinicalTrials.gov ID:NCT01541969; Hoare et al., 2013). For that trial, participants were recruited via adverts placed on the website of the Nottingham Hearing BRU or in local hearing clinics, or to publicity in the national media. Participants reflected a mix of those who had previously attended clinical appointments for their tinnitus, and those who had never sought medical help for their tinnitus. Although none of the participants were receiving any clinical interventions for their tinnitus at the time of assessment, all participants were strongly motivated to seek a specific treatment by volunteering for this clinical trial in which a novel sound therapy for tinnitus was prescribed for a period of 36 weeks of daily use. The intake assessment for eligibility onto the trial provided data for the psychometric validation analysis. Assessment included Percentage Annoyance question, a VAS of tinnitus loudness, the TFI, THI, THQ, the Beck Anxiety Inventory (BAI; Beck and Steer, 1990) and Beck's Depression Inventory (BDI-II; Beck et al., 1996), and the World Health Organisation Quality of Life (WHOQOL-BREF; The WHOQOL group, 1998). In the clinical trial, 391 were assessed for eligibility but 291 were excluded from the trial at either telephone screening or eligibility appointments because they did not meet the inclusion criteria (stated in ClinicalTrials.gov Identifier: NCT01541969, but not relevant for the present study), or withdrew. Hence, 100 participants were allocated to one of the study arms and received treatment. The data contributing to the present study comprised 294 individuals (212 male, 82 female), with an average age of 52.8 years (range: 18 to 82) and tinnitus duration of 9.0 years (range: 4 months to 50 years). We have TFI data at the initial assessment from 285 individuals (two were excluded due of missing data) and of those, 12% reported tinnitus as not a problem (range: 0–17), 27% reported tinnitus as a small problem (range: 18–31), 31% as a moderate problem (range: 32–53), 24% as a big problem and 5% as a very big problem (range: 73–100). This distribution was comparable to that reported by some of the clinical centres participating in the original development of the TFI Protocol 1 (Meikle et al.,

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

2012), with individuals spanning all categories of severity.

Data were collected in accordance with the permissions granted by the Nottingham 1 NHS Research Ethics Committee and the Sponsor (Nottingham University Hospitals NHS Trust) as part of the protocol described in Hoare et al. (2013).

2.2. Missing data

Not all participants completed all assessments and only complete questionnaire datasets were analysed. Listwise deletion is considered an effective approach to deal with missing data when only a small amount of data (<5%) is assessed as 'missing completely at random' (MCAR) (Schafer and Graham, 2002) and avoids problems associated with over-estimating factors (Tabachnick and Fidell, 2013). This was the case here.

Only those data with fully completed TFI scores on all 25 items were used for analysis of the TFI factor structure, internal consistency and responsiveness (floor and ceiling effects) and so after list-wise deletion the effective sample size was 283. TFI was not completed in 9 cases, and in 2 cases one item was missing (defined as MCAR). Furthermore, analyses of convergent and divergent validity were calculated after list-wise deletion of missing items on the different comparison assessments and so the effective sample size was 247. Forty-seven individuals did not complete all the necessary assessments.

The clinical trial required a second TFI dataset for the 100 enrolled participants, which we used here to assess reproducibility using test-retest reliability and agreement analysis to determine how close repeated measures were to each other. The clinical trial protocol did not specify a required time interval between first and second administration of the TFI, but based on the previous validation (Meikle et al., 2012) and recommendations (Terwee et al., 2007) we conservatively limited reproducibility analyses to data from a subset of 44 participants who completed the TFI twice within an average of 15 days (SD = 7).

2.3. Measures

2.3.1. Percentage annoyance

As part of the Tinnitus Case History Questionnaire (TCHQ), participants were asked to state any number between 0 and 100 that represents the percentage of time awake they were annoyed by their tinnitus.

2.3.2. Visual analogue scale of loudness (VAS-Loudness)

As part of the 'Tinnitus Tester' computerised test (Roberts et al., 2006, 2008) participants rated the loudness of their tinnitus on a Borg CR100 (VAS) scale (Borg and Borg, 2001). Participants mark the loudness of their tinnitus at any point along the numerical scale, but word descriptors, "extremely weak," "moderate," "strong," "very strong," and "extremely strong", are utilised as anchor points which predisposes subjects to interpret it as an ordinal scale. Hoare et al. (2014a) recently reported that test-retest agreement was very high for this element of the Tinnitus Tester.

2.3.3. Tinnitus Functional Index (TFI)

Participants scored each item of the 25 items according to how they felt over the past week. Each item is scored on an 11-point scale, with descriptors at either end of the scale. The procedure for scoring the TFI followed the instructions provided by Meikle et al. (2012). The sum of all scores is divided by 2.5 to give a global score out of 100. Higher scores reflect greater impact on daily functioning. Subscale scores are calculated as the sum of the relevant three or four items.

2.3.4. Tinnitus Handicap Inventory (THI)

The THI measures the effects of tinnitus on everyday function (Newman et al., 1996, 1998; Baguley et al., 2000). Each of the 25 items is rated on a categorical 3-point scale (yes/no/sometimes). The mean global score reflects the sum of all responses with a maximum score of 100 indicating the greatest impact on everyday function. Although subscales of the THI have been proposed (Newman et al., 1996) subsequent analyses have demonstrated that the THI items load predominantly onto a single factor (Baguley and Andersson, 2003; Kennedy et al., 2004) and so for the purposes of analysis here this questionnaire is considered unidimensional.

2.3.5. Tinnitus Handicap Questionnaire (THQ)

The THQ measures overall handicap associated with tinnitus, in particular the effects of tinnitus on hearing and communication, physical health, social and emotional status (Kuk et al., 1990; Robinson et al., 2003). For each of the 27 items, participants indicate their agreement with each item, by assigning a number between 0 (strongly disagree) to 100 (strongly agree). Again, the mean global score reflects the sum of all responses, averaged to give a global score out of 100. Higher scores indicate higher levels of tinnitus handicap. Kuk et al. (1990) recommended a two-factor structure for the THQ, with items relating to factor 1 (physical, emotional and social effects) and factor 2 (hearing and communication ability) considered reliable enough to be used as independent subscales.

2.3.6. Beck's Depression Inventory – II (BDI-II)

The BDI-II provides a measure of depressive symptomatology, in particular mood and physical effects (Beck et al., 1996; Dozois et al., 1998; Segal et al., 2008). Participants select statements characterising how they have felt over the previous two weeks, and each of the 21 items is rated on a categorical scale (0–3 points). Responses are summed to form a global score out of 63, with higher scores indicating higher levels of depressive symptomatology.

2.3.7. Beck's Anxiety Inventory (BAI)

The BAI is a measure of the clinical anxiety (Beck and Steer, 1990; Steer et al., 1993). It lists 21 common symptoms associated with clinical anxiety, such as nervousness and fear of losing control. Participants rate how much they were bothered by each symptom over the previous week on a categorical scale (0–3 points) and, as for the BDI, responses are summed to give a global score out of 63 (higher scores indicate greater anxiety).

2.3.8. World Health Organisation Quality of Life-BREF (WHOQOL-BREF)

The WHOQOL-BREF provides a broad reliable measurement of perceived quality of life embedded in a cultural, social and environmental context (The WHOQOL Group, 1998; Skevington et al., 2004). The WHOQOL-BREF produces four domain scores (physical health, psychological, social relationships and environment) and also includes one facet on overall quality of life and general health ("How would you rate your quality of life?"). This item has 5 response options being (1) "very poor"; (2) "poor"; (3) "neither poor nor good"; (4) "good"; and (5) "very good". The score is transformed onto a 100 point scale, using the WHOQOL-BREF conversion method (The WHOQOL Group, 1998).

2.4. Data screening

Non-normality of data can have adverse effects on the statistics conducted here, in particular the Confirmatory Factor Analysis, so as a first step the TFI data were screened for outliers, linearity and multicollinearity. There was no evidence of univariate outliers in

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

the boxplots and histograms. However Mahalanobis distance statistic indicated that there were nine multivariate outliers with the greatest distance from the rest of the data points (Mahalanobis d-squared: 90.72 to 59.15, $p \leq 0.0001$). Kurtosis and skewness did not exceed the recommended cut-off points (for kurtosis = 2.00; skewness = 7.00; Curran et al., 1996). However, Mardia's normalised coefficient estimate was 37, exceeding the recommended value of <5 (Bentler, 2006; Mardia, 1971). This indicates some non-normality in the distribution of the data, requiring control.

The data for all questionnaires (global and subscales scores) met the assumptions relating to multicollinearity and linearity; analysis of tolerance indices and Variance Inflation Factor (VIF) all met the cut-off points of >0.10 and <10, respectively (Menard, 2002; Myers, 2000).

2.5. Statistical analysis

2.5.1. Confirmation of the 8-factor structure of the TFI

Confirmatory Factor Analysis was performed in Mplus 7 (Muthén and Muthén, 2012). It was conducted on TFI data to test how the variables observed for our research population fit the 8-factor structure devised by Meikle et al. (2012, Fig. 1). The initial 8-factor model was defined by four properties: (i) The latent constructs: eight first-order factors corresponding to the TFI subscales and one second-order factor corresponding to the global measure "Functional impact of tinnitus"; (ii) Each item (observed variable) loaded only on to its designated first factor without any cross-loading (i.e. constrained zero loadings on the other factors); (iii) Residual variance (error/uniqueness terms) associated with each variable (25 items, 8 first-order factors) were assumed to be uncorrelated and random (constrained to zero); (iv) The variance of the second order factor was fixed at 1 as it was assumed that the first-order factors are completely explained by the relationship to the second-order factor.

Data were treated as continuous rather than categorical, as the response scale was large (0–10 points) (Muthén and Muthén, 2012). To adjust for non-normality in the data and to ensure robust standard errors for parameter estimates and goodness of fit indices, the model was estimated using maximum likelihood parameter estimation adjusted with Satorra–Bentler scaled Chi-square (S–B χ^2; Satorra and Bentler, 1994; Bentler, 2006; Hu and Bentler, 1999). Caution is needed when interpreting the significance of S–B χ^2 as it is strongly influenced by sample size and variability in the data (Hu and Bentler, 1998; Brown, 2006).

Factor intercorrelations were performed to indicate the degree to which the factors are related to one another and are potentially overlapping in content. These are examined first before the model included the second-order factor. A degree of overlap is expected between factors such as these as they are purported to be measuring the same underlying construct (functional impact of tinnitus). However, highly correlated factors (>0.85) were taken to indicate that they are not measuring distinct constructs from each other (poor discriminant validity). Weakly correlated factors (<0.30) were taken to indicate that they were highly distinct from each other, and potentially measuring an alternative underlying construct (Brown and Moore, 2012; Brown, 2006).

The criterion for goodness of fit was determined using absolute fit indices S–B χ^2 (Satorra and Bentler, 1994) and Standardised Root Mean Square Residual (SRMR; Hu and Bentler, 1998; Bentler, 2006) to access the discrepancies between the implied correlations (predicted by the model) and observed covariances. The S–B χ^2 is assessed relative to the degrees of freedom, and this estimate has a critical ratio cut-off of $\leq$2.0. Alongside this, a large S–B χ^2 with $p < 0.05$ and SRMR that exceeds 0.07 (ideally less than 0.06) were

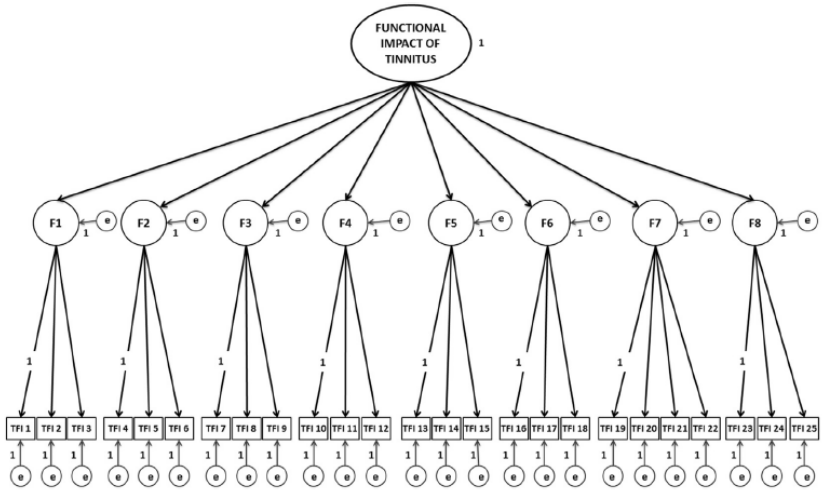

Fig. 1. Illustrative diagram of the theoretical 8-factor structure of the TFI assessed by Confirmatory Factor Analysis. The model represents the proposed relationships between the observed variables (items i.e. TFI 1), the first order factors (F1 to F8) and the second-order factor (Functional impact of tinnitus). The model has the following properties: (i) Second-order latent construct: "Functional impact of tinnitus" with the variance fixed at 1. Here, the unidirectional black arrows (⟶) from the second-order factor to the first order factors represent the direct effects of the second-order latent construct onto those factors; (ii) Eight first-order latent constructs: F1: Intrusiveness; F2: Sense of control; F3: Cognition, F4: Sleep; F5: Auditory; F6: Relaxation; F7: Quality of life; F8: Emotional with the variance explained by second-order factor. In this case, the unidirectional black arrows (⟶) represent the direct effects of the first-order constructs onto the observed measures; (iii) 25 observed variables: TFI item 1 to TFI item 25 with the variance of the first item on each factor fixed at 1, and all items have zero loadings on the other factors; (iv) The unidirectional grey arrows (⟶) represent the residual variance (e) associated with each variable (25 items; 8 first-order factors), which were constrained to zero. F1 = Intrusiveness; F2 = Sense of control; F3 = Cognition, F4 = Sleep; F5 = Auditory; F6 = Relaxation; F7 = Quality of life; F8 = Emotional; e = residual variance (error and uniqueness terms).

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

taken to together indicate poor fit and that the model should be rejected. Approximation fit indices were also used. Tucker–Lewis Index (TLI; Tucker and Lewis, 1973) and Comparative Fit Index (CFI; Bentler, 1990) assessed the model fit to baseline. Values for both should exceed 0.90, and preferably exceed 0.95 (Hu and Bentler, 1999). Root Mean Square Error of Approximation (RMSEA; Steiger and Lind, 1980) measured the discrepancy per degree of freedom. Ideally, RMSEA should be less than 0.05, but values up to 0.08 are considered reasonable when the SRMR value is ≤0.06. RMSEA confidence intervals should also fall within the desired criteria (Brown, 2006; Hu and Bentler, 1999, 1998).

Standardised parameter estimates (β; factor loadings) provided an indication of the magnitude and pattern of the relationship between the latent constructs and the observed variables. Our assumption was that the item–factor relationship is entirely explained by the influence of the latent construct. Factor loadings exceeding 0.7 are were taken to mean that the majority of the shared variance was explained by the latent construct. Loadings below 0.4 are associated with measurement error or poor explained variance and were taken to indicate a potential source of poor model fit (Brown and Moore, 2012; Floyd and Widaman, 1995).

The Modification Index (MI) and Expected Parameter Change (EPC) were used to identify any misspecification in the parameters of the model. Large modification indices exceeding 3.84 were taken to indicate that if a parameter was freely estimated, rather than fixed or constrained, the overall model fit would significantly improve (Brown and Moore, 2012). The EPC value was used to provide an approximation of the direction or magnitude by the parameter would change in subsequent analysis. Together, they were used to decide, where supported by conceptual foundations, which parameter should be adjusted (Brown and Moore, 2012; MacCallum et al., 1992).

2.5.2. Psychometric properties of the TFI

All statistical analyses were performed in SPSS (v.21.0). Reproducibility, validity and responsiveness of the TFI were assessed.

2.5.2.1. Reproducibility of the TFI.

Reproducibility was assessed using three methods; internal consistency, reliability and agreement across testing sessions. *Internal consistency* assesses the extent to which each item in a factor measures the same underlying construct. Cronbach's alpha (α) estimates between 0.7 and 0.9 were taken to indicate acceptable internal consistency (Peterson, 1994; Terwee et al., 2007). *Reliability* compares the degree to which people with tinnitus can be distinguished from each other across two testing sessions, despite measurement error, i.e. the similarity in the variability in scores. Reliability was assessed using Intra-Class Correlations (ICC), with scores >0.70 indicating high reliability (Terwee et al., 2007). *Agreement* relates to the measurement error, and the degree to which each individual's scores collected on two separate time points are in agreement with each other. Agreement was assessed using two methods identifying the limits of agreement (Bland and Altman, 1986) and the Smallest Detectable Change. The limits of agreement method (Bland and Altman, 1986) assumes the mean change score (difference) between repeated measures is zero, and that 95% of mean changes should be within ±1.96 standard deviations of the zero difference score (Bland and Altman, 1986). Limits of agreement were calculated as

$$\textit{limits of agreement} = \overline{d} \pm 1.96 \times SD_{diff}$$

where $\overline{d}$ represents the mean difference in scores between the two administrations, the ±1.96 represents two standard deviations, whilst the SD_{diff} represents the mean difference in standard deviation. This allows for examination of the mean change scores in relation to the change in standard deviation, taking into account the random measurement error. 95% agreement was taken as an indication of high test-retest agreement.

Smallest Detectable Change reflects the extent of expected measurement error and was derived from the Standard Error of Measurement (SEM) between repeated measures Smallest Detectable Change (de Vet et al., 2011; Terwee et al., 2007; de Vet et al., 2006a), where: $SEM_{consistency} = SD_{diff}/\sqrt{2}$

$$\text{Smallest Detectable Change} = 1.96 \times \sqrt{2} \times SEM$$

The Smallest Detectable Change score should be comparable to the limits of agreement score to be deemed an acceptable score.

2.5.2.2. Validity of the TFI.

Convergent and discriminant validity (the extent to which a questionnaire is measuring the construct it purports to measure; Haynes et al., 1995; Streiner and Norman, 2008) was assessed as Pearson bivariate correlations. To evaluate *convergent validity*, the global TFI scores were compared to THQ and THI global scores in the same population. The TFI was assumed to measure a similar construct and so it was predicted to have high convergent validity with both questionnaires (correlation > 0.60). We predict that the TFI global score will show a weak convergent validity (correlation < 0.6) with VAS-Loudness and Percentage Annoyance, in the same way that THI does (Adamchic et al., 2012).

We expect that general health and quality of life questionnaires measure general constructs of health, not the tinnitus-specific construct measured by the TFI. To evaluate *discriminant validity*, TFI global scores were compared with scores on our general health questionnaires (BAI, BDI-II, WHOQOL-BREF) in the same participants. It was predicted that there would be weak to moderate correlations (<0.6) indicating acceptable discriminant validity.

Secondary analyses on the strength of the relationships between the individual TFI subscales and other questionnaires and their subscales were assessed. Previous evaluations suggest the THI and THQ global scores would correlate with the emotional subscale of the TFI (Kennedy et al., 2004; Baguley et al., 2000; Newman et al., 1996; Kuk et al., 1990). We also predicted that the BDI-II and BAI would moderately correlate with scores on the emotional subscale of the TFI, and that WHOQOL-BREF scores would moderately correlate with the Quality of life subscale of the TFI.

2.5.2.3. Responsiveness of the TFI.

With respect to responsiveness, this refers to items that are sensitive to change and confirmation that the questionnaire is able to detect important change (above measurement error; Terwee et al., 2007). Responsiveness was assessed in terms of the number of questions exhibiting floor and/or ceiling effects (having limited capacity for change), and to the value corresponding to the Smallest Detectable Change. Response frequency distributions were examined at item level to detect floor and ceiling effects. Potentially problematic items were predefined as those rated at the lowest or highest possible response option (i.e. 0 or 10 on 10-point scales) by more than 15% of respondents (Terwee et al., 2007). The SEM and Smallest Detectable Change scores were calculated using test-retest data (method described in section 2.5.2.1).

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

Table 1
Descriptive statistics and internal consistency. The maximum score is 100, except for BDI and BAI where the maximum score is 63. Values presented in bold indicate poor internal consistency below the recommended criteria ($\alpha < 0.7$).

Questionnaire/subscale	# Items	Descriptive statistics			Internal consistency	Sample size
		Mean	SD	Range	α	N
Tinnitus Functional Index (TFI)[a]	25	40.6	20.1	4–93	0.80	283
Intrusiveness	3	52.8	21.1	6–93	**0.58**	
Sense of control	3	53.9	23.2	0–100	0.75	
Cognition	3	35.8	27.1	0–100	0.95	
Sleep	3	39.6	32.3	0–100	0.94	
Auditory	3	34.0	27.3	0–100	0.95	
Relaxation	3	54.6	29.2	0–100	0.93	
Quality of life	4	28.2	25.4	0–100	0.90	
Emotional	3	30.3	26.3	0–100	0.91	
Tinnitus Handicap Inventory (THI)[b]	25	37.6	20.1	0–90	0.91	247
Tinnitus Handicap Questionnaire (THQ)[c]	27	41.3	17.9	3–90	0.91	247
Social, emotional and physical functioning	15	39.4	23.2	1–91	0.94	
Hearing ability and unease	8	40.4	22.7	0–98	0.86	
Beck's Depression Inventory-II (BDI-II)[d]	21	8.4	8.2	0–51	0.92	247
Beck's Anxiety Inventory (BAI)[e]	21	5.0	6.4	0–44	0.90	
WHOQOL-BREF *global item 1*[f]	1	39.1	8.0	10–50	–	
Tinnitus loudness VAS-L	–	50.1	22.0	1–100	–	
Tinnitus annoyance rating	–	39.8	30.4	1–100	–	

SD = standard deviation; α = Cronbach's alpha estimates.
[a] (Meikle et al., 2012).
[b] (Newman et al., 1996).
[c] (Kuk et al., 1990).
[d] (Beck et al., 1996).
[e] (Beck and Steer, 1990).
[f] (WHOQOL group, 1998).

3. Results

3.1. Inspection of the distribution of scores

Descriptive statistics for all questionnaire measures, including the TFI subscales are shown in Table 1. Scores on tinnitus severity questionnaires were moderate (~40/100 in each case). For depression and anxiety, mean scores were low, although the range was broad. Cumulative frequency distributions for global TFI, THI and THQ are given in Fig. 2. THI global scores were slightly positively skewed towards the lower end of the scales (i.e. 70% of participants scored below 50). THQ global scores had very few higher value scores with all participants scoring less than 70. Compared with these two questionnaires, the TFI global scores appear to be more evenly distributed across the scale, and cover a broad range of scores.

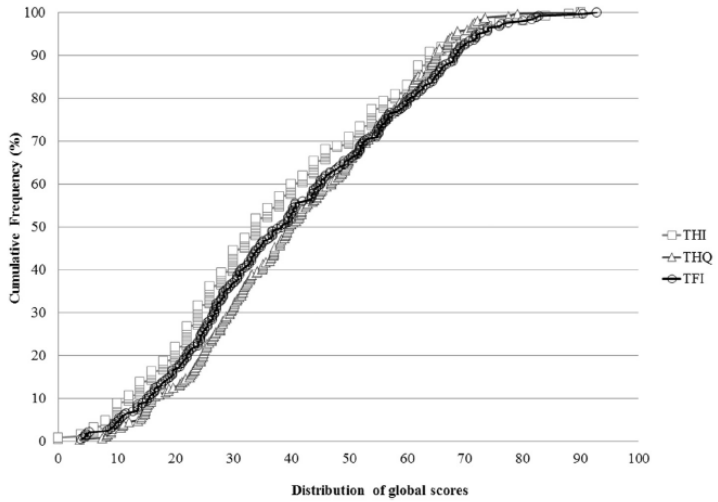

Fig. 2. Cumulative frequency distributions of Tinnitus Functional Index (TFI), Tinnitus Handicap Inventory (THI), and Tinnitus Handicap Questionnaire (THQ) global scores. The percentage of responses for 247 participants on the three different tinnitus questionnaires completed. The graph indicates that the TFI global scores are evenly distributed across the scale, i.e. 100% of participants scored below 90, whilst the THI and THQ global scores distributed towards the lower end, i.e. 70% of participants scored below 50 on the THI and all participants scored less than 70 on the THQ.

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

3.2. Confirmation of the 8-factor structure of the TFI

The initial 8-factor model shown in Fig. 1 was subjected to Confirmatory Factor Analysis.

3.2.1. Factor intercorrelations

Correlation between the first-order factors ranged from very weak ($r = 0.11$) to extremely strong ($r = 0.85$), but most were strong, with 85% above 0.60 (Table 2). The Auditory factor showed unacceptably weak correlations with all the other factors, from an extremely weak correlation with Sleep ($r = 0.11$) to a moderate correlation with Quality of life ($r = 0.43$).

3.2.2. Original model fit

S–B χ^2 was large and significant (χ^2: 578.95; $p < 0.001$) suggesting poor model fit. However, the S–B χ^2 relative to the degrees of freedom (df = 267) was only marginally higher (2.1) than the critical ratio cut-off (≤2.0), suggesting the fit could improve with modifications (Schreiber et al., 2006). The SRMR indicated an acceptable fit. Approximation fit indices also suggested that the model was acceptable albeit less than optimal (Table 3). The TLI and CFI scores were both acceptable, whilst the RMSEA score indicated reasonable fit. Consequently, at this stage, factor loading estimates and modification indices were examined to identify the potential source of the "less than optimal" model fit. The identified parameters were re-specified accordingly, if they improved the model fit and if they were conceptually justified.

3.2.3. Factor loading estimates

The standardised and unstandardised parameter estimates, R-square values and the standard errors are summarised in Table 4. Standardised parameter estimates for the model revealed high factor loading estimates (>0.70) for all the items with their designated factor, except for items 1 and 4, which had factor loadings of 0.68 and 0.57, respectively.

The Auditory and Sleep factors had the weakest factor loadings with the second-order factor. The Auditory factor (F5 in Table 4) loading estimate was only 0.31 indicating a very weak relationship to the second-order factor. The squared factor loadings mirrored these findings (see R^2 in Table 4). For instance, the Sense of control factor only accounted for 33% of the variance in Item 4. The second-order factor of only accounted for 39% of the variance in the Sleep factor (F4 in Table 4) and of most concern, only 9% of the variance in the Auditory factor. The rest of the squared factor loadings for the factors and items ranged from 0.45 to 0.95. From this we conclude that the Auditory factor makes considerably less contribution to the global 'Functional impact of tinnitus' construct than do the other seven factors.

Table 2
Correlations between first-order factors in the Confirmatory Factor Analysis. The correlations between the first-order factors were in general strong, with 85% above 0.60. The Auditory factor showed the weakest correlations with all the other factors. 1 = Intrusiveness; 2 = Sense of control; 3 = Cognition; 4 = Sleep; 5 = Auditory; 6 = Relaxation; 7 = Quality of life; 8 = Emotional. Values presented in bold are below or above the recommended criteria (<0.30 to >0.85).

Factor	1	2	3	4	5	6	7	8
(1) Intrusiveness	1							
(2) Sense of control	0.842	1						
(3) Cognitive	0.640	0.795	1					
(4) Sleep	0.507	0.570	0.562	1				
(5) Auditory	0.328	**0.223**	0.330	**0.114**	1			
(6) Relaxation	0.655	0.814	0.725	0.613	**0.239**	1		
(7) Quality of life	0.655	0.733	0.782	0.465	0.413	0.687	1	
(8) Emotional	0.676	**0.855**	0.784	0.543	**0.197**	0.722	**0.855**	1

3.2.4. Modification index (MI) and expected parameter change (EPC)

Findings indicated the presence of three large MIs that were constrained in the initial 8-factor model. Error covariance (uniqueness) was identified between item 16 "*How much has your tinnitus interfered with your quiet resting activities?*" and item 18 "*How much has your tinnitus interfered with your ability to enjoy 'peace and quiet'?*" (MI: 35.62; EPC: 1.45) on the relaxation subscale, and between item 19 "*How much has your tinnitus interfered with your enjoyment of social activities?*" and item 21 "*How much has your tinnitus interfered with your relationships with family, friends and other people?*" (MI: 25.72; EPC: 1.05) on the Quality of life subscale. Inspection of these items indicated that the large error variance might be attributable to the similarity of the question wording. Therefore, these were freely estimated in the re-specified model (Table 4).

Cross-loading was identified for item 22 (MI: 25.93; EPC: 1.22). Even though item 22 strongly loaded (0.70) onto the Quality of life factor in the initial model; results indicated that it also loaded onto the Cognitive factor. Item 22 asks "*How often did your tinnitus cause you to have difficulty performing your work or other tasks, such as home maintenance, school work, or caring for children or others?*". In this context, the focus is on assessing "difficulties in performing work or tasks" which could be attributed to cognitive processes. There is logic to this cross-loading and although this might marginally lower the loading estimates these parameters were freely estimated in the re-specified model.

3.2.5. Model fit for re-specified model

The SRMR improved and the approximation fit indices were all within desirable limits (Table 3), although S–B χ^2 remained <0.001, the χ^2/df ratio was now 1.89 so within the critical cut-off of <2.0. RMSEA improved slightly (to 0.056), while TLI and CFI were similar to those of the original model (Table 3). Re-specification of the parameters identified as error covariance marginally reduced the factor loading estimate for those items associated with the error, suggesting that the items loading estimates were previously inflated with unique variance. Although factor loading estimates were expected to marginally fall due to the cross-loading, re-specification of the parameters to adjust for cross-loading item 22 substantially reduced the loading estimates for this item on both factors (to 0.4 and 0.43, Table 4). This was unexpected. The standardised parameter estimates and R-square values for the final model are given in Fig. 3.

3.3. Psychometric properties of the TFI

3.3.1. Reproducibility of the TFI

Inter-item correlations ranged 0.055 to 0.904 (Appendix A). Most notably, the Auditory subscale items 14 and 15 exhibited extremely low correlations ($r \sim 0.1$) with the Sleep subscale items 10, 11 and 12. Otherwise items generally showed low to moderate correlations with one another, indicating expected variability in item content. Alpha estimates for the global TFI scores were high ($\alpha = 0.80$, Table 1). Alpha estimates for the TFI subscales were also extremely high, except for the Intrusiveness subscale which was low (0.58), and considerably lower than that reported by Meikle for prototype 2 where $\alpha = 0.85$. This lower alpha estimate further indicates poor fitting items within this dataset.

Table 5 summarises test-retest reliability and agreement between two repeated measures. ICC for the TFI global score was 0.91,

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

Table 3
Summary of the model fit. Model based on proposed factor structure and re-specified model for final factor structure with modifications. Following modifications, model fit improved with all fit statistics, but the S–B χ^2, within the desired limits. Therefore the re-specified model represents the best fit of this population data. S–B χ^2 = Satorra & Bentler adjusted Chi-square; SRMR = Standardised Root Mean Square Residual; TLI = Tucker–Lewis Index; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation.

Models	Modifications	S–B $\chi2$ (*df*)	$\chi2/df$	p-value	TLI	CFI	SRMR	RMSEA (95% CI)
Original model	None	578.947 (267)	2.17	<0.001	0.939	0.946	0.06	0.064 (0.057–0.071)
Re-specified model	Error covariance, cross-loading (Q22 with F3)	498.484 (264)	1.89	<0.001	0.954	0.959	0.056	0.056 (0.048–0.064)

Table 4
Parameter estimates, R-squared values and Standard Error for the proposed Confirmatory Factor Analysis Model and Re-specified Model. The factor loadings (standardised/unstandardized), standard errors and squared factor loadings (R-squared) for all 25 observed variables (Items) and the eight first-order factor (factor loadings). Two loading estimates representing the cross-loading for Item 22 are given for the re-specified model. The values presented in bold have poor associations with their designated factor, all below the recommended cut-off <0.40. β = Standardised parameter estimate; B = Unstandardised parameter estimate; SE = Standard Error; R_2 = R-squared. TFI = Tinnitus functional Index; F1 = Intrusiveness; F2 = Sense of control; F3 = Cognition, F4 = Sleep; F5 = Auditory; F6 = Relaxation; F7 = Quality of life; F8 = Emotional.

First order factor	Observed variable	Original model				Re-specified model			
		β	B	SE	R^2	β	B	SE	R^2
Intrusiveness	TFI 1	0.68	1.00		0.45	0.67	1.00		0.45
Intrusiveness	TFI 2	0.69	0.77	0.08	0.48	0.69	0.78	0.08	0.48
Intrusiveness	TFI 3	0.79	1.16	0.11	0.63	0.80	1.17	0.12	0.63
Sense of control	TFI 4	0.57	1.00		**0.33**	0.57	1.00		**0.33**
Sense of control	TFI 5	0.92	1.16	0.11	0.84	0.92	1.16	0.10	0.84
Sense of control	TFI 6	0.72	1.06	0.11	0.52	0.72	1.05	0.11	0.52
Cognitive	TFI 7	0.94	1.00		0.89	0.94	1.00		0.89
Cognitive	TFI 8	0.93	0.96	0.03	0.87	0.93	0.96	0.03	0.87
Cognitive	TFI 9	0.91	0.90	0.03	0.82	0.91	0.90	0.03	0.82
Sleep	TFI 10	0.88	1.00		0.78	0.88	1.00		0.78
Sleep	TFI 11	0.98	1.13	0.04	0.95	0.98	1.13	0.04	0.95
Sleep	TFI 12	0.91	1.04	0.04	0.82	0.91	1.04	0.04	0.82
Auditory	TFI 13	0.92	1.00		0.85	0.92	1.00		0.85
Auditory	TFI 14	0.98	1.10	0.03	0.97	0.98	1.10	0.03	0.97
Auditory	TFI 15	0.89	1.09	0.03	0.79	0.89	1.09	0.03	0.79
Relaxation	TFI 16	0.93	1.00		0.93	0.88	1.00		0.78
Relaxation	TFI 17	0.94	0.98	0.02	0.94	0.98	1.08	0.03	0.97
Relaxation	TFI 18	0.82	0.92	0.04	0.82	0.75	0.89	0.04	0.57
Quality of life	TFI 19	0.83	1.00		0.83	0.80	1.00		0.64
Quality of life	TFI 20	0.91	1.14	0.05	0.91	0.94	1.23	0.07	0.89
Quality of life	TFI 21	0.85	0.95	0.06	0.85	0.81	0.94	0.06	0.65
Quality of life	TFI 22	0.76	0.91	0.06	0.76	**0.43**	0.53	0.09	0.60
Cognitive	TFI 22	–	–	–	–	**0.40**	0.42	0.07	–
Emotional	TFI 23	0.89	1.00		0.89	0.89	1.00		0.80
Emotional	TFI 24	0.90	1.07	0.04	0.90	0.90	1.07	0.04	0.82
Emotional	TFI 25	0.83	0.87	0.04	0.83	0.83	0.87	0.04	0.68
Second order factor									
Functional impact of tinnitus	F1	0.80	1.48	0.14	0.62	0.78	1.47	0.14	0.62
	F2	0.92	1.71	0.17	0.83	0.91	1.71	0.16	0.83
	F3	0.87	2.38	0.10	0.75	0.87	2.37	0.10	0.75
	F4	0.62	1.83	0.15	**0.39**	0.62	1.84	0.15	**0.39**
	F5	**0.31**	0.79	0.15	**0.1**	0.30	0.77	0.16	**0.09**
	F6	0.83	2.36	0.12	0.69	0.84	2.26	0.12	0.70
	F7	0.87	2.10	0.13	0.75	0.86	2.01	0.14	0.74
	F8	0.91	2.28	0.12	0.83	0.92	2.29	0.12	0.84

indicating excellent reliability, and all subscale scores showed similarly high reliability with ICCs ranging 0.81 to 0.95.

In terms of agreement, the Smallest Detectable Change and limits of agreement values for the global and each of the subscale scores were largely comparable. For example, the TFI global scores had a Smallest Detectable Change score of 22.4, whereas the limits of agreement score was 22.2. The Smallest Detectable Change scores are all slightly different than the limits of agreement scores because the $SEM_{consistency}$ score (i.e. $SEM_{consistency}$ of 8.1) is considered in the calculation of the Smallest Detectable Change, but not in the calculation of the limits of agreement.

Some of the repeated measure change scores in TFI global and subscale scores were not within the identified agreement limits. For three participants, the differences between the TFI global scores were outside the defined limits of agreement (more than 22.2 points below the mean difference; Fig. 4). 95% agreement between scores was observed for only one of the eight TFI subscales, Sense of Control, but not the global score (Table 5).

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

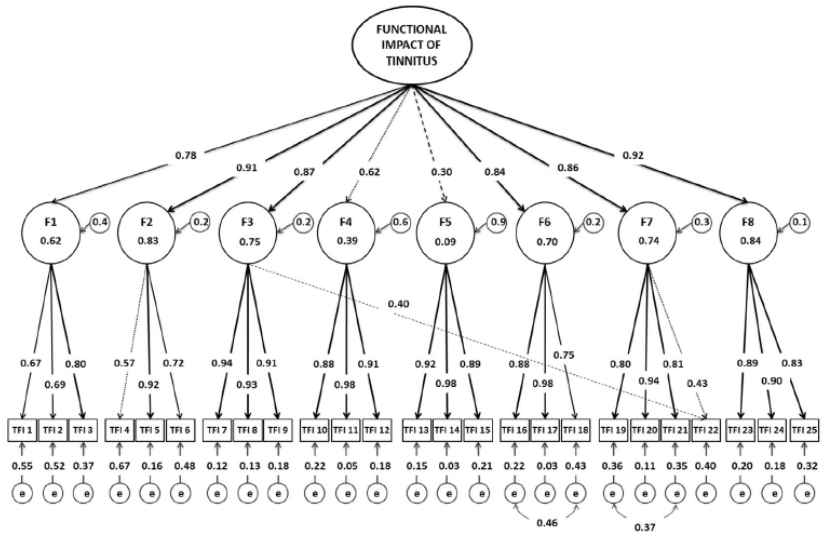

Fig. 3. Illustrative diagram of the re-specified 8-factor model including standardised parameter estimates and r-squared values. The diagram represents the re-specified model results. The standardised parameter estimates indicate the strength of the association between the observed variables, first-order factors and the second-order factor. The unidirectional arrows represent the direct effects of the latent constructs. The solid black unidirectional arrow (⟶) indicates a very strong association (>0.70). The dotted unidirectional arrows (⟶) indicate moderate associations with loading values below 0.65. The dash line unidirectional arrows (·····•) indicate poor associations below the recommended cut-off (<0.40). The residual variance (e) represents the error and unique variance associated with each of the items and the factors. The bidirectional curved arrows (◡◠) represent the association between the error variance. The dotted unidirectional arrow (⟶) from first-order factors; Sense of control (F3) and Quality of life (F7) to the observed variable TFI22 indicates the cross-loading for item 22. F1 = Intrusiveness; F2 = Sense of control; F3 = Cognition, F4 = Sleep; F5 = Auditory; F6 = Relaxation; F7 = Quality of life; F8 = Emotional; e = residual variance (error and uniqueness terms).

Table 5
Reproducibility of Tinnitus Functional Index (TFI) scores: Intra-class correlations (ICC) and limits of agreement between two administrations. The TFI showed excellent stability over time as indicated by the high ICC values and acceptable test-retest agreement. Although most of the subscales were below 95% limits of agreement, it only suggested marginal measurement error. The smallest detectable change scores for the global TFI and subscales are comparable to the limits of agreement. ICC = Intra-class correlations; Mean diff = the mean difference scores between the repeated measure; SEM = Standard error of measurement; SDC = Smallest detectable change.

N = 44	Mean (±SD)		Reliability	Agreement				
Scale	Baseline	Retest	ICC (95%CI)	Mean diff	SEM	SDC	Limits of agreement	% of agreement
Tinnitus Functional index	45.3 (±20.1)	45.6(±19.4)	0.91 (0.84–0.95)	−0.3	8.1	22.4	22.2–22.7	93.2%
Intrusiveness	57.1 (±19.1)	58.8 (±21.3)	0.92 (0.82–0.96)	−1.7	7.6	21.1	19.4–22.7	93.2%
Sense of control	58.1 (±22.8)	57.6 (±20.9)	0.81 (0.65–0.90)	0.5	12.5	34.8	35.3–34.2	95.5%
Cognitive	39.2 (±38.2)	41.9 (±24.3)	0.89 (0.79–0.94)	−2.6	11.8	32.8	30.2–35.5	93.2%
Sleep	41.9 (±31.6)	41.2 (±30.1)	0.91 (0.83–0.95)	0.7	12.8	35.5	36.2–34.8	93.2%
Auditory	33.9 (±29.7)	36.1 (±30.2)	0.95 (0.90–0.97)	−2.3	9.6	26.6	24.3–28.9	93.2%
Relaxation	64.6 (±25.9)	62.9 (±25.3)	0.83 (0.69–0.91)	1.7	13.9	38.5	40.3–36.8	88.6%
Quality of life	35.1 (±26.1)	34.0 (±24.6)	0.86 (0.75–0.92)	1.1	12.6	34.9	36.0–33.8	93.2%
Emotional	36.0 (±28.1)	36.6 (±27.5)	0.87 (0.77–0.93)	−0.6	13.3	36.8	36.2–37.4	91.0%

3.3.2. *Validity of the TFI*

Pearson's correlation coefficients between the global scores on all measures (TFI, THI, THQ, VAS-Loudness, Percentage Annoyance, BDI-II, BAI and global WHOQOL-BREF) are displayed in Table 6.

For convergent validity, results were as predicted. TFI global scores showed strong positive correlations with the THI and THQ global scores ($r = 0.82$ in both cases) and moderate positive correlations with the VAS-Loudness ($r = 0.46$) and Percentage Annoyance ($r = 0.58$). Therefore, the TFI demonstrates acceptable convergent validity indicating that it measures a tinnitus construct that is similar to that measured by other multi-item tinnitus questionnaires.

For most of the TFI subscales, moderate to strong positive pairwise correlations were observed with THI and the THQ global scores (see values for r reported in Table 7). However, when the influence of the remaining subscales were held constant, partial correlation coefficients demonstrated that only the Emotional subscale remained meaningful with a moderate to weak correlation (THI, $pr = 0.31$ and THQ, $pr = 0.29$, respectively) and the Auditory subscale with a moderate correlation (THQ $pr = 0.41$). To confirm the strength of the association between the TFI subscales and the THI and THQ global scores, a series of

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

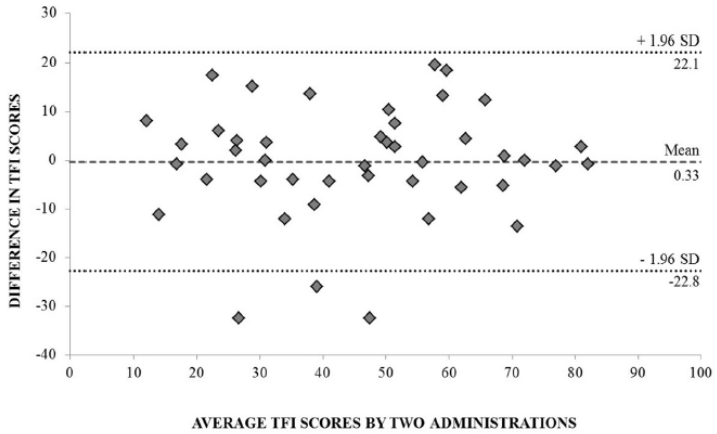

Fig. 4. Bland–Altman plot of test-retest agreement for repeated measures of the TFI global scores. The limits of agreement are represented as ±2 standard deviations from the standard error of measurement. The dotted line denotes the 95% limits of agreement for the TFI global scores. 93% of scores are within the limits of agreement, suggesting marginal measurement error between the repeated measures. Dashed line = mean difference. Dotted lines = limits of agreement (1.96 × SD of the mean difference).

Table 6
Correlations between global scores of all eight measures. The correlations between all eight measures indicate acceptable construct validity for the TFI. The strong correlations (>0.60) between the tinnitus questionnaires show high convergent validity, whilst the moderate correlations (>0.30) with the general health questionnaires show acceptable discriminant validity. TFI: Tinnitus Functional Index = THI; Tinnitus Handicap Inventory = THQ = Tinnitus Handicap Questionnaire, VAS-L = Visual analogue scale for loudness, PR-A = Percentage Rating Annoyance, BDI-II = Beck's Depression Inventory-II, BAI = Beck's Anxiety Inventory, WHOQOL-BREF = World Health Organisation Quality of Life-Bref.

	TFI	THI	THQ	VAS-L	PR-A	BDI	BAI	WHOQOL
TFI	1							
THI	0.82	1						
THQ	0.82	0.79	1					
VAS-L	0.46	0.41	0.29	1				
PR-A	0.58	0.58	0.41	0.42	1			
BDI	0.57	0.60	0.53	0.27	0.31	1		
BAI	0.39	0.43	0.43	0.20	0.19	0.67	1	
WHOQOL	−0.48	−0.52	−0.44	- 0.16	−0.37	−0.55	−0.35	1

multiple linear regression analyses were also conducted (see estimated values for β reported in Table 7). These beta values (β) mirrored the same pattern as shown by the partial correlations indicating that the TFI is measuring similar properties of emotional distress as in the THI and THQ and of auditory difficulties as in the THQ.

Finally, correlations between TFI subscales and the two major subscales of the THQ were examined (Table 8). The THQ subscale 1 assesses the physical, emotional and social effects of tinnitus, while the THQ subscale 2 assesses hearing and communication ability. THQ subscale 1 scores correlated strongly with most TFI subscales, while THQ subscale 2 scores correlated moderately or strongly with all TFI subscales. However, when the influence of

Table 8
Correlation coefficients (*r*), partial correlation coefficients (*pr*) and beta (β) values for the Tinnitus Functional Index (TFI) subscales and the two major subscales of the Tinnitus Handicap Questionnaire (THQ). *r* = Pearson's correlation coefficient; *Pr* = partial correlation coefficient; β = Standardised Beta values.

	THQ	*factor*	1	*THQ*	*factor*	2
	r	*pr*	β	*r*	*pr*	β
Intrusiveness	0.48	−0.13	−0.09	0.27	−0.15	−0.12
Sense of control	0.65	0.04	0.04	0.25	−0.02	−0.02
Cognition	0.75	0.21	0.19	0.42	0.10	0.11
Sleep	0.64	0.31	0.21	0.16	−0.02	−0.02
Auditory	0.21	−0.01	−0.01	0.77	0.71	0.68
Relaxation	0.68	0.14	0.11	0.26	−0.03	−0.03
Quality of life	0.73	0.19	0.18	0.52	0.25	0.27
Emotional	0.81	0.36	0.37	0.31	0.01	0.23

Table 7
Correlation coefficients (*r*), partial correlation coefficients (*pr*) and beta (β) values for the Tinnitus Functional Index (TFI) subscales and the Tinnitus Handicap Inventory (THI) global score, Tinnitus Handicap Questionnaire (THQ) global score, Beck's Depression Inventory-II (BDI-II), Beck's Anxiety Inventory (BAI); and World Health Organisation Quality of Life–BREF (WHOQOL-BREF). *r* = Pearson's correlation coefficient; *Pr* = partial correlation coefficient; β = Standardised Beta values.

TFI subscale	*THI*			*THQ*			*BDI-II*			*BAI*			*WHOQOL*		
	r	*pr*	β	*r*	*pr*	β	*r*	*pr*	β	*r*	*pr*	β	*r*	*pr*	β
Intrusiveness	0.58	0.13	0.10	0.49	0.15	−0.11	0.29	−0.09	−0.10	0.14	−0.16	−0.19	−0.29	−0.00	−0.00
Sense of control	0.64	0.00	0.00	0.60	0.02	0.02	0.35	−0.19	−0.24	0.23	0.10	−0.14	−0.34	0.10	0.15
Cognition	0.72	0.09	0.09	0.73	0.19	0.17	0.58	0.25	0.34	0.39	0.14	0.22	−0.42	−0.01	−0.02
Sleep	0.58	0.19	0.14	0.54	0.21	0.15	0.40	0.07	0.07	0.28	0.09	0.10	−0.34	−0.03	−0.03
Auditory	0.22	0.06	−0.03	0.46	0.41	0.28	0.20	0.07	0.06	0.20	0.16	0.16	−0.04	0.13	0.12
Relaxation	0.66	0.10	0.09	0.63	0.10	0.09	0.44	0.06	0.07	0.27	−0.01	−0.02	−0.43	−0.12	−0.17
Quality of life	0.75	0.27	0.28	0.75	0.22	0.22	0.53	0.02	0.03	0.35	−0.02	−0.04	−0.47	−0.13	−0.20
Emotional	0.79	0.31	0.33	0.74	0.29	0.30	0.59	0.30	0.45	0.24	0.24	0.42	−0.53	−0.22	−0.36

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

remaining subscales were held constant, partial correlation coefficients demonstrated that only the TFI Auditory subscale remained meaningfully associated with THQ subscale 2, with a strong correlation ($pr = 0.71$). TFI Emotional and Sleep subscales remained meaningfully associated with THQ subscale 1, with a moderate correlation ($pr = 0.36$ and $pr = 0.31$ respectively). Acceptable convergent validity was therefore only shown by the TFI Auditory subscale and the THQ hearing and communication subscale.

For discriminant validity, results were also as predicted. TFI global scores correlated moderately with BDI-II ($r = 0.57$), BAI ($r = 0.39$), and WHOQOL-BREF global item scores ($r = 0.48$). Therefore, the TFI demonstrates acceptable discriminant validity and is concluded to measures construct(s) that are distinct from those measured by more general health domains.

Partial correlations between individual TFI subscales and general health, with the remaining subscales held constant, yielded a distinct pattern of results. As predicted, the TFI Emotional subscale correlated significantly with all three general health questionnaires (Table 7). Against our prediction, the Quality of life subscale showed only a *weak* negative correlation with WHOQOL-BREF ($pr = -0.13$). The only other notable correlation was the weak correlation between the BDI-II and the TFI Cognitive subscale ($pr = 0.25$). Beta values (β) estimated as part of a series of multiple linear regression mirrored findings from the partial correlation analyses, although they were marginally higher. The Emotional subscale again had the highest β, showing moderate associations with the BDI-II, BAI and WHOQOL-BREF (Table 7). The Cognitive subscale showed a moderate association with the BDI-II, perhaps indicating some sensitivity to aspects of cognitive difficulty associated with generalised depression. Overall, these results suggest an acceptable degree of discriminant validity. The partial correlations and beta values indicate as expected that the BDI-II and BAI are greatly associated with the emotional subscale, whilst unexpectedly the WHOQOL-BREF only showed a small association with the Quality of life subscale.

3.3.3. Responsiveness of the TFI

Response frequency distributions for each item on the TFI were examined for floor and ceiling effects (Fig. 5: Appendix B). Seventeen out of 25 items failed to meet the *a priori* definition of non-significant floor or ceiling effects (i.e. ratings of either 0 points (floor effect) or 10 point (ceiling effect) being observed in no more than 15% of respondents on the 11-point scale). More specifically 15 items showed floor effects, with '0' being observed for between 16 and 41% of participants (items 24, 13, 10, 9, 8, 11, 12, 15, 23, 14, 20, 19, 22, 21, and 25, respectively). Two items showed a ceiling effect, with responses of 10 being observed for 22% and 25% of the population (items 4 and 18, respectively).

Smallest Detectable Change scores were identified for the TFI global and subscale scores (Table 5). For the TFI global score, the Smallest Detectable Change score was above or below 22.4. Change scores above 22.4 were taken to detect true changes related to worsening or improvement of tinnitus. For example, if a change in TFI global score of 23 was observed, it is reasonable to assume that this reflects real change rather than measurement error. For the TFI subscales, Smallest Detectable Change scores were in general larger than the global score Smallest Detectable Change, ranging from 21.1 (Intrusiveness subscale) to 38.5 (Relaxation subscale). Therefore, the subscale scores would have to have large changes before a "true change" is represented.

4. Discussion

Although only recently developed, the TFI has been implemented as a baseline assessment and outcome measure in numerous research studies (including Henry et al., 2015; Krings et al., 2015; Michiels et al., 2014; Shekhawat et al., 2014; Wilson

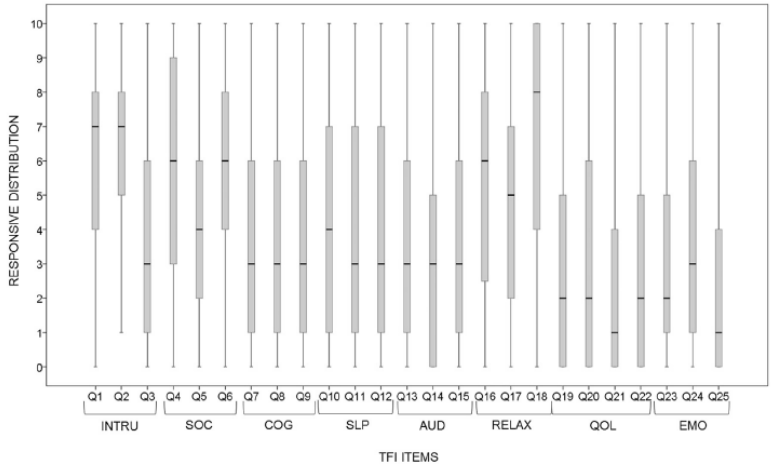

Fig. 5. Response frequency distributions for each Tinnitus Functional Index item within their subscales allowing for examination of floor and ceiling effects. Ceiling effects are evident from the position of the upper quartile and medium on the upper end of the scale, i.e. on response options 9 and 10. Item 4 and item 18 both show ceiling effects. For example, the upper quartile for item 18 is at the end of the scale, indicating that 25% of people endorsed the highest category (10) and the medium indicates that over 50% of participants selected the response options 8, 9, and 10. The floor effects are evident by the position of the first quartile and medium on the lower end of the scale, i.e. on response options 0 and 1. Fifteen items showed floor effects. For example, the lower quartile and medium for item 25 indicates that 50% of participants selected response options 1 and 0. This suggests that these items are limited in their detection of change in tinnitus severity, reducing the responsiveness of the TFI. TFI = Tinnitus Functional Index; INTRU = Intrusiveness; SOC = Sense of control; COG = Cognition; SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional.

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

et al., 2015). The psychometric evaluation performed here however provides the first account of how reliably the TFI measures tinnitus severity and how well it distinguishes between individual differences in tinnitus-related distress in a research population. We raise a number of important points for discussion and reach a number of specific conclusions on the use of the TFI in a UK research population:

4.1. The global TFI is a composite measure of the functional impact of tinnitus

According to our psychometric evaluation, the TFI generally performed adequately as a good measure of functional impact of tinnitus. It has good construct validity and converged on the same construct of tinnitus severity as other multi-item tinnitus questionnaires. In particular, the emotional aspects as measured by the TFI were strongly associated with the global THI and THQ. From the discriminant validity findings, the TFI score is clearly a different measure from those of generalised depression, anxiety, or quality of life.

Confirmatory Factor Analysis broadly confirmed consistency with the eight-factor structure proposed by Meikle et al. (2012). However, there was some evidence of poor fit to the initial model and this improved when the questionnaire was re-specified to account for error covariance between two pairs of items and cross loading of one item onto two factors. Hence, an alternative TFI structure that slightly differed from that proposed by Meikle et al. (2012) was required to best explain the data captured in the general tinnitus population. The next section discusses several other properties in which discrepancies with the original TFI validation were observed, or new concerns are raised.

4.2. The TFI auditory subscale does not reliably contribute to the functional impact of tinnitus

Inspection of the first-order factors (corresponding to the subscales) revealed a problem with the Auditory factor in so far as it appeared to be unrelated to the other factors and in turn the underlying global construct of the functional impact of tinnitus. Hence, scores on the auditory subscale provide little additional information about the functional impact of tinnitus and in fact are likely to undermine the global TFI score. Internal consistency and reliability of the Auditory factor were both high, indicating that the items measure the same underlying construct, and that the factor can differentiate between individuals. It would therefore be reasonable to consider the auditory subscale as a stand-alone measurement tool. In our research population, the TFI therefore seems to be measuring two distinct theoretical constructs (a composite measure of the functional impact of tinnitus and a specific auditory domain).

Despite the different tinnitus populations, our finding is consistent with the analyses of Meikle et al. (2012) who also observed weak intercorrelations between the Auditory factor and the other seven factors. The authors suggested that there is perhaps, either "a general tinnitus severity factor underlying all eight subscales…[or] a general tinnitus severity factor underlying seven of the eight subscales, with the Auditory subscale representing an underlying specific factor" (p.20). A general issue may be the difficulty patients sometimes have in determining their tinnitus problems as distinct from the problems they have because of hearing loss (Ratnayake et al., 2009).

4.3. There is mixed evidence that the TFI Intrusiveness subscale is a reliable unitary construct and the items that tend to be used most as single-item visual analogue scales are poorly associated with the global construct (functional impact of tinnitus)

Our findings indicate that the Intrusiveness subscale had unacceptably low internal consistency indicating that the three items (TFI 1–3) do not measure the same underlying construct, but instead may be distinct from each other. Questions relate to percentage of time that the respondent is consciously aware or annoyed by the tinnitus (TFI 1 and 3, respectively), and a rating of how strong or loud is the tinnitus (TFI 2). There is no further evidence of this discrepancy in the inter-item correlations or the CFA; all the items had acceptably high loading values.

Some researchers use variants of these questions as single-item visual analogue scales to assess tinnitus severity and to measure treatment-related change (TFI 2 and 3 are good examples). Correlations between global TFI score and the VAS-Loudness and Percentage Annoyance were moderate at best. From this, we conclude that single item measures are not sufficient to capture the complexity of tinnitus symptomatology captured by multi-item instruments. The limitations with single items are widely recognised, they are variably reported to be psychometrically weak, with poor validity, low reliability and poor responsiveness (Adamchic et al., 2012; Hobart et al., 2007; Goebel and Hiller, 1994; Nunnally, 1967) yet are sometimes used as diagnostic or outcome measures in research (e.g. Tass et al., 2012; Vanneste et al., 2013). We recommend single-item measures are not used to measure the therapeutic effectiveness of interventions.

4.4. The TFI quality of life subscale does not assess the full multi-attribute nature of quality of life

Here we observed that the TFI Quality of life subscale did not converge with the single item facet on overall quality of life and general health. It is therefore unlikely that the TFI Quality of life subscale is a surrogate marker for the generic construct of Quality of Life used in health research. Health-related QoL is a ubiquitous concept that has different philosophical, political and health-related definitions, but the World Health Organization (1997) describe it as "individuals' perceptions of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns". Correspondingly, the WHOQOL-BREF measures four domains associated with quality of life; physical health, psychological health, social relationships, and environment. To avoid the risk of making a Type 1 error by making multiple comparisons between the TFI and these different domains, we evaluated only the single item. However, these findings enable us to draw the preliminary conclusion that health-related QoL is unlikely to be captured by the items in the TFI Quality of life subscale. This is explicable given the development of the TFI which collapsed only 'Social Distress', 'Leisure', and 'Work' domains to create the Quality of life subscale (Meikle et al., 2012), certainly leaving out physical health.

4.5. The global TFI score may be poorly responsive to treatment-related change in a research population

Arguably, the single most important factor for clinical trials is the assessment of outcome. Primary outcomes provide the means to determine what interventions are effective and hence to influence therapeutic management strategies. It is essential to identify a primary outcome tool that measures symptom categories and

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

changes that are expected to occur according to the aims of the treatment under investigation (Landgrebe et al., 2012; Langguth et al., 2007).

Substantial floor effects on many items indicated that the TFI would be somewhat limited in its responsiveness to detecting treatment-related benefits in this study population. From our sample of research participants, scores on the majority of the items were close to floor, particularly for items in the Cognitive, Sleep, Auditory and Quality of life subscales. This could be an indication that the items are not related to the underlying construct or that the wording of the items may be misleading indicating a "no problem" response (Terwee et al., 2009; Streiner and Norman, 2008). However, the latter is not indicated by any of the other findings from this study. Further research is warranted to replicate our findings and if necessary to reassess the items for inclusion or their wording. It may be that the TFI is suboptimal for use as a tinnitus outcome instrument in a research volunteer population.

Statistically significant differences in treatment effects provide information only on the error rate between the two interventions. Identification of a minimal change that is clinically meaningful is fundamental in health research and clinical trials. Following Jaeschke et al. (1989), our operational definition of a minimal clinically important difference is the smallest difference in score in the domain of interest which *patients* perceive as beneficial. Generally, a minimal clinically important difference involves patient perception. An important step towards determining minimal important differences is to evaluate the smallest change above measurement error, i.e. the Smallest Detectable Change (Landgrebe et al., 2012; Terwee et al., 2009; Revicki et al., 2008; de Vet et al., 2006a; de Vet et al., 2006b).

Test-retest data was used to identify a Smallest Detectable Change score and results indicated that a change in the TFI global score of at least 22.4 points would be required to represent a true change above measurement error. The magnitude of this change is considerably larger than the 13-point difference proposed by Meikle et al. (2012) as a clinically meaningful change. This discrepancy was larger than expected. It is possible that the statistical method used by Meikle et al. (2012) provided a too conservative estimate. Meikle et al. (2012) used an anchor-based approach and Lipsey's criterion group approach (Lipsey, 1983, 1990), using grouped responses from a global question on self-reported change to anchor the changes on the TFI. Such anchor-based methods do not account for measurement precision which could potentially result in unrealistically low cutoffs that sit within the measurement error (de Vet et al., 2006b; Crosby et al., 2004). Consequently, a change score of 13 points might not be a realistic reflection of true change in score and may still include measurement error.

Given the potential for conflicting results simply arising from whether anchor-based or distribution-based methods are used to calculate the clinically meaningful change score, we recommend an integrated approach using both to identify a clinically meaningful change score that is comparable across methods (Crosby et al., 2004).

5. Conclusions and recommendations

This study provides an overview of the psychometric properties of the TFI when used in research. Our findings lead us to draw the following conclusions:

5.1. Not all of the TFI subscales contribute equally to the composite measure of the functional impact of tinnitus. In particular, the auditory subscale score does not contribute to the functional impact of tinnitus

Generally speaking, the TFI provides an adequate composite measurement tool for evaluating the functional impact of tinnitus. However, researchers should remain aware that not all of the TFI subscales contributed equally to the global TFI scores measured in this tinnitus population. In particular, the Auditory subscale appeared to be measuring something different from that of the other subscales. Further improvements in the TFI that tailors this measurement tool are warranted. We note that Meikle et al. (2012) also observed a similar pattern in their clinical population. One priority area for future research would therefore be to explore the impact of removing the auditory subscale. For example, the Auditory subscale score could be calculated and reported separately.

5.2. The TFI quality of life subscale does not assess generic quality of life

Our current recommendation is to include a multi-attribute health-related QoL measure in research that asks questions about quality of life, and not to rely on this particular TFI subscale for a meaningful interpretation of generic quality of life. Future studies should consider the inclusion of a well-established quality of life scale that generates a global score which seems at least to be responsive to treatment-related change in a clinical population of patients with tinnitus. The HUI3 would seem to be a good candidate (Maes et al., 2011).

5.3. The global TFI score and subscale scores may be poorly responsive to treatment-related change in a research population

We provide a cautious recommendation that the TFI is suboptimal for use as a tinnitus outcome instrument in a research volunteer population. However, this warrants further independent replication. Poor responsiveness could be mitigated to some degree by specifying a lower cut-off score as a participant inclusion criterion, one that is at least as large (if not greater) than the Smallest Detectable Change score. As for making a recommendation about the Smallest Detectable Change score that is clinically meaningful and which considers measurement precision, our recommendation is to use the Smallest Detectable Change score of 23 until further research suggests otherwise.

Psychometric validation is an ongoing process that requires continuous evaluations in a variety of populations to provide the much needed evidence that the measurement tool is appropriate and performs as anticipated (Noble, 1998). For the TFI, the various evaluations are ongoing internationally and so we look forward to better understanding and optimising the use of this questionnaire for research and clinical practice alike.

Acknowledgements

This report is independent research by the National Institute for Health Research Biomedical Research Unit Funding Scheme. The original research study for which the data were collected (RESET2) was part funded by The Tinnitus Clinic. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or The Tinnitus Clinic.

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

Appendix

Appendix A

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
Q1	1.00																								
Q2	0.55	1.00																							
Q3	0.56	0.48																							
Q4	0.26	0.32	0.35	1.00																					
Q5	0.38	0.53	0.60	0.53	1.00																				
Q6	0.50	0.59	0.51	0.41	0.66	1.00																			
Q7	0.39	0.45	0.49	0.42	0.69	0.50	1.00																		
Q8	0.34	0.40	0.48	0.43	0.66	0.46	**0.89**	1.00																	
Q9	0.35	0.46	0.48	0.48	0.70	0.50	**0.84**	**0.84**	1.00																
Q10	0.25	0.34	0.39	0.28	0.52	0.37	0.50	0.50	0.52	1.00															
Q11	0.27	0.34	0.40	0.25	0.50	0.37	0.50	0.49	0.50	0.86	1.00														
Q12	0.34	0.35	0.44	0.27	0.50	0.39	0.49	0.49	0.48	0.79	0.89	1.00													
Q13	0.25	0.26	0.21	0.13	0.18	0.22	0.26	0.30	0.26	0.09	0.07	0.10	1.00												
Q14	0.25	0.23	0.22	0.14	0.18	0.18	0.29	0.34	0.27	**0.11**	**0.11**	**0.14**	**0.90**	1.00											
Q15	0.25	0.25	0.18	0.09	0.17	0.18	0.28	0.30	0.25	0.06	0.07	**0.11**	0.82	0.87	1.00										
Q16	0.34	0.46	0.48	0.44	0.64	0.58	0.61	0.58	0.62	0.52	0.50	0.51	0.23	0.21	0.19	1.00									
Q17	0.35	0.47	0.53	0.47	0.67	0.60	0.69	0.65	0.71	0.59	0.59	0.60	0.22	0.22	0.20	0.87	1.00								
Q18	0.26	0.45	0.41	0.43	0.57	0.53	0.50	0.46	0.52	0.44	0.42	0.43	0.21	0.19	0.17	0.81	0.74	1.00							
Q19	0.34	0.33	0.47	0.25	0.52	0.40	0.54	0.56	0.55	0.31	0.32	0.34	0.39	0.39	0.35	0.50	0.56	0.37	1.00						
Q20	0.31	0.42	0.56	0.39	0.66	0.46	0.64	0.66	0.67	0.38	0.40	0.40	0.30	0.30	0.28	0.60	0.67	0.51	0.76	1.00					
Q21	0.31	0.28	0.51	0.28	0.53	0.37	0.59	0.62	0.58	0.36	0.36	0.39	0.37	0.39	0.33	0.45	0.55	0.37	0.78	0.75	1.00				
Q22	0.29	0.35	0.43	0.34	0.58	0.38	0.67	0.68	0.66	0.38	0.42	0.41	0.27	0.32	0.31	0.47	0.53	0.41	0.59	0.68	0.64	1.00			
Q23	0.29	0.43	0.52	0.39	0.68	0.51	0.65	0.63	0.65	0.44	0.46	0.49	0.14	0.17	0.13	0.56	0.65	0.48	0.57	0.73	0.66	0.67	1.00		
Q24	0.34	0.49	0.54	0.45	0.70	0.60	0.66	0.62	0.71	0.46	0.46	0.50	0.16	0.16	0.17	0.58	0.67	0.51	0.56	0.72	0.60	0.60	**0.81**	1.00	
Q25	0.24	0.39	0.48	0.39	0.65	0.46	0.61	0.60	0.66	0.43	0.41	0.42	0.20	0.20	0.19	0.50	0.60	0.45	0.51	0.69	0.61	0.55	0.74	0.74	1

Values presented in bold are below or above the recommended criteria (<0.30 to >0.85).

Appendix B

	Scale items	Percentage of responses for items on the TFI											Mean (±SD)	
		0	1	2	3	4	5	6	7	8	9	10		
Int1	What percentage of your time awake were you consciously aware of your tinnitus?	0.4	6.4	7.1	8.1	7.1	10.2	8.1	11.0	17.3	12.4	12	6.20	(2.79)
Int2	How strong or loud was your tinnitus?	0.0	1.4	3.2	9.5	10.2	12.4	12.0	20.8	19.8	7.4	3	6.14	(2.10)
Int3	What percentage of your time awake were you annoyed by your tinnitus?	9.2	23.3	13.1	12.0	4.2	11.7	5.7	9.5	6.7	3.2	1	3.58	(2.76)
SOC4	Did you feel in control in regard to your tinnitus?	5.3	5.3	8.1	11.0	6.7	9.9	5.3	7.4	10.2	9.2	**22**	5.95	(3.27)
SOC5	How easy was it for you to cope with your tinnitus?	6.7	8.8	10.6	14.5	10.2	17.0	9.9	15.5	5.3	0.7	1	4.23	(2.38)
SOC6	How easy was it for you to ignore with your tinnitus?	2.1	5.3	6.7	10.2	6.4	13.4	10.2	12.4	14.1	9.9	9	5.84	(2.74)
Cog7	How much did your tinnitus interfere with your ability to concentrate?	**16.3**	11.7	14.1	8.1	5.3	9.5	11.0	12.0	7.1	3.2	2	3.86	(2.91)
Cog8	How much did your tinnitus interfere with your ability to think clearly?	**22.6**	12.4	10.2	11.0	7.1	11.3	7.8	7.8	7.1	1.8	1	3.35	(2.81)
Cog9	How much did your tinnitus interfere with your ability to focus attention on other things beside your tinnitus?	**19.1**	14.8	11.3	13.4	7.4	8.5	8.1	7.8	7.1	2.1	0	3.32	(2.73)
Slp10	How often did your tinnitus make it difficult to fall asleep or stay asleep?	**19.1**	9.9	12.0	7.8	8.1	5.3	4.9	10.6	10.2	4.2	8	4.18	(3.35)
Slp11	How often did your tinnitus cause you difficulty in getting as much sleep as you needed?	**23.0**	11.3	10.6	9.2	4.2	5.7	7.4	7.8	7.8	5.3	8	3.92	(3.42)
Slp12	How much of the time did your tinnitus keep you from sleeping as deeply or as peacefully as you would have liked?	**23.3**	10.6	11.7	10.6	3.9	7.8	3.9	6.4	8.1	7.4	6	3.83	(3.39)
Aud13	How much did your tinnitus interfere with your ability to hear clearly?	**18.4**	13.4	10.2	11.3	7.8	13.1	7.8	8.1	7.1	1.4	1	3.51	(2.76)
Aud14	How much did your tinnitus interfere with your ability to understand people who are talking?	**25.4**	11.0	11.3	11.7	8.1	7.8	7.4	6.4	8.5	1.1	1	3.20	(2.85)
Aud15	How much did your tinnitus interfere with your ability to follow conversations in a group or at meetings?	**23.3**	9.9	10.2	10.6	10.2	6.0	6.4	9.2	5.3	5.3	4	3.61	(3.10)
Relx16	How much did your tinnitus interfere with your quiet resting activities?	8.8	7.4	8.8	8.1	7.1	9.2	9.5	12.4	14.8	6.7	7	5.17	(3.06)
Relx17	How much did your tinnitus interfere with your ability to relax?	9.5	11.0	9.9	8.1	9.9	11.0	6.4	14.5	8.8	7.1	4	4.62	(2.98)
Relx18	How much did your tinnitus interfere with your ability to peace and quiet?	4.9	4.2	5.3	7.4	6.0	6.0	5.3	10.6	12.0	12.7	**25**	6.62	(3.18)
QOL19	How much did your tinnitus interfere with your enjoyment of social activities?	**32.9**	11.0	12.4	9.5	4.9	7.4	5.3	7.4	5.3	1.8	2	2.84	(2.92)
QOL20	How much did your tinnitus interfere with your enjoyment of life?	**25.8**	13.1	11.7	11.3	6.0	6.4	6.4	7.4	4.6	3.9	4	3.23	(3.05)
QOL21	How much did your tinnitus interfere with your relationships with family, friends and other people?	**37.5**	17.3	9.5	7.4	4.2	8.8	4.6	3.5	3.5	2.5	1	2.33	(2.72)
QOL22	How often did your tinnitus cause you to have difficulty performing your work or other tasks, such as home maintenance, school work, or caring for children or others?	**32.2**	15.5	11.7	8.1	4.2	7.4	3.5	8.1	5.3	2.8	1	2.74	(2.90)
Emo23	How anxious or worried has your tinnitus made you feel?	**23.7**	16.6	14.8	10.6	3.5	8.5	5.7	8.1	4.9	1.8	2	2.99	(2.81)
Emo24	How bothered or upset have you been because of your tinnitus?	14.8	15.9	14.8	7.4	6.7	12.0	5.7	8.8	5.3	6.0	2	3.72	(2.97)
Emo25	How depressed were you because of your tinnitus?	**41.0**	14.1	10.2	8.1	3.9	7.8	5.7	3.9	2.5	2.5	0	2.20	(2.63)

Values presented in bold exceed the recommended criteria (endorsed by >15% of respondents).

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

References

Adamchic, I., Tass, P.A., Langguth, B., Hauptmann, C., Koller, M., Schecklmann, M., Zeman, F., Landgrebe, M., 2012. Linking the tinnitus questionnaire and the subjective clinical global impression: which differences are clinically important? Health Qual. Life Outcomes 10, 79.

Andersson, G., 2003. Tinnitus loudness matching in relation to annoyance and grading of severity. Auris Nasus Larynx 30, 129–130.

Baguley, D.M., Andersson, G., 2003. Factor analysis of the tinnitus handicap inventory. Am. J. Audiol. 12, 31–34.

Baguley, D.M., Humphriss, R.L., Hodson, C.A., 2000. Convergent validity of the tinnitus handicap inventory and the tinnitus questionnaire. J. Laryngol. Otol. 114, 840–843.

Beck, A.T., Steer, R.A., 1990. Manual for the Beck Anxiety Inventory. The Psychological Corporation, San Antonio, TX.

Beck, A.T., Steer, R.A., Brown, G.K., 1996. Manual for the Beck Depression Inventory—II, second ed. Psychological Corporation, San Antonio, TX.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychol. Bull. 107, 238–246.

Bentler, P.M., 2006. EQS Structural Equations Program Manual. Multivariate Software, Encino, CA.

Bland, M.J., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurements. The Lancet 327, 307–310.

Borg, G., Borg, E., 2001. A new generation of scaling methods: level anchored ratio scaling. Psychologica 28, 15–45.

Boyen, K., Langers, D.R.M., de Kleine, E., van Dijk, P., 2013. Gray matter in the brain: differences associated with tinnitus and hearing loss. Hear. Res. 295, 67–78.

Brown, T.A., 2006. Confirmatory Factor Analysis for Applied Research. Guilford Press.

Brown, T.A., Moore, M.T., 2012. Confirmatory factor analysis. In: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. Guilford Press, pp. 361–379.

Crosby, R.D., Kolotkin, R.L., Williams, G.R., 2004. An integrated method to determine meaningful changes in health-related quality of life. J. Clin. Epidemiol. 57 (11), 1153–1160.

Curran, P.J., West, S.G., Finch, J.F., 1996. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol. Methods 1, 16–29.

Dauman, R., Tyler, R.S., 1992. Some considerations on the classification of tinnitus. In: Aran, J.M., Dauman, R. (Eds.), Tinnitus 91-Proceedings of the Fourth International Tinnitus Seminar. Kugler Publications, Amsterdam, pp. 225–229.

de Vet, H.C., Terwee, C.B., Knol, D.L., Bouter, L.M., 2006a. When to use agreement versus reliability measures. J. Clin. Epidemiol. 59, 1033–1039.

de Vet, H.C., Terwee, C.B., Ostelo, R.W., Beckerman, H., Knol, D.L., Bouter, L.M., 2006b. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual. Life Outcomes 4, 54.

de Vet, H.C., Terwee, C.B., Mokkink, L.B., Knol, D.L., 2011. Measurement in Medicine: a Practical Guide. Cambridge University Press.

Department of Health, 2009. Provision of Services for Adults with Tinnitus. A Good Practice Guide. Central Office of Information, London.

Dozois, D.J., Dobson, K.S., Ahnberg, J.L., 1998. A psychometric evaluation of the beck depression inventory–II. Psychol. Assess. 10, 83.

Fackrell, K., Hall, D.A., Barry, J., Hoare, D.J., 2014. Tools for tinnitus measurement: development and validity of questionnaires to assess handicap and treatment effects. In: Signorelli, F., Turjman, F. (Eds.), Tinnitus: Causes, Treatment and Short & Long-term Health Effects. Nova Science Publishers Inc, New York, pp. 13–60.

Floyd, F.J., Widaman, K.F., 1995. Factor analysis in the development and refinement of clinical assessment instruments. Psychol. Assess. 7, 286–299.

Goebel, G., Hiller, W., 1994. Tinnitus-Fragebogen (TF). Standardinstrument zur Graduierung des Tinnitusschweregrades. Erbebnisse einer Multicenterstudie mit dem Tinnitus-Fragebogen (TF). HNO – Hals-, Nasen-, Ohrenärzte 42, 166–172.

Haynes, S.N., Richard, D.C.S., Kubany, E.S., 1995. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol. Assess. 7, 238–274.

Henry, James A., Frederick, Melissa, Sell, Sara, Griest, Susan, Abrams, Harvey, 2015. Validation of a novel combination hearing aid and tinnitus therapy device. Ear Hear. 36 (1), 42–52.

Hoare, D.J., Pierzycki, R.H., Thomas, H., McAlpine, D., Hall, D.A., 2013. Evaluation of the acoustic coordinated reset (CR®) neuromodulation therapy for tinnitus: study protocol for a double-blind randomized placebo-controlled trial. Trials 14, 207.

Hoare, D.J., Van Labeke, N., McCormack, A., Sereda, M., Smith, S., Al Taher, H., Kowalkowski, V.L., Sharples, M., Hall, D.A., 2014a. Gameplay as a source of intrinsic motivation in a randomized controlled trial of auditory training for tinnitus. PLoS One 9, e107430.

Hoare, D.J., Edmondsn-Jones, M., Gander, P.E., Hall, D.A., 2014b. Agreement and reliability of tinnitus loudness matching and pitch likeness rating. PLoS One 9, e114553.

Hobart, J.C., Cano, S.J., Zajicek, J.P., Thompson, A.J., 2007. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 6, 1094–1105.

Hu, L.T., Bentler, P.M., 1998. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol. Methods 3, 424.

Hu, L.T., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. A Multidiscip. J. 6, 1–55.

Jaeschke, R., Singer, J., Guyatt, G.H., 1989. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin. Trials 10 (4), 407–415.

Kennedy, V., Wilson, C., Stephens, D., 2004. Quality of life and tinnitus. Audiol. Med. 2, 29–40.

Krings, J.G., Wineland, A., Kallogjeri, D., Rodebaugh, T.L., Nicklaus, J., Lenze, E.J., Piccirillo, J.F., 2015. A novel treatment for tinnitus and tinnitus-related cognitive difficulties using computer-based cognitive training and D-Cycloserine. JAMA Otolaryngol. – Head Neck Surg. 141, 18–26.

Kuk, F.K., Tyler, R.S., Russell, D., Jordan, H., 1990. The psychometric properties of a tinnitus handicap questionnaire. Ear Hear. 11, 434–445.

Landgrebe, M., Azevedo, A., Baguley, D., Bauer, C., Cacace, A., Coelho, C., et al., 2012. Methodological aspects of clinical trials in tinnitus: a proposal for an international standard. J. Psychosom. Res. 73, 112–121.

Langguth, B., Goodey, R., Azevedo, A., Bjorne, A., Cacace, A., Crocetti, A., et al., 2007. Consensus for tinnitus patient assessment and treatment outcome measurement: tinnitus research initiative meeting, regensburg, July 2006. Prog. Brain Res. 166, 525–536.

Langguth, B., Kleinjung, T., Landgrebe, M., 2011. Tinnitus: the complexity of standardization. Eval. Health Prof. 34 (4), 429–433, 0163278710394337.

Lipsey, M.W., 1983. A scheme for assessing measurement sensitivity in program evaluation and other applied research. Psychol. Bull. 94, 152.

Lipsey, M.W., 1990. Design Sensitivity: Statistical Power for Experimental Research. Sage, Newbury Park, CA.

MacCallum, R.C., Roznowski, M., Necowitz, L.B., 1992. Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol. Bull. 111, 490.

Maes, I.H., Joore, M.A., Cima, R.F., Vlaeyen, J.W., Anteunis, L.J., 2011. Assessment of health state in patients with tinnitus: a comparison of the EQ-5D and HUI mark III. Ear Hear. 32, 428–435.

Mardia, K.V., 1971. The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika 58, 105–121.

McCombe, A., Baguley, D., Coles, R., McKenna, L., McKinney, C., Windle-Taylor, P., 2001. Guidelines for the grading of tinnitus severity: the results of a working group commissioned by the British Association of Otolaryngologists, Head and Neck Surgeons, 1999. Clin. Otolaryngol. 26, 388–393.

Meikle, M.B., Stewart, B.J., Griest, S.E., Martin, W.H., Henry, J.A., Abrams, H.B., et al., 2007. Assessment of tinnitus: measurement of treatment outcomes. Prog. Brain Res. 166, 511–521.

Meikle, M.B., Stewart, B.J., Griest, S.E., Henry, J.A., 2008. Tinnitus outcomes assessment. Trends Amplif. 12, 223–235.

Meikle, M.B., Henry, J.A., Griest, S.E., Stewart, B.J., Abrams, H.B., McArdle, R., Myers, P.J., Newman, C.W., Vernon, J.A., 2012. The tinnitus functional index: development of a new clinical measure for chronic, intrusive tinnitus. Ear Hear. 33, 153–176.

Melcher, J.R., Knudson, I.M., Levine, R.A., 2013. Subcallosal brain structure: correlation with hearing threshold at supra-clinical frequencies (>8 kHz), but not with tinnitus. Hear. Res. 295, 79–86.

Menard, S. (Ed.), 2002. Applied Logistic Regression Analysis, vol. 106. Sage.

Michiels, S., De Hertogh, W., Truijen, S., Van de Heyning, P., 2014. Physical therapy treatment in patients suffering from cervicogenic somatic tinnitus: study protocol for a randomized controlled trial. Trials 15, 297.

Muthén, L.K., Muthén, B.O., 2012. Mplus User's Guide, seventh ed. Muthén & Muthén, Los Angeles, CA.

Myers, R.H., 2000. Classical and Modern Regression with Applications (Duxbury Classic), second ed. Duxbury Press, Pacific Grove.

Newman, C.W., Sandridge, S.A., 2004. Tinnitus questionnaires. In: Snow Jr., J.B. (Ed.), Tinnitus: Theory and Management. BC Decker Inc, Ontario, pp. 237–254.

Newman, C.W., Jacobson, G.P., Spitzer, J.B., 1996. Development of the tinnitus handicap inventory. Arch. Otolaryngol. – Head Neck Surg. 122, 143–148.

Newman, C.W., Sandridge, S.A., Jacobson, G.P., 1998. Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcome. J. Am. Acad. Audiol. 9, 153–160.

Newman, C.W., Wharton, J.A., Jacobson, G.P., 1995. Retest stability of the tinnitus handicap questionnaire. Ann. Otol. Rhinol. Laryngol. 104 (9/1), 718–723.

Noble, W., 1998. Self-assessment of Hearing and Related Function. Whurr, London.

Nondahl, D.M., Cruickshanks, K.J., Huang, G.H., Klein, B.E.K., Klein, R., Nieto, F.J., Tweed, T.S., 2011. Tinnitus and its risk factors in the Beaver Dam Offspring Study. Int. J. Audiol. 50, 313–320.

Nunnally, J.C., 1967. Psychometric Theory. McGraw-Hill, New York.

Peterson, R.A., 1994. A meta-analysis of Cronbach's coefficient alpha. J. Consum. Res. 21, 381–391.

Pierce, K.J., Kallogjeri, D., Piccirillo, J.F., Garcia, K.S., Nicklaus, J.E., Burton, H., 2012. Effects of severe bothersome tinnitus on cognitive function measured with standardised tests. J. Clin. Exp. Neuropsychol. 34, 126–134.

Ratnayake, S.A., Jayarajan, V., Bartlett, J., 2009. Could an underlying hearing loss be a significant factor in the handicap caused by tinnitus? Noise Health 11 (44), 156–160.

Revicki, D., Hays, R.D., Cella, D., Sloan, J., 2008. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J. Clin. Epidemiol. 61, 102–109.

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

Roberts, L.E., Moffat, G., Bosnyak, D.J., 2006. Residual inhibition functions in relation to tinnitus spectra and auditory threshold shift. Acta Oto Laryngol. 126, 27–33.
Roberts, L.E., Moffat, G., Baumann, M., Ward, L.M., Bosnyak, D.J., 2008. Residual inhibition functions overlap tinnitus spectra and the region of auditory threshold shift. J. Assoc. Res. Otolaryngol. 9, 417–435.
Robinson, S.K., McQuaid, J.R., Viirre, E.S., Betzig, L.L., Miller, D.L., Bailey, K.A., Harris, J.P., Perry, W., 2003. Relationship of tinnitus questionnaires to depressive symptoms, quality of well-being, and internal focus. Int. Tinnitus J. 9, 97–103.
Satorra, A., Bentler, P.M., 1994. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A.E., Clogg, C.C. (Eds.), Latent Variables Analysis: Applications for Developmental Research. Sage Publications, SAGE Publications Inc, Thousand Oaks, CA, pp. 399–419.
Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychol. Methods 7, 147.
Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A., King, J., 2006. Reporting structural equation modeling and confirmatory factor analysis results: a review. J. Educ. Res. 99, 323–338.
Segal, D.L., Coolidge, F.L., Cahill, B.S., O'Riley, A.A., 2008. Psychometric properties of the Beck Depression Inventory—II (BDI-II) among community-dwelling older adults. Behav. Modif. 32, 3–20.
Shekhawat, G.S., Searchfield, G.D., Stinear, C.M., 2014. Randomized trial of transcranial direct current stimulation and hearing aids for tinnitus management. Neurorehabilitation Neural Repair 28, 410–419.
Skevington, S.M., Lotfy, M., O'Connell, K.A., 2004. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual. Life Res. 13, 299–310.
Song, Jae-Jin, Punte, A.K., De Ridder, D., Vanneste, S., Van de Heyning, P., 2013. Neural substrates predicting improvement of tinnitus after cochlear implantation in patients with single-sided deafness. Hear. Res. 299, 1–9.
Steer, R.A., Ranieri, W.F., Beck, A.T., Clark, D.A., 1993. Further evidence for the validity of the beck anxiety inventory with psychiatric outpatients. J. Anxiety Disord. 7, 195–205.
Steiger, J.H., Lind, J.C., 1980. Statistically Based Tests for the Number of Common Factors. Paper Presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.
Stevens, C., Walker, G., Boyer, M., Gallagher, M., 2007. Severe tinnitus and its effect on selective and divided attention. Int. J. Audiol. 46, 208–216.
Streiner, D.L., Norman, G.R., 2008. Health Measurement Scales: a Practical Guide to Their Development and Use. Oxford University Press.
Szczepek, A.J., Haupt, H., Klapp, B.F., Olze, H., Mazurek, B., 2014. Biological correlates of tinnitus-related distress: an exploratory study. Hear. Res. 318, 23–30.
Tabachnick, B.G., Fidell, L.S., 2013. Using Multivariate Statistics, sixth ed. Pearson, Boston.
Tass, P.A., Adamchic, I., Freund, H.J., von Stackelberg, T., Hauptmann, C., 2012. Counteracting tinnitus by acoustic coordinated reset neuromodulation. Restor. Neurol. Neurosci. 30, 137–159.
Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A., Knol, D.L., Dekker, J., Bouter, L.M., de Vet, H.C., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34–42.
Terwee, C.B., Roorda, L.D., Knol, D.L., De Boer, M.R., De Vet, H.C., 2009. Linking measurement error to minimal important change of patient-reported outcomes. J. Clin. Epidemiol. 62, 1062–1067.
Tinnitus Case history questionnaire. www.tinnitusresearch.org/en/consensus/consensus_en.php (date accessed: 27.01.15.).
Tucker, L.R., Lewis, C., 1973. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 38, 1–10.
Tunkel, D.E., Bauer, C.A., Sun, G.H., Rosenfeld, R.M., Chandrasekhar, S.S., Cunningham, E.R., Archer, S.M., Whamond, E.J., 2014. Clinical practice guideline tinnitus. Otolaryngol. – Head Neck Surg. 151, S1–S40.
Tyler, R.S., Baker, L.J., 1983. Difficulties experienced by tinnitus sufferers. J. Speech Hear. Disord. 48, 150–154.
Vanneste, S., van Dongen, M., De Vree, B., Hiseni, S., van der Velden, E., Strydis, C., Joos, K., Norena, A., Serdijn, W., De Ridder, D., 2013. Does enriched acoustic environment in humans abolish chronic tinnitus clinically and electrophysiologically? A double blind placebo controlled study. Hear. Res. 296, 141–148.
WHOQOL group, The, 1998. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol. Med. 28, 551–558.
Wilson, M.B., Kallogjeri, D., Joplin, C.N., Gorman, M.D., Krings, J.G., Lenze, E.J., Nicklaus, J.E., Spitznagel, E.E., Piccirillo, J.F., 2015. Ecological momentary assessment of tinnitus using smartphone technology a pilot study. Otolaryngol. – Head Neck Surg. 152, 897–903. Epub ahead of print: 0194599815569692.
World Health Organization, 1997. Measuring Quality of Life. World Health Organization.

Please cite this article in press as: Fackrell, K., et al., Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population, Hearing Research (2015), http://dx.doi.org/10.1016/j.heares.2015.09.009

REFERENCES

Adamchic, I., Tass, P.A., Langguth, B., Hauptmann, C., Koller, M., Schecklmann, M., Zeman, F. and Landgrebe, M. (2012a). Linking the Tinnitus Questionnaire and the subjective Clinical Global Impression: Which differences are clinically important? *Health and Quality of Life Outcomes*. 10. p.p. 79.

Adamchic, I., Langguth, B., Hauptmann, C. and Tass, P.A. (2012b). Psychometric evaluation of Visual Analog Scale for the assessment of chronic tinnitus. *American Journal of Audiology*. 21. p.pp. 215–226.

Agbo, A.A. (2014). Cronbach's Alpha: Review of Limitations and Associated Recommendations. *Journal of Psychology in Africa*. 20 (2). p.pp. 233–239.

Altman, D.G. and Bland, J.M. (2013). Measurement in Medicine : the Analysis of Method Comparison Studiest. *Journal of the Royal Statistical Society*. 32 (3). p.pp. 307–317.

Andersson, G., Baguley, D.M., McKenna, L. and McFerran, D. (2005). *Tinnitus: A multidisciplinary approach.* London: Whurr Publishers Ltd.

Andersson, G., Kaldo-Sandström, V., Ström, L. and Strömgren, T. (2003). Internet administration of the Hospital Anxiety and Depression Scale in a sample of tinnitus patients. *Journal of Psychosomatic Research*. 55 (3). p.pp. 259–262.

Andersson, G., Lyttkens, L. and Larsen, H.C. (1999). Distinguishing levels of tinnitus distress. *Clinical Otolaryngology*. 24. p.pp. 404–410.

Andresen, E.M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation*. 81 (2). p.pp. 15–20.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*. 43 (4). p.pp. 561–573.

Andrich, D. (1988). *Rasch models for measurement*. California: Sage Publications.

Andrich, D. and Van Schoubroeck, L. (1989). The General Health Questionnaire : a psychometric analysis using latent trait theory. *Psychological Medicine*. 19. p.pp. 469–485.

Andrich, D., Sheridan, B. and Lou, G. (2009). *RUMM2030*. Perth, Australia: RUMM Laboratory.

Atcherson, S.R., Zraick, R.I. and Brasseux, R.E. (2011). Readability of patient-reported outcome questionnaires for use with persons with tinnitus. *Ear and hearing*. 32 (5). p.pp. 671–673.

Atkinson, G. and Nevill, A.M. (1998). Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Medicine*. 26 (4). p.pp. 217–238.

Baguley, D.M. (2002). Mechanisms of tinnitus. *British Medical Bulletin*. 63. p.pp. 195–212.

Baguley, D.M. and Andersson, G. (2003). Factor analysis of the tinnitus handicap inventory. *American Journal of Audiology*. 12 (1). p.pp. 31–34.

Baguley, D.M., Humphriss, R.L. and Hodgson, C.A. (2000). Convergent validity of

the tinnitus handicap inventory and the tinnitus questionnaire. *The Journal of laryngology and otology*. 114 (11). p.pp. 840–843.

Bartlett, J.W. and Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 31 (4). p.pp. 466–475.

Beaton, D.E., Boers, M. and Wells, G.A. (2002). Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology*. 14 (2). p.pp. 109–114.

Beck, A.T. and Steer, R.A. (1990). *Manual for the Beck Anxiety Inventory.* San Antonio, TX: The Psychological Corporation.

Beck, A.T., Steer, R.A., Ball, R., Ciervo, C.A. and Kabat, M. (1997). Use of the Beck Anxiety and Depression Inventories for primary care with medical outpatients. *Assessment*. 4. p.pp. 211–219.

Beck, A.T., Steer, R.A. and Brown, G.K. (1996). *Manual for the Beck Depression Inventory—II*. 2nd Ed. San Antonio, TX: The Psychological Corporation.

Belli, H., Belli, S., Oktay, M.F. and Ural, C. (2012). Psychopathological dimensions of tinnitus and psychopharmacologic approaches in its treatment. *General Hospital Psychiatry*. 34 (3). p.pp. 282–289.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*. 107. p.pp. 238–246.

Bentler, P.M. (2006). *EQS 6 structural equations program manual*.

Bland, J.M. (2015). *An introduction to medical statistics*. 4th Ed. Oxford: Oxford University Press.

Bland, J.M. and Altman, D.G. (2007). Agreement between methods of measurement with multple observations per individual. *Journal of Biopharmaceutical Statistics*. 17 (4). p.pp. 571 – 582.

Bland, J.M. and Altman, D.G. (1996a). Measurement error. *British Medical Journal*. 313. p.p. 744.

Bland, J.M. and Altman, D.G. (1996b). Measurement error proportional to the mean. *BMJ (Clinical research ed.)*. 313. p.p. 106.

Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*. 8 (2). p.pp. 135–160.

Bland, J.M. and Altman, D.G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*. 47 (8). p.pp. 931–936.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurements. *The Lancet*. 327. p.pp. 307–310.

Bland, J.M. and Altman, D.G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*. 314. p.p. 572.

Bond, T. and Fox, C.M. (2007). *Applying the rasch model: Fundamental*

measurement in the human sciences. 2nd Ed. New Jersey: Lawrence Erlbaum Associates Inc.

Borg, G. and Borg, E. (2001). A new generation of scaling methods: level anchored ratio scaling. *Psychologica*. 28. p.pp. 15–45.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Brown, T.A. and Moore, M.T. (2012). Confirmatory factor analysis. *Confirmatory factor analysis for applied research*. p.pp. 1–38.

Byrne, B.M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. East Sussex: Routledge.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*. 56 (2). p.pp. 81–105.

Cano, S.J. and Hobart, J.C. (2011). The problem with health measurement. *Patient preference and adherence*. 5. p.pp. 279–90.

Cella, D., Eton, D.T., Lai, J.S., Peterman, A.H. and Merkel, D.E. (2002). Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *Journal of Pain and Symptom Management*. 24 (6). p.pp. 547–561.

Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of general psychology*. 130 (3). p.pp. 290–304.

Charter, R.A. (1997). Confidence interval procedures for retest, alternate-form, validity, and alpha coefficients. *Perceptual and Motor Skills*. 84 (3c). p.pp. 1488–1490.

Chen, W.H., Lenderking, W., Jin, Y., Wyrwich, K.W., Gelhorn, H. and Revicki, D. a. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*. 23 (2). p.pp. 485–493.

Clark, L.A. and Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*. 7 (3). p.pp. 309–319.

Clauser, B. and Linacre, J.M. (1999). Relating Cronbach and Rasch Reliabilities. *Rasch Measurement Transactions*. 13 (2). p.p. 696.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum.

Copay, A.G., Subach, B.R., Glassman, S.D., Polly, D.W. and Schuler, T.C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The spine journal : official journal of the North American Spine Society*. 7 (5). p.pp. 541–546.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*. 78 (1). p.pp. 98 – 104.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests*.

Psychometrika. 16 (3). p.pp. 297–334.

Cronbach, L.J. and Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological bulletin*. 52 (4). p.pp. 281–302.

Crönlein, T., Langguth, B., Geisler, P. and Hajak, G. (2007). Tinnitus and insomnia. *Progress in Brain Research*. 166. p.pp. 227–233.

Crosby, R.D., Kolotkin, R.L. and Williams, G.R. (2004). An integrated method to determine meaningful changes in health-related quality of life. *Journal of Clinical Epidemiology*. 57 (11). p.pp. 1153–1160.

Crosby, R.D., Kolotkin, R.L. and Williams, G.R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*. 56 (5). p.pp. 395–407.

Crummer, R. and Hassan, G. (2004). Diagnostic Approach to Tinnitus. *American Family Physician*. 69 (1). p.pp. 120–126.

Cumming, G. and Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *The American psychologist*. 60 (2). p.pp. 170–180.

Curran, P.J., West, S.G. and Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*. 1. p.pp. 16–29.

Davis, A. and El Rafaie, A. (2000). Epidemiology of tinnitus. In: R. S. Tyler (ed.). *Tinnitus Handbook*. San Diego: Singular Publishing Group, pp. 1–23.

de Vet, H.C., Terwee, C.B., Knol, D.L. and Bouter, L.M. (2006a). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*. 59. p.pp. 1033–1039.

de Vet, H.C., Terwee, C.B., Mokkink, L.B. and Knol, D.L. (2011). *Measurement in medicine: a practical guide.* Cambridge: Cambridge University Press.

de Vet, H.C.W., Terwee, C.B., Ostelo, R.W., Beckerman, H., Knol, D.L. and Bouter, L.M. (2006b). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and quality of life outcomes*. 4 (Mic). p.p. 54.

Department of Health (2009). *Provision of Services for Adults with Tinnitus. A Good Practice Guide*. London.

DeVon, H.A., Block, M.E., Moyle-Wright, P., Ernst, D.M., Hayden, S.J., Lazzara, D.J., Savoy, S.M. and Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability [Electronic Version]. *Journal of Nursing Scholarship*. 39 (2). p.pp. 155–164.

Doganay Erdogan, B., Elhan, A.H., Öztuna, D., Kûçûkdevecï, A.A. and Kutlay, Ç. (2013). Multiple Imputation of Missing Values Using the Response Function Method Based on a Data Set of the Health Assessment Questionnaire Disability Index. *Turkish Journal of Rheumatology*. 28 (1). p.pp. 2–9.

Dozois, D.J.A., Dobson, K.S. and Ahnberg, J.L. (1998). A psychometric evaluation of the Beck Depression Inventory-II. *Psychological Assessment*. 10 (2). p.pp. 83–89.

Edwards, D.W., Yarvis, R.M., Mueller, D.P., Zingale, H.C. and Wagman, W.J. (1978). Test-Taking and the Stability of Adjustment Scales Can We Assess Patient Deterioration? *Evaluation Review*. 2 (2). p.pp. 275–291.

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Pratap, S., Wentz, R. and Kwan, I. (2002). Increasing response rates to postal questionnaires: systematic review. *BMJ (Clinical research ed.)*. 324 (7347). p.p. 1183.

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L.M. and Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*. (3). p.pp. 2009–2011.

Embertson, S.E. and Reise, S.P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates Inc.

Eng, J. (2005). Receiver operating characteristic analysis: A primer. *Academic Radiology*. 12 (7). p.pp. 909–916.

Erlandsson, S.I. and Holgers, K.M. (2001). The impact of perceived tinnitus severity on health-related quality of life with aspects of gender. *Noise and Health*. 3 (10). p.pp. 39–51.

Eton, D.T., Cella, D., Yost, K.J., Yount, S.E., Peterman, A.H., Neuberg, D.S., Sledge, G.W. and Wood, W.C. (2004). A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *Journal of Clinical Epidemiology*. 57 (9). p.pp. 898–910.

Fabrigar, L.R., Porter, R.D. and Norris, M.E. (2010). Some things you should know about structural equation modeling but never thought to ask. *Journal of Consumer Psychology*. 20 (2). p.pp. 221–225.

Fackrell, K., Hall, D.A., Barry, J.G. and Hoare, D.J. (2016). Psychometric properties of the Tinnitus Functional Index (TFI): assessment in a UK research volunteer population. *Hearing Research*. 335. p.pp. 220–235.

Fackrell, K., Hall, D.A., Barry, J.G. and Hoare, D.J. (2014). Tools for tinnitus measurement: development and validity of questionnaires to assess handicap and treatment effects. In: F. Signorelli & F. Turjman (eds.). *Tinnitus: Causes, Treatment and Short & Long-Term Health Effects*. New York: Nova Science Publishers Inc, pp. 13–60.

Fan, X. and Thompson, B. (2001). *Confidence Intervals for Effect Sizes*. 61 (4). p.pp. 517–531.

Fisher, W.J. (1992). Reliability, Separation, Strata Statistics. *Rasch Measurement Transactions*. 6 (3). p.p. 238.

Fitzpatrick, A. (1983). The meaning of Content Validity. *Applied Psychological Measurement*. 7 (1). p.pp. 3–13.

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D. and Cox, D. (1992). Quality of life measures in health care. I: Applications and issues in assessment. *BMJ*. 305 (October). p.pp. 1074–1077.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003). *Statistical methods for rates and*

proportions. John Wiley.

Floyd, F.J. and Widaman, K.F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*. 7 (3). p.pp. 286–299.

Frei, A., Williams, K., Vetsch, A., Dobbels, F., Jacobs, L., Rüdell, K. and Puhan, M. a (2011). A comprehensive systematic review of the development process of 104 patient-reported outcomes (PROs) for physical activity in chronically ill and elderly people. *Health and Quality of Life Outcomes*. 9 (1). p.p. 116.

Frost, M.H., Reeve, B.B., Liepa, A.M., Stauffer, J.W., Hays, R.D. and Sloan, J. a. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health*. 10 (SUPPL. 2). p.pp. 94–105.

Goebel, G. and Hiller, W. (1998). *Tinnitus-Fragebogen (TF). Ein Instrument zurErfassung von Belastung und Schweregradbei Tinnitus Handanweisung*. Göttingen: Hogrefe.

Goebel, G., Kahl, M., Arnold, W. and Fichter, M. (2006). 15-Year Prospective Follow-Up Study of Behavioral Therapy in a Large Sample of Inpatients With Chronic Tinnitus. *Acta oto-laryngologica.* (126). p.pp. 70–79.

Gopinath, B., McMahon, C.M., Rochtchina, E., Karpa, M.J. and Mitchell, P. (2010). Risk Factors and Impacts of Incident Tinnitus in Older Adults. *Annals of Epidemiology*. 20 (2). p.pp. 129–135.

Gregory, R.J. (2014). *Psychological testing: History, principles, and applications.* 7th Ed. Pearsons.

Griffiths, P. and Murrells, T. (2010). Reliability assessment and approaches to determining agreement between measurements : Classic methods paper. *International Journal of Nursing Studies*. 47. p.pp. 937–938.

Guttman, L. (1950). The problem of attitude and opinion measurement. *Measurement and prediction*. 4. p.pp. 46–59.

Guyatt, G.H., Kirshner, B. and Jaeschke, R. (1992a). A methodologic framework for health status measures: clarity or oversimplification? *Journal of clinical epidemiology*. 45 (12). p.pp. 1353–1355.

Guyatt, G.H., Kirshner, B. and Jaeschke, R. (1992b). Measuring health status: what are the necessary measurement properties? *Journal of clinical epidemiology*. 45 (12). p.pp. 1341–1345.

Hagquist, C. (2001). Evaluating composite health measures using Rasch modelling: an illustrative example. *Sozial- und Praventivmedizin*. 46 (6). p.pp. 369–378.

Hagquist, C. and Andrich, D. (2004). Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences*. 36 (4). p.pp. 955–968.

Hagquist, C., Bruce, M. and Gustavsson, J.P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*. 46 (3). p.pp. 380–393.

Haider, H., Fackrell, K., Kennedy, V. and Hall, D.A. (in press). Dimensions of tinnitus-related complaints reported by patients and their significant others:

protocol for a systematic review. *BMJ Open*.

Hall, D.A., Haider, H., Kikidis, D., Mielczarek, M., Mazurek, B., Szczepek, A.J. and Cederroth, C.R. (2015a). Toward a Global Consensus on Outcome Measures for Clinical Trials in Tinnitus: Report From the First International Meeting of the COMiT Initiative, November 14, 2014, Amsterdam, The Netherlands. *Trends in hearing*. 19. p.pp. 1–7.

Hall, D.A., Haider, H., Szczepek, A.J., Lau, P., Rabau, S., Jones-Diette, J., Londero, A., Edvall, N.K., Cederroth, C.R., Mielczarek, M., Fuller, T., Batuecas-Caletrio, A., Brueggemen, P., Thompson, D.M., Norena, A., Cima, R.F., Mehta, R.L. and Mazurek, B. (in review). Systematic review of outcome domains and instruments used in clinical trials of tinnitus treatments in adults. *Trials*.

Hall, D.A., Szczepek, A.J., Kennedy, V. and Haider, H. (2015b). Current-reported outcome domains in studies of adults with a focus on the treatment of tinnitus: protocol for a systematic review. *BMJ Open*. 5 (11). p.pp. e009091–e009091.

Hallam, R.S. (1996). *Manual of the Tinnitus Questionnaire (TQ).* London: The Psychological Corporation.

Hallam, R.S. (2008). *TQ Manual of the tinnitus questionnaire: Revised and updated.* London: Polpresa Press.

Hallam, R.S., Jakes, S.C. and Hinchcliffe, R. (1988). Cognitive variables in tinnitus annoyance. *British Journal of Clinical Psychology*. 27. p.pp. 213–222.

Hambleton, R.K. and Jones, R.W. (1993). Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice*. 12 (3). p.pp. 38–47.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*. 23. p.pp. 52–62.

Hamon, A. and Mesbah, M. (2002). Questionnaire Reliability Under the Rasch Model. *Statistical Methods for Quality of Life Studies*. p.pp. 155–168.

Hankins, M. (2008). How discriminating are discriminative instruments? *Health and Quality of Life Outcomes*. 6. p.p. 36.

Haynes, S.N., Richard, D.C.S. and Kubany, E.S. (1995). Content Validity in Psychological Assessment : A Functional Approach to Concepts and Methods Introduction to Content Validity. *Psychological Assessment*. 7 (3). p.pp. 238–247.

Hays, R.D., Morales, L.S. and Reise Steve P (2000). Item Response Theory and health outcomes measurement in the 21st century. *Medical care*. 38 (9 Suppl). p.pp. 1128–1142.

Hays, R.D. and Woolley, J.M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *PharmacoEconomics*. 18 (5). p.pp. 419–423.

Henry, J.A., Dennis, K.C. and Schechter, M.A. (2005). General Review of Tinnitus Prevalence, Mechanisms, Effects, and Management. *Journal of speech, language, and hearing research*. 48 (5). p.pp. 1204–1235.

Henry, J.A., Griest, S., Thielman, E., McMillan, G., Kaelin, C. and Carlson, K.F.

(2015). Tinnitus Functional Index: Development, validation, outcomes research, and clinical application. *Hearing Research*.

Henry, J.A., Griest, S., Zaugg, T.L., Thielman, E.J., Kaelin, C., Galvez, G. and Carlson, K.F. (2014). Tinnitus and Hearing Survey: A screening tool to differentiate bothersome tinnitus from hearing difficulties. *American Journal of Audiology*. 24. p.pp. 66–77.

Hesser, H. (2010). Methodological considerations in treatment evaluations of tinnitus distress: A call for guidelines. *Journal of Psychosomatic Research*. 69 (3). p.pp. 305–307.

Hesser, H., Weise, C., Rief, W. and Andersson, G. (2011). The effect of waiting: A meta-analysis of wait-list control groups in trials for tinnitus distress. *Journal of Psychosomatic Research*. 70 (4). p.pp. 378–384.

Hiller, W. and Goebel, G. (1992). A psychometric study of complaints in chronic tinnitus. *Journal of Psychosomatic Research*. 36 (4). p.pp. 337–348.

Hiller, W. and Goebel, G. (2006). Factors influencing tinnitus loudness and annoyance. *Archives of otolaryngology--head & neck surgery*. 132 (12). p.pp. 1323–1330.

Hiller, W. and Goebel, G. (2004). *Rapid assessment of tinnitus-related psychological distress using the Mini-TQ Evaluación rápida del estrés psicológico relacionado*. p.pp. 600–604.

Hiller, W., Goebel, G. and Rief, W. (1994). Relaibility of self-rated tinnitus distress and association with psychological symptom patterns. *British Journal of Clinical Psychology*. 33. p.pp. 231 – 239.

Hoare, D.J., Broomhead, E., Stockdale, D. and Kennedy, V. (2015). Equity in provision of tinnitus services in United Kingdom National Health Service audiology departments. *European Journal for Person Centered Heathcare*. 3 (3). p.pp. 318–326.

Hoare, D.J., Edmondson-Jones, M., Sereda, M., Akeroyd, M. a and Hall, D. (2014). Amplification with hearing aids for patients with tinnitus and co-existing hearing loss. *Cochrane Database of Systematic Reviews*. 1 (10). p.p. CD010151.

Hoare, D.J., Pierzycki, R.H., Thomas, H., McAlpine, D. and Hall, D. a (2013). Evaluation of the acoustic coordinated reset (CR ®) neuromodulation therapy for tinnitus: study protocol for a double-blind randomized placebo-controlled trial. *Trials*. 14 (1). p.p. 207.

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

Hobart, J.C., Cano, S.J., Zajicek, J.P. and Thompson, A.J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *The Lancet Neurology*. 6 (12). p.pp. 1094–1105.

Hoffman, H.J., George, M.A. and Reed, W. (2004). Epidemiology of tinnitus. *Tinnitus Theory and Management*. p.pp. 16–41.

Hogan, T.P., Benjamin, A. and Brezinski, K.L. (2000). Frequency of Use of Various

Types. *Educational and Psychological Measurement*. 60 (4). p.pp. 523–531.

Holmes, S. and Padgham, N.D. (2011). Ringing in the ears'': narrative review of tinnitus and its impact. *Biological research for nursing*. 13 (1). p.pp. 97–108.

Hu, L. and Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 6. p.pp. 1–55.

Hu, L. and Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*. 3 (4). p.pp. 424–453.

Husted, J.A., Cook, R.J., Farewell, V.T. and Gladman, D.D. (2000). Methods for assessing responsiveness: a critical review and recommendations. *Journal of clinical epidemiology*. 53 (5). p.pp. 459–468.

Iacobucci, D. and Duhachek, A. (2003). Advancing Alpha: Measuring Reliability With Confidence. *Journal of Consumer Psychology*. 13 (4). p.pp. 478–487.

Ireland, C.E., Wilson, P.H., Tonkin, J.P. and Platt-hepworth, S. (1985). An evaluation of relaxation training in the treatment of tinnitus. *Behaviour Research and Therapy*. 23 (4). p.pp. 423–430.

Jacobson, G.P. and Newman, C.W. (1990). The development of the Dizziness Handicap Inventory. *Archives of Otolaryngology - Head and Neck Surgery*. 116. p.pp. 424–427.

Jacobson, N.S., Roberts, L.J., Berns, S.B. and McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *Journal of consulting and clinical psychology*. 67 (3). p.pp. 300–307.

Jaeschke, R., Singer, J. and Guyatt, G.H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled clinical trials*. 10 (4). p.pp. 407–415.

Jakes, S., Hallam, R., Chambers, C. and Hinchcliffe, R. (1985a). A factor analytical Study of Tinnitus Complaint Behaviour. *Audiology*. 24. p.pp. 195–206.

Jakes, S.C., Hallam, R.S., Chambers, C. and Hinchcliffe, R. (1985b). A Factor Analytical Study of Tinnitus Complaint Behaviour. *International journal of audiology*. 24 (3) p.pp. 195–206.

Jones, P.W. and Kaplan, R.M. (2003). Methodological issues in evaluating measures of health as outcomes for COPD. *The European respiratory journal. Supplement*. 41. p.p. 13s–18s.

Juniper, E.F., Guyatt, G.H., Willan, a and Griffith, L.E. (1994). Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *Journal of clinical epidemiology*. 47 (1). p.pp. 81–87.

Kaplan, R.M., Ganiats, T.G. and Seiber, W.J. (1996). *The Quality of Well Being Scale, Self-Administered (QWBS-SA).* San Diego: University of California-San Diego.

Kaye, S. and Darke, S. (2002). Determining a diagnostic cut-off on the Severity of Dependence Scale (SDS) for cocaine dependence. *Addiction*. 97. p.pp. 737–731.

Kennedy, V., Wilson, C. and Stephens, D. (2004). Quality of life and tinnitus. *Audiological Medicine*. 2. p.pp. 29–40.

Keszei, A.P., Novak, M. and Streiner, D.L. (2010). Introduction to health measurement scales. *Journal of Psychosomatic Research*. 68 (4). p.pp. 319–323.

Kirshner, B. and Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of chronic diseases*. 38 (1). p.pp. 27–36.

Kleinstäuber, M., Frank, I. and Weise, C. (2015). A confirmatory factor analytic validation of the Tinnitus Handicap Inventory. *Journal of Psychosomatic Research*. 78 (3). p.pp. 277–284.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M. and Streiner, D.L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*. 48 (6). p.pp. 661–671.

Kottner, J. and Streiner, D.L. (2010). Internal consistency and Cronbach's alpha: A comment on Beeckman et al. (2010). *International Journal of Nursing Studies*. 47 (7). p.pp. 926–928.

Krings, J.G., Wineland, A., Kallogjeri, D., Rodebaugh, T.L., Nicklaus, J., Lenze, E.J. and Piccirillo, J.F. (2015). A novel treatment for tinnitus and tinnitus-related cognitive difficulties using computer-based cognitive training and D-Cycloserine. *JAMA Otolaryngology--Head & Neck Surgery*. 141. p.pp. 18–26.

Kuder, G.F. and Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika*. 2 (3). p.pp. 151–160.

Kuk, F.K., Tyler, R.S., Russell, D. and Jordan, H. (1990). The psychometric properties of a Tinnitus Handicap Questionnaire*. *Amplification and Aural Rehabilitation*. 11 (6). p.pp. 434–445.

Kurtaiş, Y., Őztuna, D., Küçükdeveci, A.A., Kutlay, Ş., Hafiz, M. and Tennant, A. (2011). Reliability, construct validity and measurement potential of the ICF comprehensive core set for osteoarthritis. *BMC musculoskeletal disorders*. 12 (1). p.p. 255.

Lamoureux, E.L., Pallant, J.F., Pesudovs, K., Hassell, J.B. and Keeffe, J.E. (2006). The Impact of Vision Impairment Questionnaire: an evaluation of its measurement properties using Rasch analysis. *Investigative ophthalmology & visual science*. 47 (11). p.pp. 4732–4741.

Landgrebe, M., Azevedo, A., Baguley, D., Bauer, C., Cacace, A., Coelho, C., Dornhoffer, J., Figueiredo, R., Flor, H., Hajak, G., van de Heyning, P., Hiller, W., Khedr, E., Kleinjung, T., Koller, M., Lainez, J.M., Londero, A., Martin, W.H., Mennemeier, M., Piccirillo, J., De Ridder, D., Rupprecht, R., Searchfield, G., Vanneste, S., Zeman, F. and Langguth, B. (2012). Methodological aspects of clinical trials in tinnitus: A proposal for an international standard. *Journal of Psychosomatic Research*. 73 (2). p.pp. 112–121.

Langguth, B., Kleinjung, T. and Landgrebe, M. (2011). Tinnitus: The Complexity of Standardization. *Evaluation & the Health Professions*. 34 (4). p.pp. 429–433.

Linacre, J.M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*. 2 (3). p.pp. 266–283.

Linacre, J.M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*. 3 (1). p.pp. 85–106.

Linacre, M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*. 7 (4) p.p. 328.

Lipsey, M. and Cordray, D. (2000). Evaluation methods for social intervention. *Annual Review of Psychology*. 51. p.pp. 345–75.

Lipsey, M.W. (1983). A scheme for assessing measurement sensitivity in program evaluation and other applied research. *Psychological bulletin*. 94 (1). p.pp. 152–165.

Lipsey, M.W. and Hurley, S.M. (2008). Design Sensitivity: Statistical Power for Applied Experimental Research. *The SAGE handbook of applied social research methods*. p.pp. 44–76.

Lockwood, A.H., Salvi, R.J. and Burkard, R.F. (2002). Tinnitus. *The New England Journal of Medicine*. 347 (12). p.pp. 904–910.

Lohr, K.N., Aaronson, N.K., Alonso, J., Burnam, M.A., Patrick, D. l and Perrin, E.B. (1996). Evaluating Quality of Life and Health status instruments: development of scientific review criteria. *Clinical Therapeutics*. 18 (5). p.pp. 979–992.

MacCallum, R., Browne, M. and Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*. 1 (2). p.pp. 130–149.

MacCallum, R.C., Roznowski, M. and Necowitz, L.B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological bulletin*. 111. p.pp. 490–504.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R.D., Brod, M., Snyder, C., Boers, M. and Cella, D. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research*. 21 (5). p.pp. 739–746.

Main, C.J. (1983). The Modified Somatic Perception Questionnaire (MSPQ). *Journal of Psychosomatic Research*. 27. p.pp. 503–514.

Marais, I. and Andrich, D. (2008). Formalizing dimension and response violations of local independence in the tridimensional Rasch model. *Journal of Applied Measurement*. 9 (3). p.pp. 200–215.

Marx, R.G., Menezes, A., Horovitz, L., Jones, E.C. and Warren, R.F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology*. 56. p.pp. 730–735.

Masters, G.N. (1982). A rasch model for partial credit scoring. *Psychometrika*. 47 (2). p.pp. 149–174.

Mayhew, A., Cano, S., Scott, E., Eagle, M., Bushby, K., Muntoni, F. and On Behalf of the NorthStar Clinical Network for Paediatric Neuromuscular Disease (2011). Moving towards meaningful measurement: Rasch analysis of the North Star Ambulatory Assessment in Duchenne muscular dystrophy. *Developmental Medicine and Child Neurology*. 53 (6). p.pp. 535–542.

McCombe, A., Baguley, D., Coles, R., McKenna, L., McKinney, C. and Windle-

Taylor, P. (2001). Guidelines for the grading of tinnitus severity: The results of a working group commissioned by the British Association of Otolaryngologists, Head and Neck Surgeons, 1999. *Clinical Otolaryngology and Allied Sciences*. 26 (5). p.pp. 388–393.

McFadden, D. (1982). *Tinnitus. Facts, Theories, and Treatments.* Washington, DC: National Academy Press.

Meikle, M., Stewart, B., Griest, S.E., Martin, W.H., Henry, J., Abrams, H.B., McArdle, R., Newman, C.W. and Sandridge, S. (2007). Assessment of tinnitus: Measurement of treatment outcomes. *Progress in Brain Research*. 166. p.pp. 511–521.

Meikle, M. and Taylor-Walsh, E. (1984). Characteristics of tinnitus and related observations in over 1800 tinnitus clinic patients. *Journal of laryngology and Otology, (Suppl.)*. 9. p.pp. 17–21.

Meikle, M.B., Henry, J. a, Griest, S.E., Stewart, B.J., Abrams, H.B., McArdle, R., Myers, P.J., Newman, C.W., Sandridge, S., Turk, D.C., Folmer, R.L., Frederick, E.J., House, J.W., Jacobson, G.P., Kinney, S.E., Martin, W.H., Nagler, S.M., Reich, G.E., Searchfield, G., Sweetow, R. and Vernon, J.A. (2012). The Tinnitus Functional Index: development of a new clinical measure for chronic, intrusive tinnitus. *Ear and hearing*. 33 (2). p.pp. 153–76.

Meikle, M.B., Stewart, B.J., Griest, S.E. and Henry, J. a (2008). Tinnitus outcomes assessment. *Trends in amplification*. 12 (3). p.pp. 223–235.

Menard, S. (2002). *Applied logistic regression analysis*. Vol. 106. Sage publishers.

Michiels, S., De Hertogh, W., Truijen, S. and Van de Heyning, P. (2014). Physical therapy treatment in patients suffering from cervicogenic somatic tinnitus: study protocol for a randomized controlled trial. *Trials*. 15. p.p. 297.

Miguel, G.S., Yaremchuk, K., Roth, T. and Peterson, E. (2014). The effect of insomnia on tinnitus. *The Annals of otology, rhinology, and laryngology*. 123 (10). p.pp. 696–700.

Mokkink, L.B., Terwee, C.B., Knol, D.L., Stratford, P.W., Alonso, J., Patrick, D.L., Bouter, L.M. and de Vet, H.C.W. (2010a). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC medical research methodology*. 10. p.p. 22.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M. and de Vet, H.C.W. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*. 19 (4). p.pp. 539–549.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M. and de Vet, H.C.W. (2012). *The COSMIN checklist manual*. Amsterdam: VU University Medical Centre.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M. and de Vet, H.C.W. (2010c). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of*

Clinical Epidemiology. 63 (7). p.pp. 737–745.

Mulcahy, J. and Vaughan, B. (2015). Exploring the construct validity of the Patient Perception Measure – Osteopathy (PPM-O) using classical test theory and Rasch analysis. *Chiropractic & Manual Therapies*. 23 (1). p.pp. 1–12.

Muthén, L. and Muthén, B. (2012). *Mplus User's Guide*. 7th Ed. Los Angeles, CA: Muthén & Muthén.

Myers, R. (2000). *Classical and modern regression with applications*. 2nd Ed. Pacific Grove: Duxbury Press.

Newman, C.W., Jacobson, G.P. and Spitzer, J.B. (1996). *Development of the Tinnitus Handicap Inventory*. 122. p.pp. 143–148.

Newman, C.W. and Sandridge, S.A. (2004). Tinnitus questionnaires. *Tinnitus : theory and management*. p.pp. 237–254.

Newman, C.W., Sandridge, S.A. and Jacobson, G.P. (2014). Assessing Outcomes of Tinnitus Intervention. *Journal of the American Academy of Audiology*. 25 (1). p.pp. 76–105.

Newman, C.W., Sandridge, S.A. and Jacobson, G.P. (1998). Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluation treatment outcome. *Journal of the American Academy of Audiology*. 9 (2). p.pp. 153–160.

Newman, C.W., Wharton, J.A. and Jacobson, G.P. (1995). Retest stability of the Tinnitus Handicap Questionnaire. *The Annals of Otology, Rhinology and Laryngolgy*. 104. p.pp. 718–723.

NHS White Paper (2010). *Department of Health, Equity and excellence: Liberating the NHS*. London.

Nondahl, D.M., Cruickshanks, K.J., Huang, G.-H., Klein, B.E.K., Klein, R., Javier Nieto, F. and Tweed, T.S. (2011). Tinnitus and its risk factors in the Beaver Dam offspring study. *International journal of audiology*. 50 (5). p.pp. 313–320.

Nunnally, J.C. (1978). *Psychometric theory*. 2nd Editio. New York: McGraw-Hill.

Ooms, E., Meganck, R., Vanheule, S., Vinck, B., Watelet, J.-B. and Dhooge, I. (2011). Tinnitus Severity and the Relation to Depressive Symptoms: A Critical Study. *Otolaryngology -- Head and Neck Surgery*. 145 (2). p.pp. 276–281.

Ostini, R. and Nering, M.L. (2006). Polytomous Rasch models. In: *Polytomous Item Response Theory models*. London: Sage Publications, pp. 26–64.

Pallant, J.F., Miller, R.L. and Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC psychiatry*. 6. p.p. 28.

Pallant, J.F. and Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *The British journal of clinical psychology / the British Psychological Society*. 46 (Pt 1). p.pp. 1–18.

Peterson, R.A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*. 21. p.pp. 381–391.

Pierce, K.J., Kallogjeri, D., Piccirillo, J.F., Garcia, K.S., Nicklaus, J.E. and Burton, H. (2012). Effects of severe bothersome tinnitus on cognitive function measured

with standardized tests. *Journal of Clinical and Experimental Neuropsychology*. 34 (2). p.pp. 126–134.

Rabau, S., Wouters, K. and Heyning, P. Van De (2014). *Validation and translation of the Dutch Tinnitus Functional Index*. p.pp. 251–258.

Rasch, G. (1966a). An Individualistic Approach to Item Analysis. *Readings in Mathematical Social Science.* p.pp. 89–107.

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*. 19 (1). p.pp. 49–57.

Ratnayake, S., Jayarajan, V. and Bartlett, J. (2009). Could an underlying hearing loss be a significant factor in the handicap caused by tinnitus? *Noise and Health*. 11 (44). p.pp. 156–160.

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

Redelmeier, D.A., Guyatt, G.H. and Goldstein, R.S. (1996). Assessing the minimal important difference in symptoms: A comparison of two techniques. *Journal of Clinical Epidemiology*. 49 (11). p.pp. 1215–1219.

Reeve, B.B., Wyrwich, K.W., Wu, A.W., Velikova, G., Terwee, C.B., Snyder, C.F., Schwartz, C., Revicki, D. a., Moinpour, C.M., McLeod, L.D., Lyons, J.C., Lenderking, W.R., Hinds, P.S., Hays, R.D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P.M., Cella, D., Brundage, M., Ahmed, S., Aaronson, N.K. and Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research*. 22 (8). p.pp. 1889–1905.

Revicki, D., Hays, R.D., Cella, D. and Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*. 61 (2). p.pp. 102–109.

Riddle, D.L., Stratford, P.W. and Binkley, J.M. (1998). Sensitivity to Change of the Roland- Morris Back Pain Questionnaire : Part 2. *Physical Therapy*. 78 (2). p.pp. 1197–1207.

Roberts, L.E., Moffat, G., Baumann, M., Ward, L. and Bosnyak, D. (2008). Residual inhibition functions overlap tinnitus spectra and the region of auditory threshold shift. *Journal of the Association for Research in Otolaryngology*. 9. p.pp. 417–435.

Roberts, L.E., Moffat, G. and Bosnyak, D.J. (2006). Residual inhibition functions in relation to tinnitus spectra and auditory threshold shift. *Acta Oto-laryngologica*. 126. p.pp. 27–33.

Robinson, S.K., McQuaid, J.R., Viirre, E.S., Betzig, L.L., Miller, D.L., Bailey, K.A., Harris, J.P. and Perry, W. (2003). Relationship of tinnitus questionnaires to depressive symptoms, quality of well-being and internal focus. *The International Tinnitus Journal*. 9 (2). p.pp. 97–103.

Sanchez, L. and Stephens, S.D.G. (1997). A tinnitus problem questionnaire in a clinic population. *Ear and Hearing*. 18 (3). p.pp. 210–217.

Sanchez, L. and Stephens, S.D.G. (2000). Perceived problems of tinnitus clinic

clients at long-term follow up. *Journal of Audiological Medicine*. 9 (2). p.pp. 94–103.

Satorra, A. and Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. *Latent Variable Analysis*. p.pp. 339–419.

Schafer, J.L. and Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods*. 7. p.p. 147.

Schmidt, C.J., Kerns, R.D., Griest, S., Theodoroff, S.M., Pietrzak, R.H. and Henry, J.A. (2014). *Toward Development of a Tinnitus Magnitude Index*. p.pp. 476–484.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*. 8 (4). p.pp. 350–353.

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

Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A. and King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*. 99 (6). p.pp. 323–338.

Schumacker, R.E. and Smith, E. V. (2007). A Rasch Perspective. *Educational and Psychological Measurement*. 67 (3). p.pp. 394–409.

Scientific Advisory Committee of the Medical Outcomes Trust (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*. 11 (3). p.pp. 193–205.

Segal, D.L., Coolidge, F.L., Cahill, B.S. and Riley, A. a O. (2008). *Behavior Modification of the Beck Depression Older Adults*. p.pp. 3–20.

Shekhawat, G.S., Searchfield, G.D. and Stinear, C.M. (2014). Randomized trial of transcranial direct current stimulation and hearing aids for tinnitus management. *Neurorehabilitation and Neural Repair*. 28. p.pp. 410–419.

Shevlin, M., Miles, J.N.., Davies, M.N.. and Walker, S. (2000). Coefficient alpha: a useful indicator of reliability? *Personality and Individual Differences*. 28 (2). p.pp. 229–237.

Shoukri, M.M., Asyali, M.H. and Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res*. 13. p.pp. 251–271.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass Correlations : Uses in Assessing Rater Reliability. *Psychological Bulletin*. 86 (2). p.pp. 420–428.

Skevington, S.M., Lotfy, M. and O'Connell, K.A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. *Quality of Life Research*. 13. p.pp. 299–310.

Smith, E. V (2002). Detecting and evaluating the imapct of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*. 3 (2). p.pp. 205–230.

Smith, G.T. (2005). On construct validity: issues of method and measurement.

Psychological assessment. 17 (4). p.pp. 396–408.

Smith, R.M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*. 1 (2). p.pp. 199–218.

Speer, D.C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology*. 60 (2). p.pp. 402–408.

Steer, R.A., Ranieri, W.F., Beck, A.T. and Clark, D.A. (1993). Further evidence for the validity of the beck anxiety inventory with psychiatric outpatients. *Journal of Anxiety Disorders*. 7 (3). p.pp. 195–205.

Steiger, J.H. and Lind, J.C. (1980). Statistically based tests for the number of common factors. In: *Paper Presented at the Annual Meeting of the Psychometric Society*. 1980, Iowa City.

Stephens, D. (2000). A history of tinnitus. In: R. Tyler (ed.). *Tinnitus Handbook*. San Diego: Singular Publishing Group, pp. 437–448.

Stevens, C., Walker, G., Boyer, M. and Gallagher, M. (2007). Severe tinnitus and its effect on selective and divided attention. *International journal of audiology*. 46 (5). p.pp. 208–216.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.)*. 103 (2684). p.pp. 677–680.

Stouffer, J.L. and Tyler, R.S. (1990). Characterization of tinnitus by tinnitus patients. *Journal of Speech and Hearing Disorders*. 55 (August 1990). p.pp. 439–453.

Stratford, P.W. and Goldsmith, C.H. (1997). Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Physical therapy*. 77 (7). p.pp. 745–750.

Streiner, D.L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*. 80 (1). p.pp. 99–103.

Streiner, D.L. and Norman, G.R. (2008). *Health Measurement Scales : A practical guide to their development and use*. 4th Ed. Oxford: Oxford University Press.

Stucki, G., Daltroy, L., Katz, J.N., Johannesson, M. and Liang, M.H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*. 49 (7). p.pp. 711–717.

Sullivan, M., Katon, W., Russo, J., Dobie, R. and Sakai, C. (1993). A randomized trial of Nortriptyline for severe chronic tinnitus. Effects on depression, disability, and tinnitus symptoms. *Archives of Internal Medicine*. 153. p.pp. 2251–2259.

Tabachnick, B.G. and Fidell, L.S. (2013). *Using Multivariate Statistics*. 6th Ed. Boston: Pearsons.

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*. 2. p.pp. 53–55.

Tennant, A. and Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what

should one look for in a Rasch paper? *Arthritis Care and Research*. 57 (8). p.pp. 1358–1362.

Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A.W.M., Knol, D.L., Dekker, J., Bouter, L.M. and de Vet, H.C.W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*. 60 (1). p.pp. 34–42.

Terwee, C.B. and Prinsen, S. (2014). How to measure? Selecting measurement instruments for core outcome sets. In: *Core Outcome Measures in Effectiveness Trials (COMET) IV*. 2014, Rome, Italy.

Terwee, C.B., Roorda, L.D., Knol, D.L., De Boer, M.R. and De Vet, H.C.W. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology*. 62 (10). p.pp. 1062–1067.

The WHOQOL group (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine*. 28. p.pp. 551–558.

Till, A.G. (1989). *Measuring Health — a Guide To Rating Scales and Questionnaires*.

Tucker, L.R. and Lewis, C. (1973). *A reliability coefficient for maximum likelihood factor analysis**. (38). p.pp. 1–10.

Tunkel, D.E., Bauer, C.A., Sun, G.H., Rosenfeld, R.M., Chandrasekhar, S.S., Cunningham, E.R., Archer, S.M., Blakley, B.W., Carter, J.M., Granieri, E.C., Henry, J.A., Hollingsworth, D., Khan, F.A., Mitchell, S., Monfared, A., Newman, C.W., Omole, F.S., Phillips, C.D., Robinson, S.K., Taw, M.B., Tyler, R.S., Waguespack, R. and Whamond, E.J. (2014). Clinical Practice Guideline: Tinnitus. *Otolaryngology -- Head and Neck Surgery*. 151 (2 Suppl). p.pp. S1–S40.

Tyler, R.S. and Baker, L.J. (1983). Difficulties experienced by tinnitus sufferers. *Journal of Speech and Hearing Disorders*. 48 (May). p.pp. 150–154.

Tyler, R.S., Haihong J, Perreau, A., Witt, S., Noble, W. and Coelho, C. (2014). Development and validation of the Tinnitus Primary Function Questionnaire. *American Journal of Audiology*. 23. p.pp. 260–272.

Uijen, A.A., Heinst, C.W., Schellevis, F.G., van den Bosch, W.J.H.M., van de Laar, F.A., Terwee, C.B. and Schers, H.J. (2012). Measurement properties of questionnaires measuring continuity of care: A systematic review. *PLoS ONE*. 7 (7).

Uslu, R.I., Kapci, E.G., Oncu, B., Ugurlu, M. and Turkcapar, H. (2008). Psychometric properties and cut-off scores of the beck depression inventory-II in Turkish adolescents. *Journal of Clinical Psychology in Medical Settings*. 15 (3). p.pp. 225–233.

Ventry, I.M. and Weinstein, B.E. (1982). The hearing handicap inventory for the elderly: a new tool. *Ear and hearing*. 3 (3). p.pp. 128–134.

Vernon, J., Griest, S. and Press, L. (1992). Plight of unreturned tinnitus questionnaires. *British journal of Audiology*. 26. p.pp. 137–138.

Ward, M.M., Guthrie, L.C. and Alba, M. (2014). Dependence of the minimal clinically important improvement on the baseline value is a consequence of floor and ceiling effects and not different expectations by patients. *Journal of Clinical Epidemiology*. 67 (6). p.pp. 689–696.

Warne, R.T., McKyer, E.J.L. and Smith, M.L. (2012). An Introduction to Item Response Theory for Health Behavior Researchers. *American Journal of Health Behavior*. 36 (1). p.pp. 31–43.

Watts, E., Fackrell, K., Sheldrake, J. and Hoare., D.J. (in prep). Identifying problems associated with tinnitus. *Ear and hearing*.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of strength and conditioning research / National Strength & Conditioning Association*. 19 (1). p.pp. 231–240.

Westen, D. and Rosenthal, R. (2003). Quantifying construct validity: two simple measures. *Journal of personality and social psychology*. 84 (3). p.pp. 608–618.

Wilson, M.B., Kallogjeri, D., Joplin, C.N., Gorman, M.D., Krings, J.G., Lenze, E.J., Nicklaus, J.E., Spitznagel, E.E. and Piccirillo, J.F. (2015). Ecological momentary assessment of tinnitus using smartphone technology a pilot study. *Otolaryngology--Head and Neck Surgery*. Epub ahead.

Wilson, P.H., Henry, J., Bowen, M. and Haralambous, G. (1991). Tinnitus Reaction Questionnaire: psychometric properties of a measure of distress associated with tinnitus. *Journal of speech and hearing research*. 34 (1). p.pp. 197–201.

Wilson, P.H. and Henry, J.L. (1998). Tinnitus Cognitions Questionnaire : Development and psychometric properties of a measure of dysfunctional cognitions associated with tinnitus. *International Tinnitus Journal*. 4 (1). p.pp. 23–30.

Wright, B.D. (2000). Rasch model overview. *Journal of Applied Measurement*. 1 (1). p.pp. 83 – 106.

Wright, B.D. (1996). Reliability and separation. *Rasch Measurement Transactions*. 9 (4). p.p. 472.

Wright, B.D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*. 14 (2). p.pp. 97–116.

Wright, B.D. and Masters, G.N. (2002). Number of Person or Item Strata: (4*Separation + 1)/3. *Rasch Measurement Transactions*. 16 (3). p.p. 888.

Wright, J.G. and Young, N.L. (1997). A comparison of different indices of responsiveness. *Journal of Clinical Epidemiology*. 50 (3). p.pp. 239–246.

Wyrwich, K.W. (2004). Minimal important difference thresholds and the standard error of measurement: is there a connection? *Journal of biopharmaceutical statistics*. 14 (1). p.pp. 97–110.

Wyrwich, K.W., Norquist, J.M., Lenderking, W.R. and Acaster, S. (2013). Methods for interpreting change over time in patient-reported outcome measures. *Quality of Life Research*. 22 (3). p.pp. 475–483.

Wyrwich, K.W., Tierney, W.M. and Wolinsky, F.D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual

changes in health-related quality of life. *Journal of Clinical Epidemiology*. 52 (9). p.pp. 861–873.

Wyrwich, K.W., Tierney, W.M. and Wolinsky, F.D. (2002). Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Quality of Life Research*. 11 (1). p.pp. 1–7.

Yorgason, J.G., Fayad, J.N. and Kalinec, F. (2006). Understanding drug ototoxicity: molecular insights for prevention and clinical management. *Expert Opinion on Drug Safety*. 5 (3). p.pp. 383–399.

Yost, K.J. and Eton, D.T. (2005). Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Evaluation & the health professions*. 28 (2). p.pp. 172–191.

Yost, K.J., Eton, D.T., Garcia, S.F. and Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology*. 64 (5). p.pp. 507–516.

Zeman, F., Koller, M., Figueiredo, R., Aazevedo, A., Rates, M., Coelho, C., Kleinjung, T., de Ridder, D., Langguth, B. and Landgrebe, M. (2011). Tinnitus Handicap Inventory for Evaluating Treatment Effects: Which Changes Are Clinically Relevant? *Otolaryngology -- Head and Neck Surgery*. 145 (2). p.pp. 282–287.

Zeman, F., Koller, M., Langguth, B. and Landgrebe, M. (2014). Which tinnitus-related aspects are relevant for quality of life and depression: results from a large international multicentre sample. *Health and Quality of Life Outcomes*. 12 (1). p.p. 7.

Zeman, F., Koller, M., Schecklmann, M., Langguth, B. and Landgrebe, M. (2012). Tinnitus assessment by means of standardized self-report questionnaires: Psychometric properties of the Tinnitus Questionnaire (TQ), the Tinnitus Handicap Inventory (THI), and their short versions in an international and multi-lingual sample. *Health and Quality of Life Outcomes*. 10 (1). p.p. 128.

Zou, K.H., O'Malley, A.J. and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 115 (5). p.pp. 654–657.