Performance of the Tinnitus Functional Index as a diagnostic instrument in a UK clinical population

Kathryn Fackrell[1,2*], Deborah A. Hall[1,2], Johanna G. Barry [3,4] and Derek J. Hoare [1,2]

[1] NIHR Nottingham Biomedical Research Centre, Nottingham, UK

[2] Otology and hearing group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, UK

[3] Medical Research Council Institute of Hearing Research, School of Medicine, The University of Nottingham, University Park, Nottingham, NG7 2RD

[4] Nottingham University Hospitals NHS Trust, Nottingham

Author for correspondence:

Dr Kathryn Fackrell,

Address: NIHR Nottingham Biomedical Research Centre, Ropewalk House, 113 The Ropewalk, Nottingham, NG1 5DU, UK

Email: kathryn.fackrell@nottingham.ac.uk

Running title: Diagnostic performance of the Tinnitus Functional Index

Keywords: Outcome instruments; Diagnostic tool; Confirmatory Factor Analysis; Interpretability; Convergent Validity; Reliability; Tinnitus severity

ABSTRACT

Objectives

The Tinnitus Functional Index (TFI)[*] has been optimised as a diagnostic tool for quantifying the functional impact of tinnitus in US veteran and civilian groups. However, the TFI has not been evaluated for use in other English-speaking clinical populations despite its increasingly popular uptake. Here, a prospective multi-site longitudinal validation study was conducted to evaluate psychometric properties relevant to the UK clinical population. Guided by quality criteria for the measurement properties of health-related questionnaires, we specifically evaluated three diagnostic properties relating to the degree to which the TFI (i) covers the eight dimensions proposed to be important for diagnosis, (ii) reliably distinguishes individual differences in severity of tinnitus, and (iii) reliably measures the functional impact of tinnitus. We also examine whether clinically meaningful interpretations of the scores can be produced for the UK population.

Methods

Twelve National Health Service audiology clinics across the UK recruited 255 tinnitus patients to complete questionnaires at four time-intervals, from initial clinical assessment and then over a nine-month period. Patients completed the TFI, the Tinnitus Handicap Inventory (THI), tinnitus case history questions, a Global rating of Perceived Problem with tinnitus and a Clinical Global Impression of perceived change in tinnitus. Baseline TFI data were used to examine the factor structure, construct validity and interpretability of the TFI. Follow-up TFI data were used to examine reliability.

Results

Confirmatory factor analysis suggested that of the eight subscales (factors) initially established for the TFI, the ‘Auditory’ subscale did not contribute to the overall construct

[*] Abbreviations and acronyms used throughout: AUC = Area Under the receiver operator characteristic Curve; AUD = Auditory subscale; CFA = Confirmatory Factor Analysis; CFI = Comparative Fit Index; COG = Cognition subscale; EFA = Exploratory Factor Analysis; EMO = Emotional subscale; EPC = Expected Parameter Change; ICC = IntraClass Correlations; INTR = Intrusiveness subscale; LoA = Limits of Agreement; MI = Modification Index; NHS = National Health Service; QOL = Quality of life subscale; REL = Relaxation subscale; RMSEA = Root Mean Square Error of Approximation; ROC = Receiver Operator Characteristic; S-B $\chi 2$ = Satorra-Bentler scaled Chi-square; SEM = Standard Error of Measurement; SLP = Sleep subscale; SOC = Sense of control subscale; SRMR = Standardised Root Mean Square Residual; T0 = Baseline; T1 = 3 month follow up; T2 = 6 month follow up; T3 = 9 month follow up; TFI = Tinnitus Functional Index; THI = Tinnitus Handicap Inventory; THQ = Tinnitus Handicap Questionnaire; THS = Tinnitus and Hearing Survey; TLI = Tucker-Lewis Index; TQ = Tinnitus Questionnaire; TRQ = Tinnitus Reaction Questionnaire; VA = Veteran’s Affairs.

‘functional impact of tinnitus’, and a modified seven-factor model (TFI-22) better fit the variance in the patient scores. Both the global 25-item TFI and the global TFI-22 scores showed exceptionally high internal consistency ($\alpha \geq 0.95$), high construct validity with the THI ($r = 0.80$) and high test-retest reliability (ICC = 0.87). Test-retest agreement however was only deemed to be borderline acceptable (89%). Receiver Operator Characteristic analysis indicated the 25-item TFI and TFI-22 has excellent ability to distinguish between different levels of impact (Area under the curve > 0.7).

Conclusion

The TFI was confirmed to cover multiple symptom domains, measuring a multi-domain construct of tinnitus, and satisfies a range of psychometric requirements for a good clinical measure, including having excellent reliability, stability over time and sensitivity to individual differences in tinnitus severity. However, a modified seven-factor structure without the Auditory subscale (TFI-22) is recommended for calculating a global composite score for UK patients. Using patients’ experience and Receiver Operator Characteristic analysis, a grading system was presented which identifies the distinct grades of tinnitus impact in the UK clinical population that is broadly comparable to the US-based system.

1 INTRODUCTION

The experience of tinnitus involves much more than the ‘phantom’ sensation of sound since the condition can also impact on daily functioning and cause emotional distress (Henry et al., 2016; Mohamad et al., 2016; Pierzycki et al., 2016; Szczepek et al., 2014). Thus, for those who do find tinnitus bothersome, it can be described as a multi-dimensional condition. As such, it is best captured using a multi-domain patient-reported questionnaire whereby multiple items ask about particular aspects/domains of the condition which are deemed to be important (Hall et al., 2016; Henry et al., 2016). Many tinnitus questionnaires, such as the Tinnitus Questionnaire (TQ; Hallam, 2008, 1996; Hiller and Goebel, 1992), Tinnitus Handicap Inventory (THI; Newman et al., 1996), Tinnitus Reaction Questionnaire (TRQ; Wilson et al., 1991), and Tinnitus Handicap Questionnaire (THQ; Kuk et al., 1990), have known measurement properties that are consistent with their use in clinical diagnosis i.e. good discriminative power (Kamalski et al., 2010; Kirshner and Guyatt, 1985). However, in a systematic review of the psychometric properties of tinnitus questionnaires, Kamalski and colleagues (2010) did not identify or report any evidence on whether authors had provided clinically meaningful interpretations of the scores. More recently, Fackrell and colleagues (2014) reviewed the validity, reliability, responsiveness, and interpretability of tinnitus questionnaires using an internationally recognised set of criterion (Terwee et al., 2007) and reported that the evidence for the discriminative capabilities of these tinnitus questionnaires varied widely. The evidence was limited and hard to determine for content validity of the TQ, TRQ, and THI, for structural validity of the TQ, and TRQ, and for the clinical interpretation of the scores of the TQ, TRQ, and THQ (Fackrell et al., 2014). The authors concluded that, although the THQ has provided normative data, the ability to provide clinical interpretations of the scores has only been determined for the THI, with a defined established UK-based grading system. It was noted, however, that this grading system was solely based on expert opinion and the statistical properties of the scores. As such, these grades do not necessarily reflect the actual patient experience.

Importantly, the evaluation by Fackrell et al. (2014) included the Tinnitus Functional Index (TFI; (Meikle et al., 2012). First published in 2012, the TFI differs from previous tinnitus questionnaires in a number of important and positive ways; namely its careful development, comprehensive coverage of many important tinnitus complaints, interpretability of scores and responsiveness to treatment-related change (Fackrell et al., 2014). Not surprisingly, the tinnitus community at large appears eager to embrace its use. In the period 2012-2015, the TFI has established itself as the second most commonly used tinnitus questionnaire in UK National Health Service (NHS) tinnitus services; the THI is most commonly used (Hoare et al., 2015). However, it is important for our communities to appreciate that the statistical properties of the TFI are not immutable. Whilst it might be valid, reliable, and interpretable in one target population, it may behave in quite a different way in a different population (e.g. Streiner et al., 2014). As the TFI gains in international popularity in the clinic, it is important that its discriminative properties be evaluated thoroughly for each new setting and population.

It is well documented that the TFI was developed using data collected in the US, some in specialist tinnitus clinics but principally in Veteran's Affairs (VA) hospitals (58% of patients) (Meikle et al., 2012). In VA hospitals, those patients tend to be male, with an active military background, potentially experiencing a range of service-related co-morbidities, and their tinnitus is considered as a service-related condition which may entitle them to compensation. This rather unique provenance of the TFI warrants caution in terms of how well those psychometric properties transfer to different target populations.

Since the development of the TFI (Meikle et al., 2012), several evaluations of the questionnaire have been conducted in English speaking and non-English speaking countries. These evaluations increase our understanding and optimising the use of this questionnaire for research and clinical practice alike. To date, the American-English version of the TFI has been evaluated in US Veterans (Henry et al., 2016), a general clinical population in New Zealand (Chandra et al., 2014) and a research population drawn from the general public in the UK (Fackrell et al., 2016). The psychometric exploration reported by Henry et al. (2016) has the same potential limitation (not generalizable) as was noted in the original development study (Meikle et al., 2012). Fackrell et al. (2016) raised some doubts of the stability of the 8-factor structure of the TFI when used in a UK-based research population, namely that the auditory subscale appeared not to contribute to the measure of global functional impact of tinnitus. There have been four independent evaluations in different target populations, where the TFI has been translated into Dutch (Rabau et al., 2014), Swedish (Hoff and Kähäri, 2016; Müller et al., 2016), and Polish (Wrzosek et al., 2016). In general, evaluations of these translated versions showed the TFI to have good discriminative properties. However, there was also some uncertainty over its proposed factor structure. In all of those studies, Exploratory Factor Analysis (EFA) was conducted which identified different patterns in the data, typically with only five or six factors initially identified, although all reported forced eight-factor models as being satisfactory (Rabau et al., 2014; Hoff and Kähäri, 2016; Müller et al., 2016; Chandra et al., 2014). Only the Polish study included Confirmatory Factor Analysis (CFA) to test the proposed eight-factor structure, finding it to be unsatisfactory (Wrzosek et al., 2016). Instead, their EFA indicated that a five-factor solution best explained the Polish population data. Interpretability was not assessed in any of those studies.

Meikle and colleagues (2012) have proposed interim grading systems for the TFI, but the question of whether this interpretability of the global scores, an essential requirement for the suitability of a questionnaire in clinical practice or research, is transferable to other populations is yet to be addressed in any subsequent psychometric evaluation.

In the present study, we examined the psychometric properties of the TFI for a large clinical sample of UK NHS patients treated for tinnitus. In designing this study we were guided by quality criteria for the measurement properties of health-related questionnaires as outlined by Mokkink et al. (2012) and Terwee et al. (2007). Unlike our previous work (Fackrell et al., 2016), this study was specifically designed to evaluate the TFI as a reliable and valid measure of tinnitus severity for use in a tinnitus clinical population, and to determine its responsiveness and interpretability. This study is particularly important because it is based on a study sample drawn from a general (i.e. non-military) help-seeking clinical population.

The aims of the study were to evaluate the degree to which the TFI:

i) covered the proposed eight important dimensions of tinnitus-related impact,
ii) reliably distinguished one patient from another,
iii) reliably measured the impact of tinnitus,
iv) produced a grading scheme that can give a meaningful diagnostic interpretation to the UK clinical population

2 MATERIALS AND METHODS

This was a prospective multi-site, repeated-measures validation study. Ethical approval was granted by Cornwall and Plymouth Research Ethics Committee (13/SW/0234), and Nottingham University Hospitals NHS Trust was Sponsor.

2.1 *Eligibility*

Patients (≥18 years old) were attending their first appointment with an audiologist and reporting persistent tinnitus. The inclusion criterion referred to those patients whom had not been treated for tinnitus or attended a tinnitus clinic in the previous 6 months. In addition, patients required sufficient command of English language to independently complete questionnaires.

2.2 *Recruiting sites*

Twelve NHS audiology clinics served as recruitment sites (Supplementary Table 1). At each site, a single member of staff from the clinical care team was responsible for identifying patients, consenting, and collecting the questionnaire data at the initial appointment. Patients were recruited from October 2013 to June 2014. Recruitment activities stopped when the target sample set *a priori* (see below) of 250 patients were recruited to the study. Five additional patients were enrolled because they had received invitations to participate before this date, and returned completed questionnaire packs to their initial (i.e. diagnostic and enrolment) appointment.

2.3 Sample size

To reliably assess the structure of the TFI using CFA of the baseline (T0) data, it is recommended that the sample size is > 200 (MacCallum et al., 1996). Using a ratio of 5:1 individuals per estimator parameter (Floyd and Widaman, 1995; Nunnally, 1978; Schreiber et al., 2006), a sample size of 290 patients would be required for the TFI model (53 estimated parameters). However, the large degrees of freedom for the TFI model (df 267) indicate that a sample size of 250 patients would provide sufficient power to effectively test model fit and allow for missing data (MacCallum et al., 1996). In general, for reliability analyses with follow-up data, a sample size of ≥ 50 is recommended for each element of the analysis. A dropout rate of approximately 38% was estimated for data collection at follow-up (based on Vernon et al. (1992)). At this rate, a starting sample of 250 patients would yield sufficient data to conduct the reliability analyses planned.

2.4 *Data collection schedule*

The full study involved completing the TFI and additional questionnaires on four separate occasions over a 9-month period. This could be done either at home or in a location of the patient’s choice (Supplementary Figure 1). On the first occasion (T0), patients completed the questionnaires before, or immediately after their initial appointment for diagnostic assessment. The questionnaires included a 10-item case history questionnaire requesting information about age, gender, tinnitus duration, its characteristics, and duration, and any self-reported hearing difficulty. Patients returned the first completed pack directly to the clinic as familiarity with clinical staff has been shown to increase compliance and return rate (Edwards et al., 2009, 2002). Follow-up (T1-T3) was conducted at three-month intervals from the initial appointment by mailing questionnaire packs to patients with prepaid return envelopes. Packs were mailed two weeks before their completion due date. Where questionnaires were not returned, reminders were issued after two weeks and again after a further week.

2.5 Measures

Tinnitus Functional Index The TFI measures the functional impact of tinnitus using 25 items, each rated on an 11-point Likert scale with descriptors at either end of the scale (Meikle et al., 2012). Patients rated each item according to how they have felt over the past week. The procedure for scoring the TFI followed the instructions provided by Meikle et al. (2012). The global score reflects the sum of all scores, divided by 2.5 to give a global score out of 100. Higher scores indicate the greater impact on everyday functioning. The TFI encompasses eight subscales; (i) Intrusiveness (INTR 1 - 3), (ii) Sense of control (SOC 4 -6), (iii) Cognition (COG 7 - 9), (iv) Sleep (SLP 10 - 12), (v) Auditory (AUD 13 - 15, (vi) Relaxation (REL 16 - 18), (vii) Quality of life (QOL 19 - 22), and (viii) Emotional distress (EMO 23 - 25). Each subscale can be scored separately, whereby the relevant three or four items are summed and weighted to give a score out of 100. The TFI was completed at T0, T1, T2, and T3 and scores from each were considered in the analyses reported here. As the TFI has an 11-point response scale, the data were treated as continuous rather than categorical (Muthén and Muthén, 2012). Throughout reference to the TFI refers to the 25-item questionnaire.

Tinnitus Handicap Inventory The Tinnitus Handicap Inventory (THI; Newman et al., 1996; Newman et al., 1998) measures tinnitus-related psychological distress using 25 items each rated on a categorical 3-point scale (4 = yes, 2 = sometimes, 0 = no). The mean global score reflects the sum of all responses with a maximum score of 100 indicating the greatest distress due to tinnitus. Newman et al. (1996) did not provide any guidelines on how to account for missing values in the calculation of the total score and so a decision was made to only calculate the global score if the respondent had missed 3 items or fewer. The global THI score are classified based on THI severity grading system (slight; mild; moderate; severe; catastrophic) (McCombe et al., 2001; Newman et al., 1996). The THI was completed at T0, T1, T2, and T3 and scores from each were considered in the construct validity and internal consistency analyses reported here.

Global rating of Perceived Problem with tinnitus (Perceived Problem rating) To develop a better understanding of what the global TFI scores mean to patients, each patient completed a single question at T0 asking "How much of a problem is your tinnitus?" There were five possible response options; 1 = not a problem, 2 = a small problem, 3 = a moderate problem, 4 = a big problem, and 5 = a very big problem.

Clinical Global Impression of perceived change in tinnitus (Clinical Global Impression) At each follow-up assessment, patients answered one question about the extent to which their tinnitus changed: "All things considered, how is your overall tinnitus condition now, compared to x months ago?", where x = 3, 6 and 9 months at T1, T2, and T3 respectively. Responses were made on a 7-point scale (3 = much improved, 2 = moderately improved, 1 = slightly improved, 0 = no change, -1 =slightly worse, -2 = moderately worse to -3 = much worse). Reliability analyses using "no change" subgroup reported here considered the Clinical Global Impression scores at T1 and T2. We had planned to use the T3 data in this analysis, but the 'no change' subgroup at T3 was too small.

3 ANALYSIS METHODS

The methodological approach taken here was underpinned by Classical Test theory principles, in which a person's "true score" is directly unobservable. Every observed score is assumed to be made up of measurement error and the person actual “true” attitude or attribute on the latent construct that is being measured, in this case tinnitus (Raykov and Marcoulides, 2011). The criteria for acceptable psychometric properties described below were guided by established frameworks to evaluate questionnaires (Mokkink et al., 2012; Terwee et al., 2007), in particular psychometric properties, validity, reliability, and interpretability were examined here. CFA was performed in Mplus 7 (Muthén and Muthén, 2012), while reliability and interpretability analyses were calculated in SPSS v.21.0 (IBM Corp., 2012) and Microsoft Excel. The TFI subscales have been proposed as having potential to be used as standalone measures, therefore when possible the subscales were also been subject to validity and reliability analysis.

3.1 ***Proposed eight-factor structure of the TFI***

CFA was conducted on TFI data collected at T0. Model specification was based on the original description of the structure with 25 items, eight subscales, and a composite global score (Meikle et al., 2012). The original eight-factor TFI model (Figure 1) was defined as follows: (i) eight first-order latent constructs (factors) corresponding to the TFI subscales and one second-order latent construct corresponding to the composite global score; (ii) the observed variables (i.e. the 25 TFI items) were fixed to their original TFI factor and constrained to zero loadings on the other factors in the questionnaire; (iii) the error variance (residual variance) associated with each observed variable was constrained to zero, assumed to be uncorrelated with the error variance of any other variable and random; (iv) variance in the first-order factors was assumed to be completely explained by the relationship to the second-order factor. Therefore the second-order factor variance was fixed at 1.

** Figure 1 **

For an extended description of the methodology see Fackrell et al. (2016). In brief, to adjust for non-normality in the distribution of the data (Mahalanobis d-squared: 81.5 to 55.0, $p < 0.001$), the model was estimated using maximum likelihood parameter estimation adjusted with a Satorra-Bentler scaled Chi-square (S-B χ^2; Satorra and Bentler, 1994) to ensure robust standard errors for parameter estimates and goodness of fit indices (Bentler, 2007, 2006; Hu and Bentler, 1999). Initially, the eight-factor model was estimated without the second-order factor, allowing for examination of covariance between first-order factors. The first-order factors are purported to be measuring the same underlying construct (i.e. second-order factor the functional impact of tinnitus), and therefore a degree of overlap in content is expected (between > 0.30 - < 0.85). Following this, the model was re-specified to include the second-order factor.

Goodness of fit was determined using the absolute fit indices S-B χ^2 (Satorra and Bentler, 1994) and Standardised Root Mean Square Residual (SRMR; (Bentler, 2006; Hu and Bentler, 1998) and approximation fit indices, the Tucker-Lewis Index (TLI; Tucker and Lewis, 1973), Comparative Fit Index (CFI; Bentler, 1990), Root Mean Square Error of Approximation (RMSEA; Steiger and Lind, 1980) and confidence intervals (CIs). In the event that the model was a less than optimal fit to the data, factor loading estimates, the Modification Index (MI) and Expected Parameter Change (EPC) were examined to identify any misspecifications in the parameters that might be adjusted to improve model fit (Brown and Moore, 2012; MacCallum et al., 1992).

3.2 *Validity of the TFI construct*

Convergent validity was assessed as Pearson bivariate correlations comparing the global TFI and subscale scores with THI global scores collected at T0 and all three follow ups (T1 – T3). The global TFI was assumed to measure a similar construct to the THI and so it was predicted to have high convergent validity (correlation > 0.60). There was no evidence of skewness or kurtosis in the data distribution. Pairwise deletion was conducted to ensure the largest possible sample sizes.

3.3 *Internal consistency of the TFI 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 (Clark and Watson, 1995; Cortina, 1993; Cronbach, 1951). Cronbach’s alpha was calculated on complete data from T0 and all three follow up (T1 – T3), with values between 0.7 and 0.95, desirably below 0.9, taken to indicate acceptable internal consistency (listwise-deletion is automatically conducted) (Peterson, 1994; Terwee et al., 2007).

3.4 *Reliability of the TFI*

Distinguish one patient from another

Reliability indicates the degree to which individuals who are different can be distinguished from each other, despite measurement error as assessed variation in test-retest situations (Mokkink et al., 2012; Terwee et al., 2007). Reliability was assessed using the measurement variance for the same individuals between TFI scores at T0 and follow-up, (for example T0 and T1). IntraClass Correlations (ICC) were computed for global TFI and subscale scores from a subset of patients that reported ‘no change’ on the Clinical Global Impression at T1 and T2. Patients who identified themselves as having ‘no change’ in tinnitus severity on the Clinical Global Impression but had changes in TFI scores of above 70 were considered outliers, as large change scores such as this would correspond to a change from severe tinnitus to mild tinnitus or vice versa. ICCs provide an estimate of the ratio of all variances ranging from 0 (no reliability) to 1 (perfect reliability), with 0.4 to 0.69 being acceptable, and > 0.70 being excellent (Terwee et al., 2007).

Stability across time, accounting for measurement error

Measurement error refers to the difference between an observed score and its true value (Mokkink et al., 2012). Appropriate statistics for assessing measurement error and the stability (precision of measurement) are Limits of Agreement (LoA) and the Standard Error of Measurement (SEM; SD*diff*/√2). The same subset of patients that reported ‘no change’ on the Clinical Global Impression at follow up were selected for these analyses as well. 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 (SD*diff*). The SD*diff* was then used to calculate the LoA. The Bland–Altman method (1986) for LoA calculates the mean difference in scores between two repeated visits (the 'bias'), and 95% LoA (Mean difference ± 1.96 × SD*diff*). The assumption is that if there is 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 data points would be within ±2 standard deviations of the mean difference.

3.5 *Diagnostic interpretation of global TFI scores*

To provide clinical meaning to the global TFI scores and identify diagnostic grades of symptom severity, the TFI scores from T0 data were assessed using anchor-based approaches (Perceived Problem rating categories and the THI grading system; (Crosby et al., 2003; de Vet et al., 2011) and were then subjected to Receiver Operator Characteristic (ROC) analysis (Eng, 2005). To ascertain the strength of the relationship between the TFI scores and anchor-based approaches, Spearman's rank correlation coefficient were calculated. Individual global TFI scores were then stratified according to the five Perceived Problem rating categories and the THI grading categories and distributions were visually examined and compared using general linear modelling. The ROC curve analysis combined information on sensitivity (true positive rate) and specificity (true negative rate) to detect the threshold value that best discriminates between the patients in adjacent categories of severity (Eng, 2005). The choice of anchor is crucial and determines whether the grading categories are considered from the patient perspective, questionnaire developer, or clinician (de'Vet et al., 2011, 2007). For this analysis, Perceived Problem rating categories were used as the gold standard anchor to assign patients into distinct grades that *do* have qualitative meaning related to patient experience (Crosby et al., 2003; Hays and Woolley, 2000; Revicki et al., 2008; Yost and Eton, 2005). Although priority was placed on the Perceived Problem ratings, the THI gradings were used to inform and guide any large conflicts in classification of the TFI score between their Perceived Problem rating (i.e. identifying a 'very big problem') and THI grading (mild problem) to ensure the final categories were clearly defined. In these cases, the TFI score classification was adjusted based on the patients score on the TFI and THI grading.

Sensitivity was equivalent to the probability that patients were correctly classified according to their TFI score as experiencing one of the problem categories, e.g. a "moderate problem" with tinnitus (positive cases), whilst specificity refers to the probability that patients were correctly classified as experiencing the adjacent lower problem category, e.g. a "small problem" (negative cases) (Eng, 2005; Uslu et al., 2008). ROC curve plots the sensitivity (y axis) vs 1 – specificity (x axis) with the Area Under the ROC Curve (AUC) representing the TFI's ability to discriminate between people who experience tinnitus as a "small problem" from "no problem". For example, an AUC = 0.5 denotes a 50% probability that the TFI would be unable to identify individuals with a small problem from those who do not. A more prominent curve is therefore equivalent to a more accurate test, and AUC values of above 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 as the cut-off value for the range in each diagnostic category. Traditionally the balance between sensitivity and specificity is employed to identify the optimal threshold. However, since it is more important as a diagnostic tool to identify the greater tinnitus symptomatology, sensitivity is prioritised above specificity for the optimal threshold. For each set of adjacent diagnostic categories separate ROC curves were calculated (for example, “small problem” versus “moderate problem” and “big problem” versus “very big problem”).

4 RESULTS

4.1 *Patient characteristics*

A total of 255 tinnitus patients (male: 149 (59%), female: 105 (41%)) were enrolled and completed T0 measurements. The average age was 53.6 years (SD = 13.4) with a range of 18 to 84 years. Just under 50% of patients 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. According to the Perceived Problem rating, almost half of patients described themselves as having a moderate problem with tinnitus. More than 70% of patients self-reported having problems hearing speech or other sounds. Of this number, over 40% identified having a moderate to big problems hearing speech or other sounds (Supplementary Table 2). According to the Clinical Global Impression rating, over 35% of patients reported that their tinnitus had improved at T1 – T3, less than 15% reported their tinnitus had “worsened” at T1, increasing to 35% at T3, and 50% of patients reported “no change” to their tinnitus at T1, with numbers reporting decreasing to less than 30% at T3. Descriptive statistics for the TFI global and subscales scores and the THI global score from T0 to T3 are presented in Supplementary Table 3.

Missing T0 data was less than 7% and was identified as Missing Completely At Random. Only T0 data with fully completed TFI scores on all 25 items were used for the CFA and so after list-wise deletion, this effective sample size was 239. For 13 patients, TFI data was missing for one question item and for three patients it was missing for two questions. Participant characteristics and distributions reported for the total sample were reflected in this CFA sample (Supplementary Table 2). Two patients did not complete the

Perceived Problem rating and so the effective sample size available for interpretability analysis was 253. Compliance exceeded the expected rate of 64%. At T1, 198 (78%) completed follow-up questionnaires, at T2 it was 176 (69%) and at T3 it was 166 (65%).

4.2 Validity of the eight-factor structure of the TFI

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. Notably, the Auditory (AUD) factor showed unacceptably weak correlations (< 0.3) with three of the other factors (Supplementary Table 4).

The fit indices for the original TFI model (Figure 1) were all borderline, indicating that the fit of the data was less than optimal (Table 1). The S-Bχ2 was significantly large (χ2: 577.5; p < 0.001) and the S-Bχ2-df was marginally larger (2.2) than the critical ratio cut-off (≤ 2.0) indicating problems with data fit. Consistent with this, the RMSEA score (0.07) was less than optimal (≤ 0.05). The SRMR however was just within reasonable fit criteria (≤ 0.07, ideally it should be ≤ 0.06) and both the TLI and CFI estimates indicated acceptable model fit (> 0.90).

With respect to the standardised parameter estimates, both Auditory and Sleep factors had loading estimates below the optimal value, although the Sleep factor was only marginally below (0.68) (Figure 1; Supplementary Table 5). Consistent with this, squared factor loadings revealed that the second-order factor accounted for less than 46% of the variance in the Sleep factor, and only 25% in the Auditory factor. The Auditory factor had very weak associations with the second-order factor and the other seven factors and as a consequence it makes considerably less contribution to the second-order construct.

** Table 1**

A high degree of parameter misspecification was associated with the Auditory factor. 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). This error covariance may reflect or be inflated by mis-specified error between items; INTR1 and INTR2 (MI: 16.46; standardised EPC: 0.38), REL19 and REL20 (MI: 21.43; standardised EPC: 0.45) and

EMO23 and EMO24 (MI: 11.29; standardised EPC: 1.01). The model was re-specified adjusting for the error covariance between these items (Supplemental Figure 2). This did not improve the MIs associated with the Auditory factor and the other factors (MI range: 10.7 to 44.4) and the model fit remained less than optimal (Table 1). Consequently, the Auditory factor was removed from the second-order structure.

The statistical properties of a modified 22-item seven-factor model (TFI-22, Figure 2) were examined (Auditory factor removed). This TFI-22 model was a much-improved fit to the data on all relevant statistics and although the RMSEA score still exceeded 0.05, when considered alongside SRMR ≤ 0.06, a RMSEA score of 0.06 indicates reasonable fit (Table 1). Standardised parameter estimates and squared factor loadings were comparable to the original 25-item TFI model (Figure 2; Supplementary Table 5). This confirms that the Auditory factor should not be included when calculating the composite score.

***Figure 2**

4.3 Validity of the TFI and TFI-22 construct

Convergent validity of the TFI was acceptable; TFI global scores consistently showed strong positive correlations with the THI global scores (r > 0.80). For the TFI subscales, weak (r = 0.41) to strong (r = 0.86) positive correlations were observed with the THI global scores, with the weakest correlation with Auditory subscale (Supplementary Table 3). Comparably strong correlations were also observed for the global TFI-22 and THI (r > 0.80).

4.4 Internal consistency of the TFI and TFI-22 structure

Internal consistency of global TFI and THI scores was extremely high (α > 0.95) for data from T0 and all three follow-ups, indicating overlap in content. Likewise, the estimates for the TFI subscales were extremely high with only the Intrusiveness and Sense of Control subscales consistently within the recommended criteria (Supplementary Table 3). The global TFI-22 also showed extremely high internal consistency (α > 0.96).

4.5 Reliability of the TFI and TFI-22

At T1, 101 patients reported 'no change' in their overall perception of tinnitus using the Clinical Global Impression, and of this subgroup at T2 only 51 patients still reported 'no change'. The 'no change' subgroup at T3 was too small for appropriate analysis (n = 29). Based on our *a priori* criteria, data from one patient for the TFI global and subscales change scores were removed as outliers. There were no missing data for the TFI global score, therefore the effective sample size was 50. For the subscales, data from four patients (excluding the patient mentioned above) were removed as outliers, one from the Cognitive subscale, one from QOL subscale, and two from the Sleep subscale (missing data reported in Table 2).

Participant characteristics for the 50 participants were representative of the total sample (Supplementary Table 2). The reported average age was 57 years, and the reported distributions of gender, duration of tinnitus, tinnitus severity, and hearing difficulties were similar to those observed for the total sample and the sample used for CFA. Table 2 shows the results of analyses that compared 'test' as T0 and 'retest' as a pooled set of T1 and T2 data for TFI global and subscale scores.

** Table 2 **

Distinguishing one patient from another

The ICC for the TFI global score was 0.87 (95% CI: 0.80 – 0.93), indicating excellent reliability (Table 2). Subscale scores showed similarly acceptable reliability with ICCs ranging 0.69 to 0.86, although for some subscales the 95% CIs indicated larger variability and lower reliability than the ICC estimates imply. The only reliability estimate below the recommended guidelines was for the Sleep subscale (0.69). Although, the estimate is only marginally below, the large CIs indicate that in a random sample the reliability could be markedly lower or within the recommended criteria. The ICC for the global TFI-22 score was 0.92 (95% CI: 0.87 – 0.96), again indicating excellent reliability.

Stability across time, accounting for measurement error

Whilst, the SEM estimate for the TFI global scores is minimal at 5.1 out of a possible 100, the LoA estimates for the TFI global was 14.2 (± Mean diff of -5.4) and only 88% of the data fell within the LoA (Table 2; Supplementary Figure 3). This indicates that the TFI is susceptible to some imprecision in the measurement, slightly reducing the reliability. LoA estimates were typically larger for the TFI subscales than for the global score, ranging from 22 to 32 points, and with some degree of imprecision as shown by the findings that < 95% of the data fell within +/- 2 SD of the mean difference. For the TFI-22, the LoA was 13.9 (± mean diff of -5.9), but again only 88% of the data points fell within the LoA indicting some degree of imprecision in the measure (Supplementary Figure 3).

4.6 Diagnostic interpretation of global TFI scores

The Perceived Problem rating distributions are reported in Supplementary Table 2. No patients reported that their tinnitus was not a problem and so the “no problem” category was not used in the analysis (Supplementary Table 2). Consequently, we had four categories of problem in our population and so chose to use the THI severity grading system with the four grades (slight = 0 – 16; mild = 18 – 36; moderate = 38 – 56; severe = 58 – 100; Newman et al., 1998). This allowed us to directly compare distribution and grading across Perceived Problem rating, THI and TFI. Spearman’s correlation coefficients comparing the TFI global scores with the categorical data for the four Perceived Problem rating categories and four THI grades indicate a strong positive relationship between the scores and categories (Spearman’s rho = 0.8, in both cases).

Using individual global TFI scores as the dependent variable, General Linear Modelling showed a significant main effect of problem category (F (3, 253) = 6.78, p < 0.001), with significant differences between each category, except between “big problem” and “very big problem” (p > 0.25). There was also a significant main effect of THI grading (F (3, 253) = 26.02, p < 0.001) with significant differences between each category. However, no significant differences were observed between the two categorising methods (F (7, 253) = 0.37, p = 0.92). The mean scores within each category were similar across the different approaches (Supplemental Figure 4).

For the ROC analysis, the adjusted Perceived Problem rating categories were used as per *a priori* criteria. Three ROC analyses compared TFI scores within the adjusted Perceived

Problem rating category to those within the category just below. So the “very big problem” category (n = 57) compared with the “big problem” category (n = 49), “big problem” compared with “moderate problem” (n = 107), and “moderate problem” compared with “small problem” (n = 42) (Figure 3). The AUC in all three comparisons was ≥ 0.85, exceeding the recommended criteria of > 0.7. This indicates excellent ability to discriminate patients reporting different levels of perceived problems. The sensitivity and specificity rates were plotted for multiple possible cut-off points for each analysis (Figure 3).

** Figure 3 **

Examination of the ROC curve, and the estimate cut-off values for detecting patients with “moderate problems” from those with “small problems” (Figure 3; Supplemental Table 6), indicated that a cut-off value of 28 approximates the optimal cut-off value that was sensitive to discriminating moderate problems (94%) from small problems (60%). Therefore, global TFI scores below 28 indicate a small problem with tinnitus. The estimate cut-off values for detecting “big problems” from “moderate problems”, and the corresponding ROC curve, indicate that a cut-off value of 47 points is optimal for discriminating patients who have big problems from those with moderate problems (Figure 3; Supplemental Table 7). Moderate problems with tinnitus are therefore identified by global TFI scores in the range of 28 and 46. To discriminate patients reporting “very big problems” from those reporting “big problems” an optimal cut-off value of 65 points was identified as correctly classifying 93% of patients as having very big problems and 60% as having big problems (Figure 3; Supplemental Table 8). The grading system generated from these findings is given in Table 3.

** Table 3 **

For the TFI-22, ROC analysis revealed that it had an excellent ability to discriminate patients reporting different levels of perceived problems, with AUC estimates (AUC > 0.84) exceeding the recommended criteria (AUC > 0.7; Figure 3). Optimal cut-off values for discriminating patients were estimated and were similar to those identified for TFI, varying only by a couple of points (Table 3).

5 DISCUSSION

The current study provides the first independent and comprehensive psychometric evaluation testing the diagnostic utility of the TFI in a UK clinical population, building on our previous psychometric evaluation in a UK tinnitus research volunteer population (Fackrell et al., 2016). Notably, we conclude for a UK clinical population that although the TFI proposed by Meikle et al. (2012) generally produced a reliable diagnostic tool with good discriminative properties, and good convergent validity with the THI, the original eight-factor structure was not confirmed. Instead, a modified 22-item seven-factor structure best explained the data captured in our UK clinical population, and this 22-item version also performed well on all other psychometric properties.

5.1 The Auditory domain is theoretically distinct from the functional impact of tinnitus measured by the remaining items

The original eight-factor structure proposed is not the best possible explanation for the UK clinical population data. The Auditory factor was unrelated to the underlying construct of the functional impact of tinnitus and consequently was removed to create a modified TFI structure with seven-factors (TFI-22). Including items that do not fit within the second-order construct risks unduly diluting the specificity of the composite score for the functional impact of tinnitus. Meikle et al. (2012) envisaged this possibility and suggested that "its [the Auditory subscale] underlying dimension may be of a different flavour compared with the other seven subscales" (p. 21) and that it could represent "an underlying specific factor" (p. 20). This seems to be case here for a UK clinical population.

The most likely explanation for this is because tinnitus is often co-morbid with hearing loss (Hoare et al., 2014). Our population reflected this with the majority self-reporting some degree of hearing difficulties. Some people attribute their hearing difficulties solely to tinnitus such that it is difficult to disentangle what hearing difficulty is related specifically to tinnitus and not hearing loss (Ratnayake et al., 2009). Given the nature of questions in tinnitus questionnaires, they can be susceptible to inaccuracies in measuring hearing difficulties specific to tinnitus (Kuk et al., 1990; Newman et al., 1998; Ratnayake et al., 2009). The Tinnitus and Hearing Survey (THS; Henry et al., 2014) was specifically developed to disambiguate difficulties related to hearing from those related to tinnitus and includes two subscales. The first asks about tinnitus problems that are unrelated to hearing difficulties and the second asks 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?”. Consequently, whilst the Auditory subscale should not be included in the calculation for the global TFI score in the UK, it could be used to aid clinical interpretation.

5.2 The TFI shows acceptable discriminative properties

All reported reliability estimates for the global TFI, here and in previous evaluations (Chandra et al., 2014; Fackrell et al., 2016; Hoff and Kähäri, 2016; Müller et al., 2016), have been shown to be considerably higher than the estimates reported in the original TFI development (Meikle et al., 2012). These results strengthen the conclusions originally made by the authors (Meikle et al., 2012). The TFI can therefore consistently and reliably distinguish one patient from another in a range of populations, with varying degrees of tinnitus severity and duration and in general, the same conclusions can be made about the subscales.

Conversely, the stability of the measure showed more susceptibility to larger degrees of measurement error in a patients scores (agreement below 95%) which cannot be attributed to true changes in tinnitus impact over long time intervals. The estimates reported here for the global TFI and subscales are lower than those reported previously (Fackrell et al., 2016). For this study, we were unable to conduct a traditional 2-3-week test-retest period due to variability in clinical appointment booking procedures. These estimates, based on variance observed in scores over 6 months, could have inflated the measurement error observed here by introducing additional error associated with memory recall. So far, no other studies have reported estimates for agreement for the TFI and as such we do not know whether agreement estimates are consistent across populations; further estimates are indicated.

Fundamentally, reliability and agreement tests provide different information and consequently conflicting results where, on the one hand, there is excellent reliability and on the other, large measurement error (Kottner and Streiner, 2011). Whilst reliability is interested in the variability of individual scores in comparison to the overall, the agreement is focused on the similarity between the scores over time, with the expectation of very little between subject variability. Therefore, if all patient scores were in complete agreement at 95%, evaluative properties might be excellent but if there was additionally little variability in scores, the discriminative properties would be reduced and the questionnaire would be deemed unreliable as a diagnostic tool. This highlights the contradictory nature of encompassing both discriminative and evaluative properties in a single measurement tool (Guyatt et al., 1987; Kirshner and Guyatt, 1985; Meikle et al., 2007). Yet these tests are recommended to be conducted. To overcome this conflict, it has been suggested that, although the assumption is that 95% data should fall within the limits of agreement, the degree in which those limits can vary and still be considered acceptable has not been established (Giavarina, 2015). There is possibly a need to be less rigid with this criterion. For the limits to be deemed acceptable, they should be based on the intended use of the measurement tool, i.e. clinical requirements and considerations and defined *a priori* (Giavarina, 2015). Therefore, considering that the TFI is intended to be used as both a diagnostic tool and outcome measure and was designed with both these properties in mind, the level of agreement observed here (88%) is reasonably high and deemed acceptable.

5.3 High internal consistency for the global TFI and all but two subscales indicates redundancy

Our findings indicate that the global TFI and ***most*** subscales had high internal consistency above the desirable and acceptable criteria. These findings have been observed in the development (Meikle et al. (2012) and subsequent evaluations of the TFI (Chandra et al., 2014; Fackrell et al., 2016; Müller et al., 2016; Rabau et al., 2014; Wrzosek et al., 2016). Cronbach's alpha should be interpreted with caution because the estimates could be inflated due to the presence of more than one underlying trait being measured or the heterogeneity of the population (Cortina, 1993; Kottner and Streiner, 2010; Shevlin et al., 2000). However, given that the subscales, which are proposed as unidimensional structures, also presented with high internal consistency, it does suggest that the subscales may not be a multi-item measure of the construct and that highly correlated items may be redundant, within the subscale and global TFI (Clark and Watson, 1995; Streiner and Norman, 2008). It could be proposed that this indicates a shorter 8-item version of the TFI could be created, with one item from each subscale.

However, although we do recognise the possibility of redundancy, removing items would dramatically reduce the reliability and utility of the global TFI and the subscales would no longer exist. The TFI was intended as a reliable diagnostic tool and outcome measure, with the ability to separately evaluate some important aspects (domains) of tinnitus to aid researchers and clinicians. Removing multiple items would also reduce its utility as an outcome measure (Clark and Watson, 1995; Guyatt et al., 1992). In terms of the subscales, a number of items are needed within the scale to sufficiently conceptualise the underlying construct that it is aimed to measure and ensure high reliability, (Clark and Watson, 1995; Hair et al., 2009; Raubenheimer, 2004). Reliability and responsiveness can adequately be achieved with three items or more***;*** any less and reliability estimates are more susceptible to error (Yong and Pearce, 2013).

5.4 A newly revised diagnostic grading for the TFI instrument in the UK

The developers of the TFI have published two grading systems (Henry et al., 2016; Meikle et al., 2012), summarised in Table 3. Here, we used an anchor-based method of patient Perceived Problem rating followed by ROC analysis to determine the threshold value that best discriminates between the patients in adjacent categories of severity. We prioritised a threshold value that would easily identify patients with the higher level of problem with their tinnitus. The TFI and TFI-22 showed excellent ability to discriminate patients reporting different levels of perceived problems. Compared to the proposed grading system (Henry et al., 2016; Meikle et al., 2012), the criterion range for each grade identified here are slightly different. In particular, in our sample, no patients reported tinnitus as "no problem" so therefore we can only provide a speculative range for this category based on the lower range of scores that were not identified by patients reporting a "small problem" with their tinnitus. Other than the "small problem" category, the score ranges in other categories are reasonably similar to those proposed (Henry et al., 2016; Meikle et al., 2012). Although we gathered data on patient experiences through the use of a closed question, patients' interpretations of the descriptors were not examined here nor in the development of the TFI. The inclusion of patient experience and confirmation of the ability of the TFI to discriminate patient with different levels of tinnitus problem reported here means that there is greater confidence in the reliability of these grades. Therefore, we recommend our grading system to be adopted for use in UK clinical practice and re-evaluated for use in research. We did not collect the type of qualitative data that could subsequently be used to inform clinical decision making relevant to scores or categories. It would also be of value to establish specific clinical meaning to the grades in a further study using qualitative methods such as focus groups or semi-structured interviews with a patient population.

5.5 Conclusions and recommendations

The TFI global score was shown to reliably distinguish one patient from another and discriminates different levels of tinnitus. However, based on our analyses of a large UK clinical population, we would recommend the modified seven-subscale TFI-22 for diagnostic purposes in the UK with a revised grading scale. Whilst the Auditory subscale is theoretically distinct from the other subscales, it can nevertheless provide clinically valuable information about the degree of hearing difficulty attributed to tinnitus and so we do not suggest removing it from the questionnaire but merely scoring the composite TFI-22 differently from the US-based TFI original. Further in-depth evaluations of the TFI subscales are warranted to examine their reliability as standalone measures.

6 ACKNOWLEDGEMENTS

The authors would like to thank the following NHS clinical facilities and individual clinicians, who worked tirelessly to recruit all the patients to the study: Tony Kay, Audiology services, Aintree University Hospital NHS Foundation Trust (Liverpool), Mary Mitchell, Hearing, Tinnitus & Balance Rehabilitation Service, Belfast Health and Social Care Trust (Belfast), Saskia Harden, Audiology, Brighton & Sussex University Hospitals NHS Trust (Brighton), Rachel Knappett and David Baguley, Audiology, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust (Cambridge), Deepak Rajenderkumar, Welsh Hearing Institute, University Hospital of Wales, Cardiff & Vale University Health Board (Cardiff), Danny Kearney, Audiology, Manchester Royal Infirmary, Central Manchester University Hospitals NHS Foundation Trust (Manchester), Jacqueline Jones, Audiology, Countess of Chester Hospital NHS Foundation Trust (Chester), Sandra Grimes, Hearing Rehabilitation Centre, Doncaster and Bassetlaw Hospitals NHS Foundation Trust (Doncaster), Pam Comiskey, Tinnitus and Balance Services, NHS Fife (Kirkcaldy), Julie Brady, Nottingham Audiology Services, Nottingham University Hospitals NHS Trust (Nottingham), Claire Gatenby, Audiology, Norfolk and Norwich University Hospitals NHS

Foundation Trust (Norwich), Michelle Booth, Kings Mill Hospital, Sherwood Forest Hospital NHS Foundation Trust (Mansfield).

This report is independent research by the National Institute for Health Research Biomedical Research Centre Funding Scheme. 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.

7 REFERENCES

Bentler, P.M., 2007. On tests and indices for evaluating structural models. Pers. Individ. Dif. 42, 825–829. doi:10.1016/j.paid.2006.09.024

Bentler, P.M., 2006. EQS 6 structural equations program manual, Los Angeles: BMDP Statistic Software.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychol. Bull. 107, 238–246.

Brown, T.A., Moore, M.T., 2012. Confirmatory factor analysis, in: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. The Guilford Press, New York, pp. 361–379.

Chandra, N., Lee, A., Searchfield, G., 2014. Validation of the Tinnitus Functional Index in New Zealand, in: 8th International TRI Tinnitus Conference. Auckland, pp. 31–32.

Clark, L.A., Watson, D., 1995. Constructing validity: Basic issues in objective scale development. Psychol. Assess. 7, 309–319. doi:10.1037/1040-3590.7.3.309

Cortina, J.M., 1993. What is coefficient alpha? An examination of theory and applications. J. Appl. Psychol. 78, 98–104.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334. doi:10.1007/BF02310555

Crosby, R.D., Kolotkin, R.L., Williams, G.R., 2003. Defining clinically meaningful change in health-related quality of life. J. Clin. Epidemiol. 56, 395–407. doi:10.1016/S0895-4356(03)00044-1

de'Vet, H.C., Ostelo, R.W.J.G., Terwee, C.B., Van Der Roer, N., Knol, D.L., Beckerman, H., Boers, M., Bouter, L.M., 2007. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual. Life Res. 16, 131–142. doi:10.1007/s11136-006-9109-9

de'Vet, H.C., Terwee, C.B., Mokkink, L.B., Knol, D.L., 2011. Measurement in medicine: a practical guide. Cambridge University Press, Cambridge.

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Pratap, S., Wentz, R., Kwan, I., 2002. Increasing response rates to postal questionnaires: systematic review. BMJ 324, 1183. doi:10.1136/bmj.325.7361.444

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L.M., Pratap, S., 2009. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst. Rev. Jul 8. doi:10.1002/14651858.MR000008.pub4

Eng, J., 2005. Receiver operating characteristic analysis: A primer. Acad. Radiol. 12, 909–916. doi:10.1016/j.acra.2005.04.005

Fackrell, K., Hall, D.A., Barry, J.G., Hoare, D.J., 2016. Psychometric properties of the Tinnitus Functional Index (TFI): assessment in a UK research volunteer population. Hear. Res. 335, 220–235. doi:10.1016/j.heares.2015.09.009

Fackrell, K., Hall, D.A., Barry, J.G., 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.

Giavarina, D., 2015. Understanding Bland Altman analysis. Biochem. Medica 25, 141–151. doi:10.11613/BM.2013.003

Guyatt, G., Walter, S., Norman, G., 1987. Measuring change over time: assessing the usefulness of evaluative instruments. J. Chronic Dis. 40, 171–178. doi:10.1016/0021-9681(87)90069-5

Guyatt, G.H., Kirshner, B., Jaeschke, R., 1992. Measuring health status: what are the necessary measurement properties? J. Clin. Epidemiol. 45, 1341–1345. doi:10.1016/0895-4356(92)90194-R

Hair, J.F.J., Black, W.C., Babin, B.J., Anderson, R.E., 2009. Multivariate Data Analysis, Seventh Ed. ed. Prentice Hall.

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., Mazurek, B., 2016. Systematic review of outcome domains and instruments used in clinical trials of tinnitus treatments in adults. Trials 17, 270. doi:10.1186/s13063-016-1399-9.

Hallam, R.S., 2008. TQ Manual of the tinnitus questionnaire: Revised and updated. Polpresa

Press, London.

Hallam, R.S., 1996. Manual of the Tinnitus Questionnaire (TQ). The Psychological Corporation., London.

Hays, R.D., Woolley, J.M., 2000. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 18, 419–423. doi:10.2165/00019053-200018050-00001

Henry, J.A., Griest, S., Thielman, E., McMillan, G., Kaelin, C., Carlson, K.F., 2016. Tinnitus Functional Index: Development, validation, outcomes research, and clinical application. Hear. Res. 334, 58–64. doi:10.1016/j.heares.2015.06.004

Henry, J.A., Griest, S., Zaugg, T.L., Thielman, E.J., Kaelin, C., Galvez, G., Carlson, K.F., 2014. Tinnitus and Hearing Survey: A screening tool to differentiate bothersome tinnitus from hearing difficulties. Am. J. Audiol. 24, 66–77. doi:10.1044/2014

Hiller, W., Goebel, G., 1992. A psychometric study of complaints in chronic tinnitus. J. Psychosom. Res. 36, 337–348. doi:10.1016/0022-3999(92)90070-I

Hoare, D.J., Broomhead, E., Stockdale, D., Kennedy, V., 2015. Equity and person-centeredness in the provision of tinnitus services in UK National Health Service audiology departments. Eur. J. Pers. Centered Heathcare 3, 318–326.

Hoare, D.J., Edmondson-Jones, M., Sereda, M., Akeroyd, M.A., Hall, D., 2014. Amplification with hearing aids for patients with tinnitus and co-existing hearing loss. Cochrane Database Syst. Rev. 1–31. doi:http://dx.doi.org/10.1002/14651858.CD010151.pub2

Hoff, M., Kähäri, K., 2016. A Swedish cross-cultural adaptation and validation of the Tinnitus Functional Index. Int. J. Audiol. Dec 20, 1–9. doi:doi: 10.1080/14992027.2016.1265154.

Hu, L., 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.

Hu, L., Bentler, P.M., 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychol. Methods 3, 424–453. doi:10.1037/1082-989X.3.4.424

IBM Corp., 2012. IBM SPSS Statistics for Window, Version 21.0. IBM Corp., Armonk, NY.

Kamalski, D.M., Hoekstra, C.E., Van Zanten, B.G., Grolman, W., Rovers, M.M., 2010.

Measuring disease-specific health-related quality of life to evaluate treatment outcomes in tinnitus patients: A systematic review. Otolaryngol. - Head Neck Surg. 143, 181–185. doi:10.1016/j.otohns.2010.03.026

Kirshner, B., Guyatt, G., 1985. A methodological framework for assessing health indices. J. Chronic Dis. 38, 27–36. doi:10.1016/0021-9681(85)90005-0

Kottner, J., Streiner, D.L., 2011. The difference between reliability and agreement. J. Clin. Epidemiol. 64, 701–702. doi:10.1016/j.jclinepi.2010.12.001

Kottner, J., Streiner, D.L., 2010. Internal consistency and Cronbach's alpha: A comment on Beeckman et al. (2010). Int. J. Nurs. Stud. 47, 926–928. doi:10.1016/j.ijnurstu.2009.12.018

Kuk, F.K., Tyler, R.S., Russell, D., Jordan, H., 1990. The psychometric properties of a Tinnitus Handicap Questionnaire*. Amplif. Aural Rehabil. 11, 434–445.

MacCallum, R., Browne, M., Sugawara, H., 1996. Power analysis and determination of sample size for covariance structure modeling. Psychol. Methods 1, 130–149.

MacCallum, R.C., Roznowski, M., Necowitz, L.B., 1992. Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol. Bull. 111, 490–504.

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. Allied Sci. 26, 388–393. doi:10.1046/j.1365-2273.2001.00490.x

Meikle, M., Stewart, B., Griest, S.E., Martin, W.H., Henry, J., Abrams, H.B., McArdle, R., Newman, C.W., Sandridge, S., 2007. Assessment of tinnitus: Measurement of treatment outcomes. Prog. Brain Res. 166, 511–521.

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., Vernon, J.A., 2012. The Tinnitus Functional Index: development of a new clinical measure for chronic, intrusive tinnitus. Ear Hear. 33, 153–76. doi:10.1097/AUD.0b013e31822f67c0

Mohamad, N., Hoare, D.J., Hall, D.A., 2016. The consequences of tinnitus and tinnitus severity on cognition: A review of the behavioural evidence. Hear. Res. 332, 199–209.

doi:10.1016/j.heares.2015.10.001

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C.W., 2012. The COSMIN checklist manual. VU University Medical Centre, Amsterdam.

Müller, K., Edvall, N.K., Idrizbegovic, E., Huhn, R., Cima, R., Persson, V., Leineweber, C., Westerlund, H., Langguth, B., Schlee, W., Canlon, B., Cederroth, C.R., 2016. Validation of Online Versions of Tinnitus Questionnaires Translated into Swedish. Front. Aging Neurosci. 8. doi:10.3389/fnagi.2016.00272

Muthén, L., Muthén, B., 2012. Mplus User's Guide, 7th ed. Muthén & Muthén, Los Angeles, CA.

Newman, C., 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 evaluation treatment outcome. J. Am. Acad. Audiol. 9, 153–160.

Nunnally, J.C., 1978. Psychometric theory, 2nd ed. McGraw-Hill, New York.

Peterson, R.A., 1994. A meta-analysis of Cronbach's coefficient alpha. J. Consum. Res. 21, 381–391.

Pierzycki, R.H., McNamara, A.J., Hoare, D.J., Hall, D.A., 2016. Whole scalp resting state EEG of oscillatory brain activity shows no parametric relationship with psychoacoustic and psychosocial assessment of tinnitus: A repeated measures study. Hear. Res. 331, 101–108. doi:10.1016/j.heares.2015.11.003

Rabau, S., Wouters, K., Heyning, P. Van De, 2014. Validation and translation of the Dutch Tinnitus Functional Index. B-ENT 10, 251–258.

Ratnayake, S., Jayarajan, V., Bartlett, J., 2009. Could an underlying hearing loss be a significant factor in the handicap caused by tinnitus? Noise Heal. 11, 156–160.

Raubenheimer, J.E., 2004. An item selection procedure to maximise scale reliability and validity. SA J. Ind. Psychol. 30, 59–64. doi:10.4102/sajip.v30i4.168

Raykov, T., Marcoulides, G.A., 2011. Introduction to Psychometric Theory. Routledge.

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. doi:10.1016/j.jclinepi.2007.03.012

Satorra, A., Bentler, P.M., 1994. Corrections to test statistics and standard errors in covariance structure analysis, in: von Eye, A., Clogg, C.C. (Eds.), Latent Variable Analysis: Applications to Developmental Research. Sage Publications Inc, Thousand Oaks, CA., pp. 339–419.

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. doi:10.3200/JOER.99.6.323-338

Shevlin, M., Miles, J.N.., Davies, M.N.., Walker, S., 2000. Coefficient alpha: a useful indicator of reliability? Pers. Individ. Dif. 28, 229–237. doi:10.1016/S0191-8869(99)00093-8

Steiger, J.H., Lind, J.C., 1980. Statistically based tests for the number of common factors., in: Paper Presented at the Annual Meeting of the Psychometric Society. Iowa City.

Streiner, D.L., Norman, G.R., 2008. Health Measurement Scales : A practical guide to their development and use, 4th ed. Oxford University Press, Oxford.

Streiner, D.L., Norman, G.R., Cairney, J., 2014. Health measurement scales: a practical guide to their development and use., 5th ed. Oxford University Press, Oxford.

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.

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., de Vet, H.C.W., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34–42. doi:10.1016/j.jclinepi.2006.03.012

Tucker, L.R., Lewis, C., 1973. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 38, 1–10.

Uslu, R.I., Kapci, E.G., Oncu, B., Ugurlu, M., Turkcapar, H., 2008. Psychometric properties and cut-off scores of the beck depression inventory-II in Turkish adolescents. J. Clin. Psychol. Med. Settings 15, 225–233. doi:10.1007/s10880-008-9122-y

Vernon, J., Griest, S., Press, L., 1992. Plight of unreturned tinnitus questionnaires. Br. J. Audiol. 26, 137–138.

Wilson, P.H., Henry, J., Bowen, M., Haralambous, G., 1991. Tinnitus Reaction Questionnaire: psychometric properties of a measure of distress associated with tinnitus. J. Speech Hear. Res. 34, 197–201. doi:10.1044/jshr.3401.197

Wrzosek, M., Szymiec, E., Klemens, W., Kotyło, P., Schlee, W., Modrzyńska, M., Lang-Małecka, A., Preis, A., Bulla, J., 2016. Polish Translation and Validation of the Tinnitus Handicap Inventory and the Tinnitus Functional Index. Front. Psychol. 7, 1–11. doi:10.3389/fpsyg.2016.01871

Yong, A.G., Pearce, S., 2013. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutor. Quant. Methods Psychol. 9, 79–94. doi:10.20982/tqmp.09.2.p079

Yost, K.J., Eton, D.T., 2005. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. Eval. Health Prof. 28, 172–191. doi:10.1177/0163278705275340

Zou, K.H., O'Malley, A.J., Mauri, L., 2007. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115, 654–657. doi:10.1161/CIRCULATIONAHA.105.594929

8 FIGURES

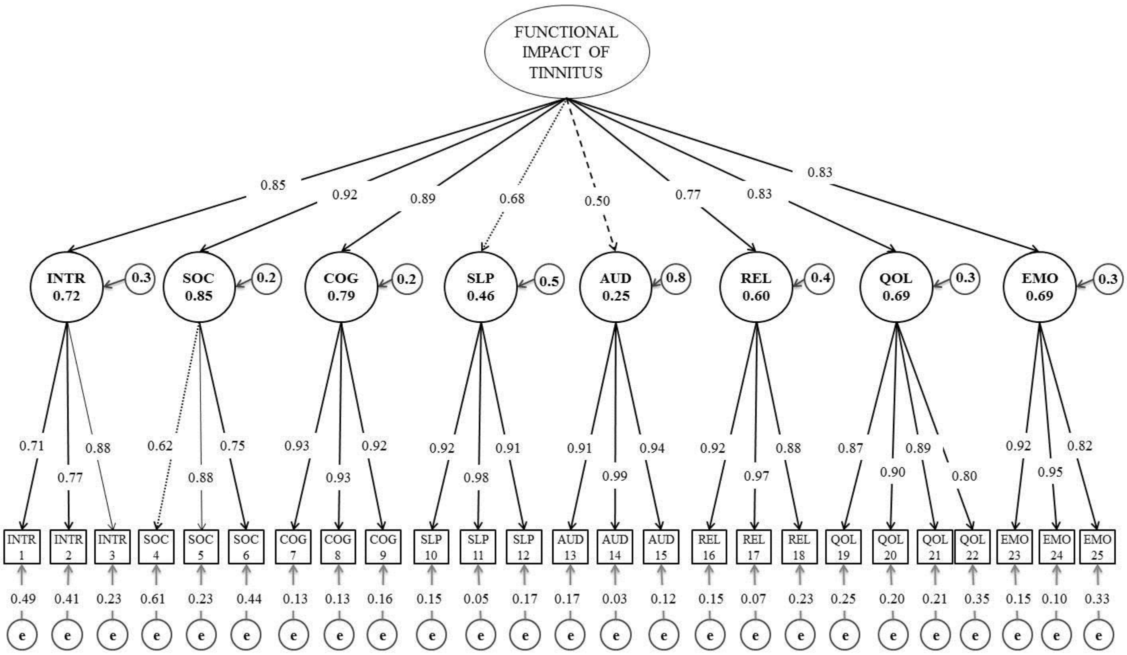


Figure 1. Original eight-factor TFI structure (25 items), as assessed by CFA, including standardised parameter estimates and r-squared values. The model represents the proposed relationships between the observed variables (items 1-25 e.g. INTR1), the first-order factors (INTR to EMO) and the second-order factor ("Functional impact of tinnitus"). The model represents: (i) a second-order latent construct with the variance fixed at 1; (ii) eight first-order latent constructs with the variance explained by second-order factor; (iii) 25 observed variables (INTR1 to EMO25) loaded on one factor only with the first item variance on each factor fixed at 1; and (iv) residual variance associated with each variable constrained to zero (represented by unidirectional grey arrows (→). The unidirectional arrows represent the direct effects of the latent constructs. The solid line arrows (⟶) indicate strong associations (> 0.70). The dotted arrows (┈┈→) indicate moderate associations with values below the desired range but still acceptable (> 0.60). The dashed line arrows (- - - - →) indicate poor associations (< 0.60). INTR = Intrusiveness; SOC = Sense of control; COG = Cognition, SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional; e = residual variance (error and uniqueness terms).

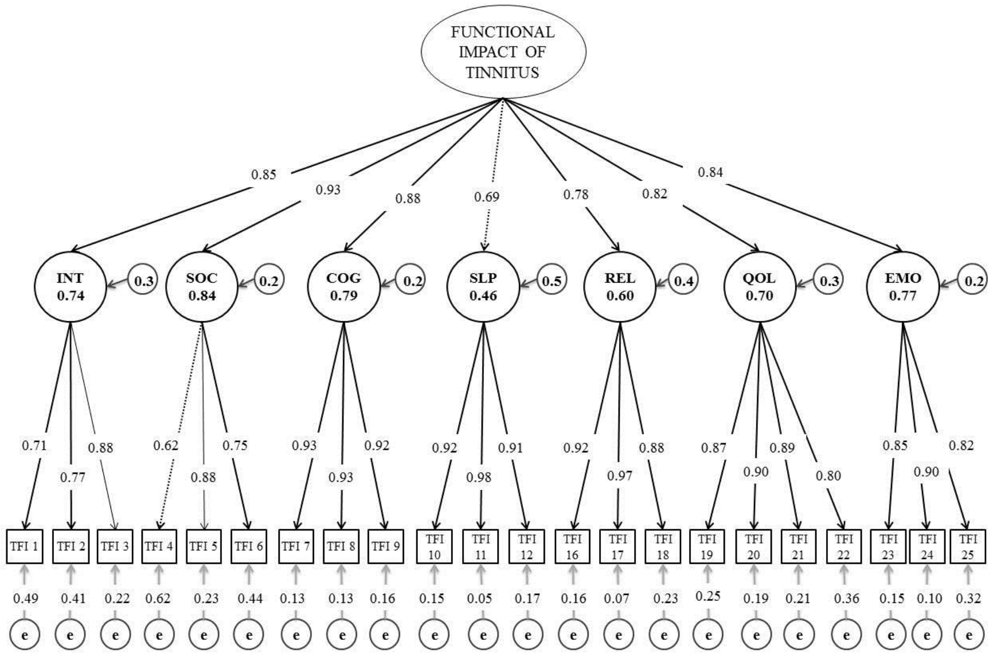


Figure 2. Our modified, re-specified TFI-22 seven-factor model including standardised parameter estimates and r-squared values. The standardised parameter estimates indicate the strength of the association between the 25 observed variables (INTR1 to EMO25), the seven first-order factors (INTR to EMO) and the second-order factor ("Functional impact of tinnitus"). The unidirectional arrows represent the direct effects of the latent constructs. The solid line arrows (⟶) indicate strong associations (> 0.70).The dotted arrows (⇢) indicate moderate associations with values below the desired range but still acceptable (> 0.60). INTR = Intrusiveness; SOC = Sense of control; COG = Cognition, SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional; e = residual variance (error and uniqueness terms)

.

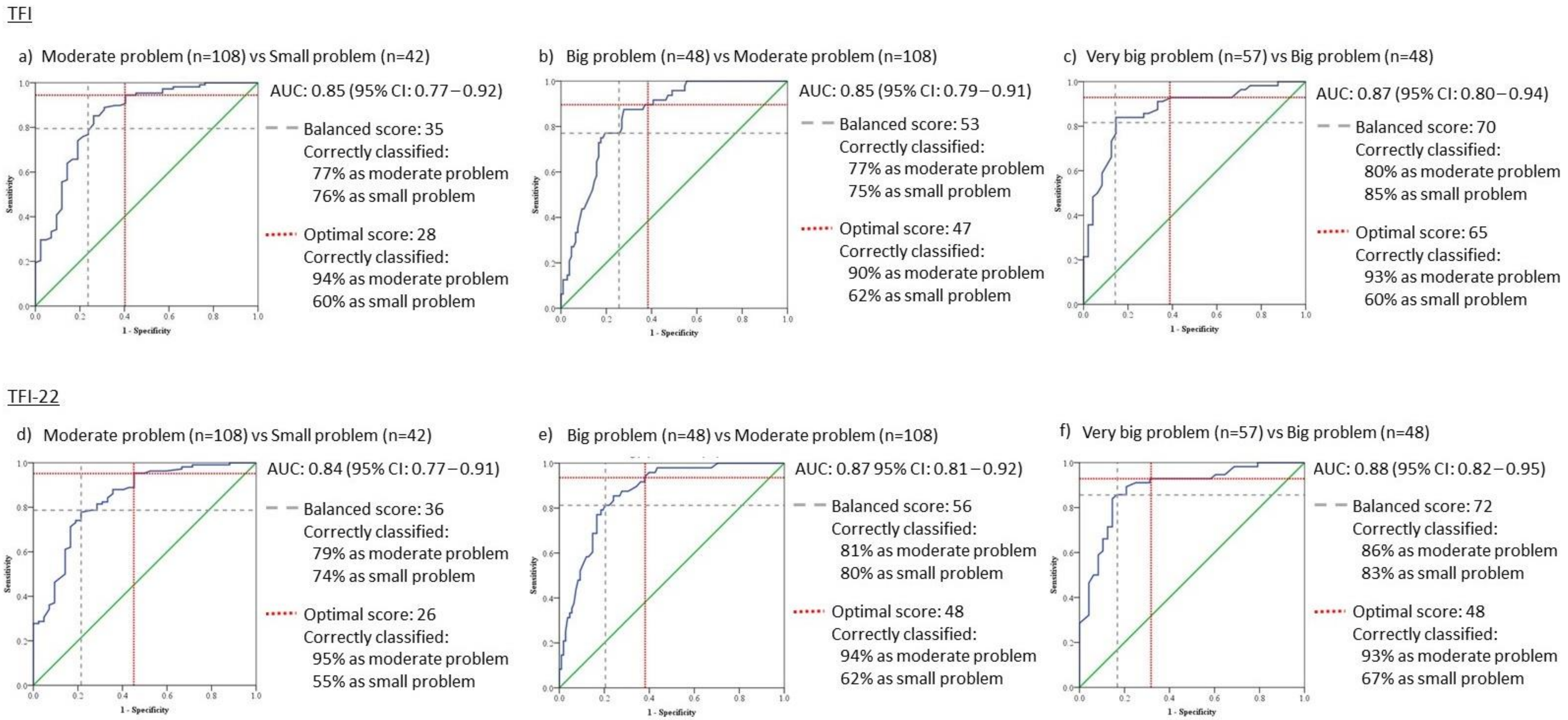


Figure 3. Receiver operating characteristic (ROC) curves for identifying optimal cut-off values for grading the original TFI and TFI22. ROC analysis was conducted using the global TFI scores and the global TFI22 scores comparing each problem category with the adjoining lower problem category (a/d) "Moderate problem" vs "Small problems". (b/e) "Big problem" vs "Moderate problem". (c/f) "Very big problem" vs "Big problem". Green line indicates 50% probability of correctly classifying improvement.

9 TABLES

Models	**Modified**	**S-B χ2 (*df*)**	**χ2/*df***	**p-value**	**TLI**	**CFI**	**SRMR**	**RMSEA (95% CI)**
TFI	None	577.50 (267)	2.16	<0.001	0.94	0.94	0.07	0.070 (0.06 – 0.08)
Re-specified TFI	Item error covariance	542.01 (264)	2.02	<0.001	0.94	0.95	0.07	0.067 (0.06 – 0.08)
TFI-22	Removed Auditory items	388.26 (202)	1.92	<0.001	0.95	0.96	0.05	0.062 (0.05 – 0.07)

Table 1. Summary of the model fit. Summary of fit statistics for the original eight-factor TFI model, re-specified TFI model adjusted for item error covariance, and the seven-factor TFI-22 model with Auditory factor removed. 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

		Mean (±SD)			Difference		Reliability	Measurement Error				
Scale	n (m)	Baseline (T0)	Follow-up (T1)	Follow-up (T2)	Mean diff	SD diff	ICC (95%CI)	SEM	Limits of Agreement			
									LoA	Lower limit (95% CI)	Upper limit (95% CI)	%
Original TFI	50	50.8 (±25.1)	45.9 (±22.8)	44.9 (±23.1)	-5.4	7.2	0.87 (0.80 - 0.93)	5.1	14.2	-19.6 (-23.2 to -16.1)	8.8 (5.2 to 12.3)	88
INTR	46 (4)	64.0 (±24.3)	55.6 (±23.2)	54.7 (±23.3)	-9.8	13.2	0.79 (0.63 - 0.88)	9.3	25.8	-35.6 (-42.3 to -28.9)	16.1 (9.3 to 22.8)	95
SOC	48 (1)	61.9 (±24.3)	56.5 (±24.2)	55.0 (±24.0)	-6.2	10.7	0.79 (0.69 - 0.87)	7.6	21.0	-27.2 (32.6 to -21.8)	14.8 (9.4 to 20.2)	90
COG	49	40.9 (±28.4)	39.3 (±27.2)	38.0 (±26.0)	-2.2	13.5	0.86 (0.79 - 0.91)	9.6	26.5	-28.7 (-35.5 to -22.0)	24.3 (17.6 to 31.0)	86
SLP	48	52.6 (±32.0)	46.4 (±29.2)	47.1 (±29.4)	-4.8	13.3	0.69 (0.57 - 0.80)	9.4	25.9	-30.8 (-37.2 to -24.3)	21.1 (14.6 to 27.6)	86
AUD	50	47.7 (±30.1)	47.5 (±28.8)	44.4 (±28.2)	-1.7	16.2	0.83 (0.74 - 0.89)	10.7	29.6	-30.9 (-38.4 to -23.5)	28.3 (20.9 to 35.8)	93
REL	50	62.0 (±29.3)	55.9 (±26.4)	53.5 (±27.4)	-7.4	14.4	0.75 (0.63 - 0.84)	10.2	28.3	-35.7 (-42.7 to -28.7)	20.8 (13.8 to 27.8)	86
QOL	49	38.6 (±32.4)	34.7 (±28.9)	33.3 (±29.0)	-4.6	11.3	0.79 (0.70 - 0.87)	8.0	22.2	-26.7 (-32.4 to -21.2)	17.6 (11.9 to 23.2)	88
EMO	50	42.8 (±31.5)	35.7 (±29.7)	38.4 (±30.2)	-5.7	14.6	0.82 (0.75 - 0.88)	10.3	28.6	-34.4 (-41.5 to -27.2)	22.9 (15.7 to 30.1)	92
TFI-22	50	51.3 (±25.9)	45.7 (±23.3)	45.0 (±23.7)	-5.9	7.1	0.90 (0.82 - 0.94)	5.0	13.9	-19.8 (-23.4 to -16.4)	8.0 (4.5 to 11.5)	88

Table 2. Reliability of Tinnitus Functional Index (TFI) scores: Intra-class correlations (ICC) and Limits of Agreement (LoA) between three administrations. The TFI showed excellent ability to distinguish between patients as indicated by the high ICC values and acceptable precision indicated by measurement error analyses. ICC = Intra-class correlations; Mean diff = the mean difference scores between administrations; SD diff = Standard Deviation of the difference; SEM = Standard Error in Measurement; LoA = Limits of Agreement.

Revised grading system for the UK population							Preliminary US-based grading systems			
	Original eight-factor 25 item TFI			Revised seven-factor TFI-22			Grading 1[a]		Grading 2[a,b]	
Diagnosis	Range	# patients (%)	Mean (±SD)	Range	# patients (%)	Mean (±SD)	Diagnosis	Range	Diagnosis	Range
No problem	0 - 7	0	-	0 - 7	0	-	No problem	0 - 17	Mild problems	< 25
Small problem	7 - 28	38 (15)	20.6 (±6.2)	7 - 26	31 (12)	18.4 (±5.4)	Small problem	18 - 31	Significant problems	25 – 50
Moderate problem	29 - 47	72 (28)	38.5 (±5.2)	27 - 48	80 (31)	38.9 (±6.4)	Moderate problem	32 - 53		
Big problem	48 - 65	70 (28)	56.5 (±5.5)	49 - 70	81 (32)	60.0 (±6.7)	Big problem	54 - 72		
Very big problem	66 - 100	75 (29)	79.1 (±9.9)	71 - 100	63 (25)	83.4 (±8.5)	Very big problem	73 -100	Severe problems	> 50

Table 3. Revised grading systems for the global TFI scores of the UK sample, compared with US-based grading systems. a (Henry et al., 2016); b (Meikle et al., 2012)

10 SUPPLEMENTARY TABLES

			Follow up questionnaires		
Procedure model	Audiology sites	Initial data	3 months	6 months	9 months
B	Aintree University Hospital NHS 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 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 Trust, Manchester	23	16	15	14
A	Countess of Chester, Chester*	10	7	6	6
A	Doncaster and Bassetlaw Hospitals NHS 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 Trust, Norwich	20	19	17	16
A	Sherwood Forest Hospital NHS Trust, Mansfield	21	19	18	18
	Total number of participants (% of total dropout)	255	198 (22%)	176 (31%)	166 (35%)

Supplementary Table 1. List of recruitment audiology sites and the number of participants providing initial and follow-up data. * To ensure the required sample size was recruited on schedule the Countess of Chester hospital was approved to recruit 10 participants in March 2014.

Participant characteristics	All data		CFA		Test-retest	
Sample size	253		239		50	
Missing	2		0		0	
Age in years						
Mean (±SD)	53.6	(±13.4)	53.3	(±13.5)	57.1	(±12.0)
Range	18 - 84		18 - 84		22 – 76	
	n	**(%)**	**n**	**(%)**	**n**	**(%)**
Gender						
Male	149	(58)	140	(59)	34	(68)
Female	105	(41)	98	(41)	16	(32)
Not reported	1	(<1)	1	(<1)	0	
How much of a problem is your tinnitus?						
No Problem	0	(0)	0	(0)	0	(0)
Small Problem	36	(14)	35	(15)	14	(28)
Moderate Problem	119	(47)	110	(46)	21	(42)
Big Problem	63	(25)	58	(24)	7	(14)
Very big problem	35	(14)	34	(14)	7	(14)
Missing	2	(1)	2	(1)	1	(2)
Duration of tinnitus						
≤ 2 years	124	(49)	117	(49)	20	(40)
3 to 10 years	73	(29)	69	(29)	16	(32)
11+ years	48	(19)	45	(19)	12	(24)
Missing data	10	(4)	8	(3)	2	(4)
Are you having any problems hearing speech or other sounds?						
No Problem	69	(27)	64	(27)	8	(16)
Small Problem	76	(30)	71	(30)	17	(34)
Moderate Problem	77	(30)	75	(31)	19	(38)
Big Problem	27	(11)	24	(10)	4	(8)
Very big problem	6	(2)	5	(2)	2	(4)

Supplementary Table 2. Participant characteristics for the total sample, the sample used in the CFA and the sample used for test-retest.

	T0				T1				T2				T3			
			Validity	Internal consistency			Validity	Internal consistency			Validity	Internal consistency			Validity	Internal consistency
Scale	n	Mean (SD)	*r* THI	α (95% CI)	n*	Mean (SD)	*r* THI	α (95% CI)	n	Mean (SD)	*r* THI	α (95% CI)	n	Mean (SD)	*r* THI	α (95% CI)
TFI	255	52.7 (21.7)	0.85	**0.96** (0.95 – 0.97)	196	44.7 (22.4)	0.83	**0.97** (0.97 – 0.98)	175	43.0 (23.7)	0.86	**0.98** (0.97 – 0.98)	165	42.9 (25.5)	0.85	**0.98** (0.98 – 0.99)
INTR	251	62.3 (22.0)	0.62	0.83 (0.79 – 0.86)	191	52.3 (23.8)	0.65	0.89 (0.86 – 0.91)	163	50.7 (25.2)	0.70	0.89 (0.85 – 0.91)	157	48.1 (25.8)	0.78	0.92 (0.89 – 0.94)
SOC	251	64.5 (21.7)	0.67	0.79 (0.74 – 0.83)	196	54.4 (24.6)	0.67	0.88 (0.85 – 0.91)	173	51.0 (25.7)	0.69	0.90 (0.87 – 0.92)	164	52.1 (27.4)	0.72	0.92 (0.90 – 0.94)
COG	255	47.1 (26.7)	0.74	0.95 (0.94 – 0.96)	193	41.0 (26.1)	0.74	**0.96** (0.95 – 0.97)	175	39.3 (27.1)	0.77	**0.96** (0.94 – 0.97)	165	38.2 (28.3)	0.75	**0.98** (0.97 – 0.98)
SLP	253	55.6 (31.9)	0.61	0.95 (0.94 – 0.96)	196	45.2 (30.6)	0.66	**0.96** (0.95 – 0.97)	175	42.4 (31.1)	0.66	**0.97** (0.96 – 0.97)	164	40.8 (33.2)	0.69	**0.97** (0.96 – 0.98)
AUD	254	42.6 (30.7)	0.41	**0.96** (0.95 – 0.97)	194	40.7 (28.4)	0.49	**0.97** (0.96 – 0.98)	175	40.7 (28.7)	0.60	**0.97** (0.96 – 0.98)	165	44.2 (30.6)	0.57	**0.98** (0.98 – 0.99)
REL	254	64.4 (27.8)	0.67	0.95 (0.93 – 0.96)	195	53.6 (26.7)	0.65	**0.96** (0.94 – 0.97)	173	51.4 (28.3)	0.74	**0.96** (0.95 – 0.97)	163	50.9 (29.4)	0.75	**0.97** (0.96 – 0.98)
QOL	255	39.9 (29.5)	0.76	0.92 (0.91 – 0.94)	196	33.7 (27.3)	0.77	0.94 (0.93 – 0.95)	175	33.8 (27.8)	0.82	0.95 (0.93 – 0.96)	165	34.2 (29.0)	0.80	**0.96** (0.95 – 0.97)
EMO	255	49.4 (30.4)	0.79	0.92 (0.91 – 0.94)	195	39.9 (29.6)	0.86	0.93 (0.92 – 0.95)	175	37.7 (30.0)	0.84	0.95 (0.94 – 0.96)	165	37.3 (30.9)	0.86	**0.96** (0.95 – 0.97)
TFI22	255	54.1 (22.4)	0.85	**0.96** (0.95 – 0.97)	195	54,1 (23.2)	0.84	**0.97** (0.97 – 0.98)	175	43.3 (24.3)	0.86	**0.98** (0.97 – 0.98)	164	42.8 (25.9)	0.86	**0.98** (0.98 – 0.99
THI	255	46.1 (23.8)	–	0.94 (0.93 – 0.95)	195	39.9 (22.5)	–	0.94 (0.92 – 0.95)	175	38.2 (23.6)	–	0.94 (0.93 – 0.96)	165	37.2 (23.5)	–	0.95 (0.93 – 0.96)

Supplementary Table 3. Descriptive statistics, convergent validity and internal consistency for the TFI and THI. The maximum score is 100. Values presented in bold indicate extremely high internal consistency (a > 0.95) above the recommended criteria (a < 0.95). α = Cronbach's Alpha estimates; SD = Standard Deviation; TFI = Tinnitus Functional Index (Meikle et al., 2012); THI = Tinnitus Handicap Inventory (Newman et al., 1996); T0 = baseline; T1 = 3 month follow-up; T2 = 6 month follow-up; T3 = 9 month follow-up.

Factor	1	2	3	4	5	6	7	8
(1) INTR	1							
(2) SOC	0.88	1						
(3) COG	0.74	0.79	1					
(4) SLP	0.61	0.62	0.59	1				
(5) AUD	0.48	0.43	0.53	**0.16**	1			
(6) REL	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) EMO	0.65	0.81	0.72	0.55	**0.28**	0.67	0.73	1

Supplementary Table 4. Correlations between first-order factors in the Confirmatory Factor Analysis. The correlations between the first-order factors were in general strong, with 70% above 0.60. The Auditory factor showed the weakest correlations with the other factors. Values presented in bold exceed recommended criteria.

First order factor	**Observed variable**	**TFI**				**TFI-22**			
		β	**B**	**SE**	**R^2**	**β**	**B**	**SE**	**R^2**
INTR	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
SOC	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
COG	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
SLP	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
AUD	AUD 13	0.91	1.00		0.83	removed			
	AUD 14	0.99	1.08	0.03	0.97	removed			
	AUD 15	0.94	1.10	0.03	0.88	removed			
REL	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
EMO	EMO 23	0.92	1.00		0.85	0.92	1.00		0.85
	EMO 24	0.95	0.93	0.03	0.90	0.95	0.93	0.03	0.90
	EMO 25	0.82	0.94	0.05	0.67	0.82	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**	removed			
	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

Supplementary Table 5. Parameter estimates, R-squared values and Standard Error for the original eight-factor TFI model and the seven-factor TFI-22 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. INTR = Intrusiveness; SOC = Sense of control; COG = Cognitive, SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional.

		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

Supplementary Table 6. Optimal grading, cut-off score, sensitivity and specificity rates for diagnosing small problems with tinnitus using the original global TFI. Bold values indicate the optimal threshold that prioritised sensitivity above specificity. Underlined values indicate the traditional threshold that is the balance between sensitivity and specificity.

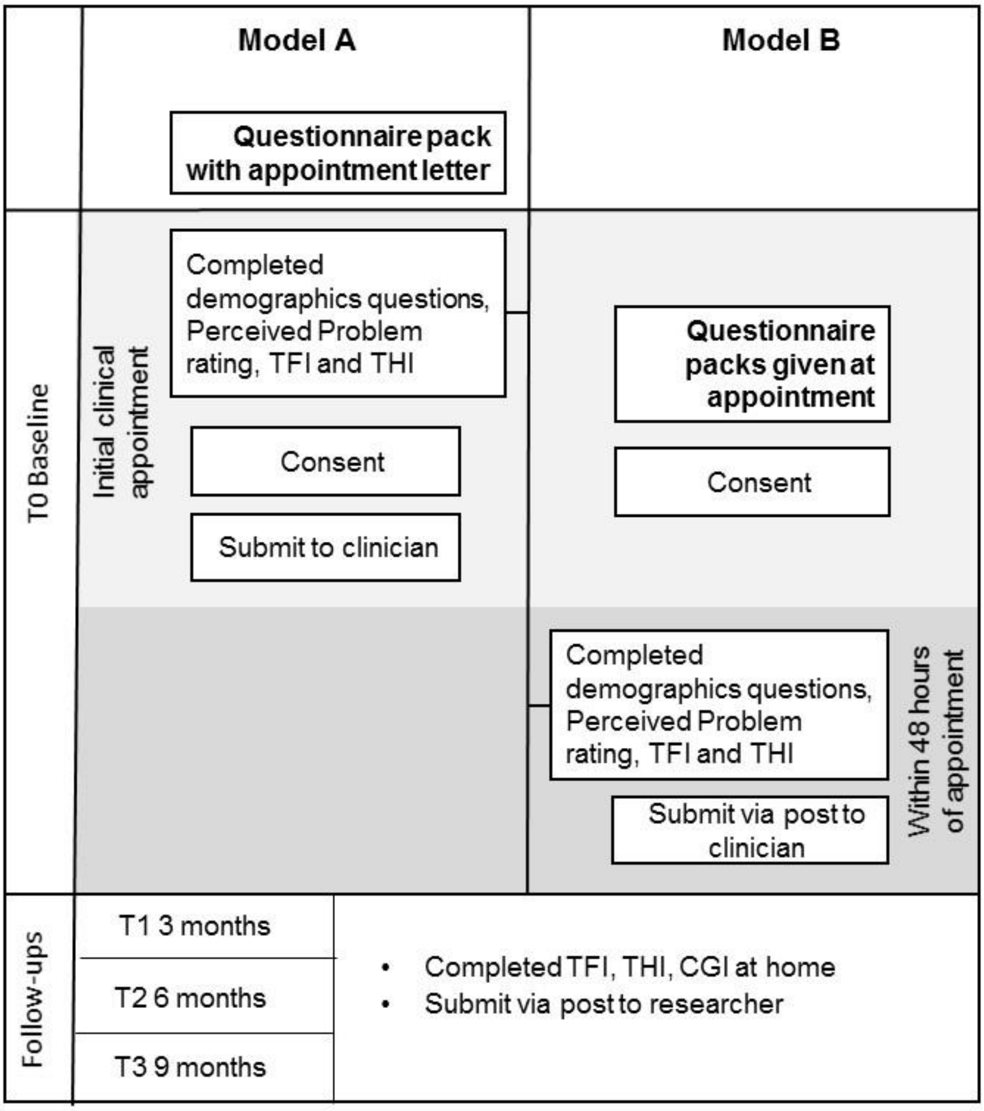
Moderate problem				
Optimal grading	Cut off score	Sensitivity	Specificity	1-Specificity
↕	29	1.00	0.00	0.90
	30	1.00	0.11	0.89
	31	1.00	0.15	0.85
	32	1.00	0.16	0.84
	33	1.00	0.19	0.81
	34	1.00	0.22	0.78
	35	1.00	0.24	0.76
	36	1.00	0.26	0.74
	37	1.00	0.27	0.73
	38	1.00	0.32	0.68
	39	1.00	0.35	0.65
	40	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

Supplementary Table 7. Optimal grading, cut-off score, sensitivity and specificity rates for diagnosing moderate problems with tinnitus using original global TFI. Bold values indicate the optimal threshold that prioritised sensitivity above specificity. Underlined values indicate the traditional threshold that is the balance between sensitivity and specificity.

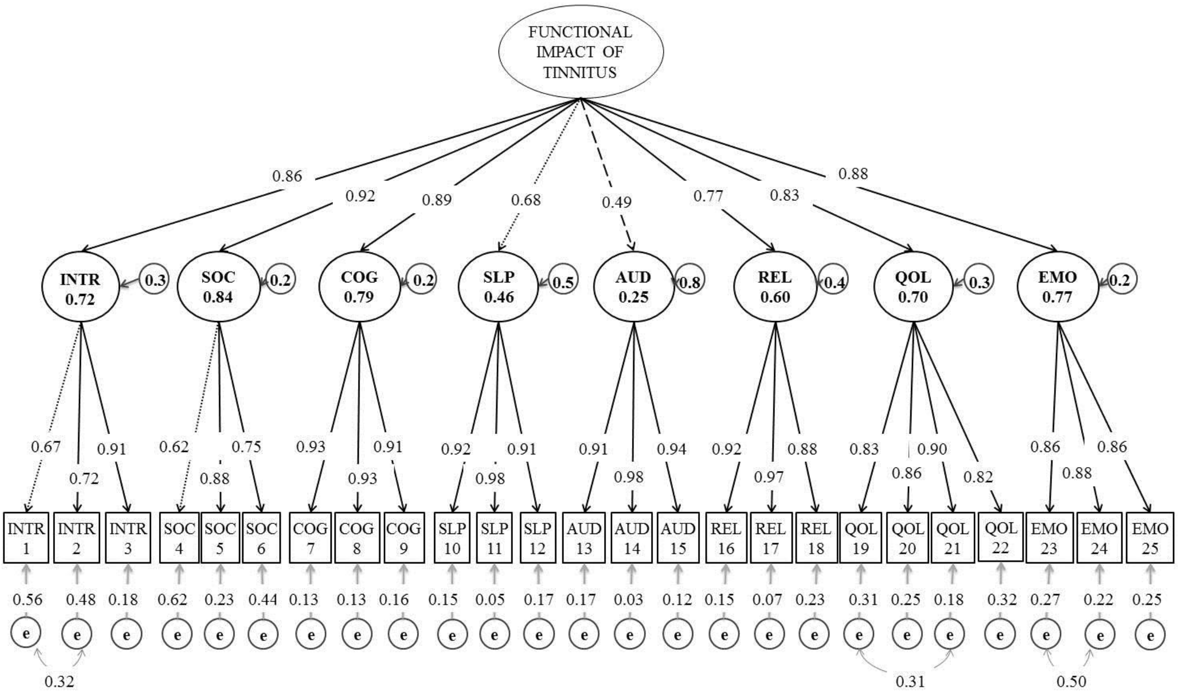
Big problem				
Optimal grading	Cut off score	Sensitivity	Specificity	1-Specificity
↕	48	1.00	0.00	0.90
	49	1.00	0.13	0.88
	51	0.98	0.13	0.88
	52	0.98	0.17	0.83
	53	0.98	0.21	0.79
	54	0.98	0.23	0.77
	57	0.98	0.25	0.75
	58	0.96	0.29	0.71
	59	0.93	0.35	0.65
	60	0.93	0.42	0.58
	61	0.93	0.48	0.52
	62	0.93	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

Supplementary Table 8. Optimal grading, cut-off score, sensitivity and specificity rates for diagnosing big problems with tinnitus using original global TFI. Bold values indicate the optimal threshold that prioritised sensitivity above specificity. Underlined values indicate the traditional threshold that is the balance between sensitivity and specificity.

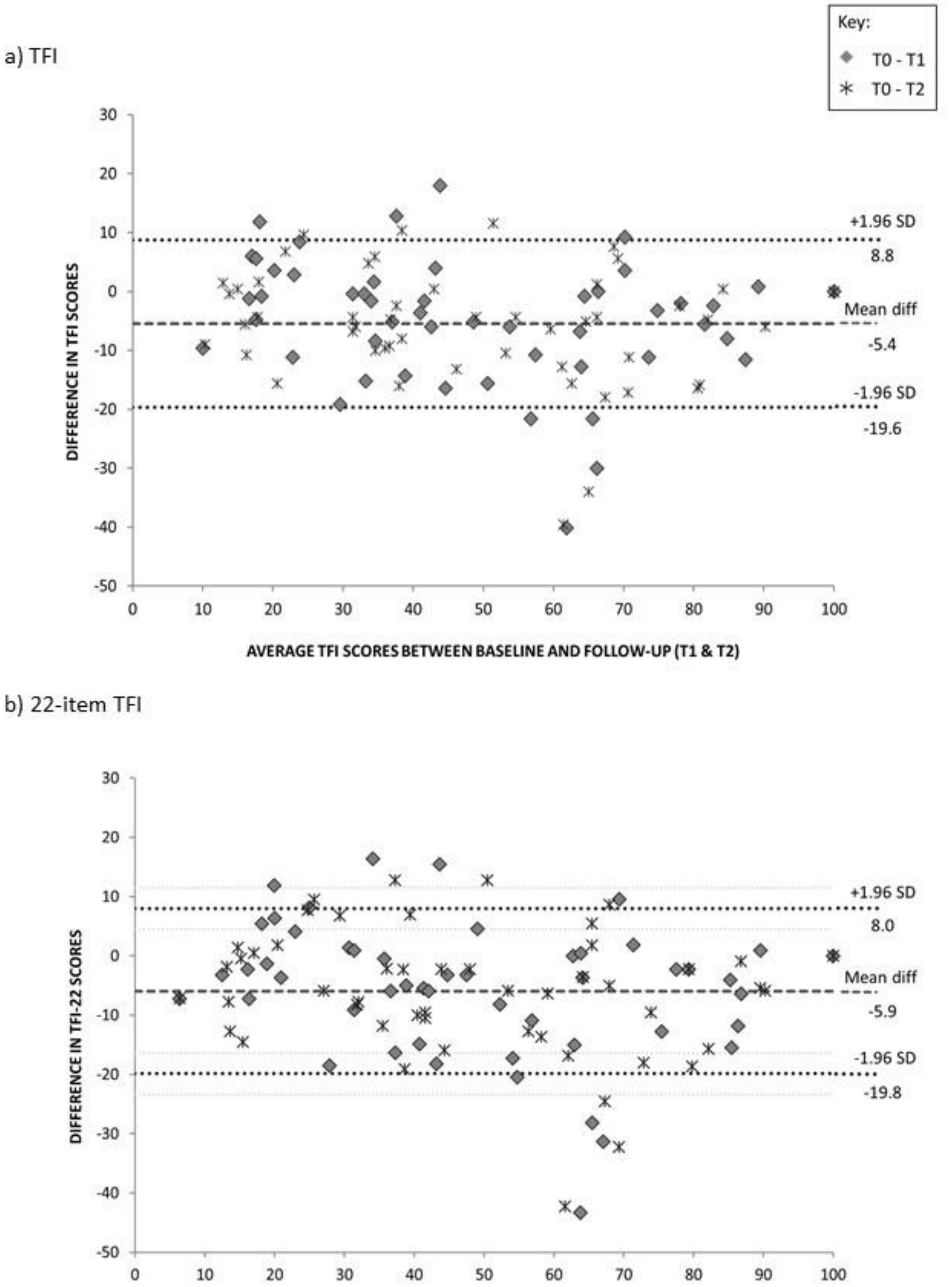
11 SUPPLEMENTARY FIGURES


Model A
Model B
Questionnaire pack with appointment letter
T0 Baseline
Initial clinical appointment
Completed demographics questions, Perceived Problem rating, TFI and THI
Consent
Submit to clinician
Questionnaire packs given at appointment
Consent
Completed demographics questions, Perceived Problem rating, TFI and THI
Submit via post to clinician
Within 48 hours of appointment
Follow-ups
T1 3 months
T2 6 months
T3 9 months
• Completed TFI, THI, CGI at home
• Submit via post to researcher

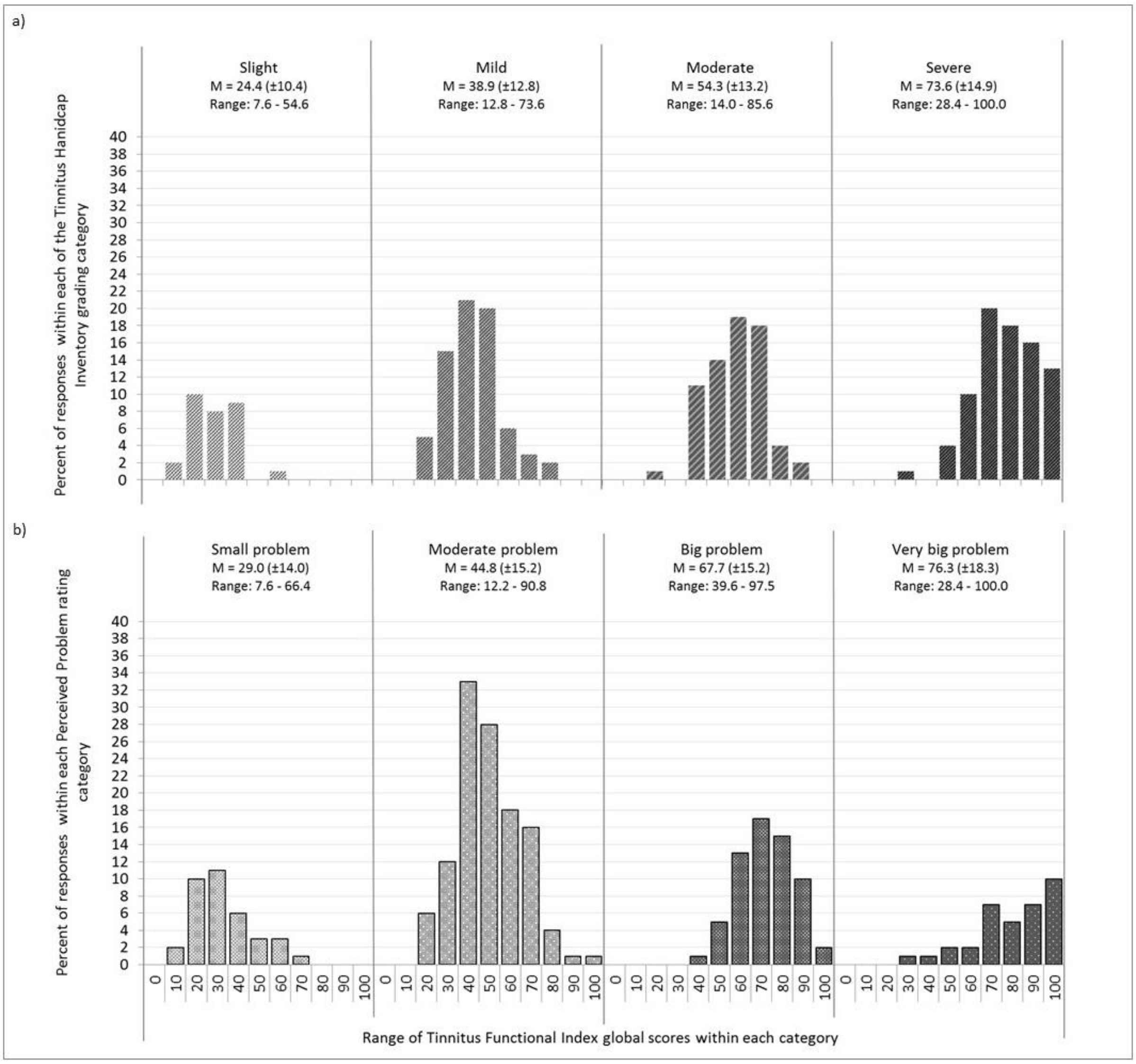
Supplementary Figure 1. Timeline of the study. Data collection at baseline (T0) followed one of two models to accommodate site differences in clinical appointment booking procedures. In Model A, the questionnaire packs were mailed to all prospective tinnitus patients with their initial appointment letters. Patients completed and returned the pack on the day of their initial appointment. At the assessment appointment, the clinician obtained written consent. For Model B, at the initial appointment, prospective tinnitus patients wishing to participate were consented and given the (T0) questionnaire pack and asked to complete and return the pack within 48 hours of the appointment. Follow-up questionnaire packs were sent to participants at 3 months (T1), 6 months (T2) and 9 months (T3) from their initial appointment date and were returned directly to the researcher.



Supplementary Figure 2. Re-specified eight-factor structure (25 items) including standardised parameter estimates and r-squared values. Standardised parameter estimates indicate the strength of the association between the 25 observed variables (items 1-25 e.g. INTR1), the eight first-order factors (INTR to EMO) and the second-order factor ("Functional impact of tinnitus"). The unidirectional arrows represent the direct effects of the latent constructs. The solid black line arrows (→) indicate strong associations (> 0.70). The dotted arrows (⋯→) indicate moderate associations with values below the desired range but still acceptable (> 0.60). The dashed line arrows (- - - →) indicate poor associations (< 0.60). The unidirectional black arrows indicate strong associations (> 0.70). The residual variance (e) represents the error and unique variance associated with each of the items and the factors residual and are represented by unidirectional grey arrows (→).The grey bidirectional curved arrows (↶↷) represent the association between the error variance of items. INTR = Intrusiveness; SOC = Sense of control; COG = Cognition, SLP = Sleep; AUD = Auditory; REL = Relaxation; QOL = Quality of life; EMO = Emotional; e = residual variance (error and uniqueness terms).



Supplementary Figure 3. Bland-Altman plot of measurement error for repeated measures (baseline, 3 months, 6 months) of the global TFI scores (a) and global TFI-22 scores (b) for self-defined "stable" participants. The Limits of Agreement (LoA) are represented as the mean difference ±1.96 times the standard deviation of the difference. The dashed line denotes the mean difference score. The dotted line denotes the 95% limits of agreement for the global scores. For the both, the TFI and TFI-22, 88% of the global scores are within the limits of agreement, suggesting a degree of measurement error between the repeated measures.



Supplementary Figure 4. Distribution of the original TFI global scores corresponding to the (a) Tinnitus Handicap Inventory grades of tinnitus severity and (iii) Perceived Problem rating categories. The distribution of the global TFI scores stratified according to two anchor-based approaches (a) Tinnitus Handicap Inventory grades, in which individual THI scores were assigned the appropriate grading (slight = 0 – 16; mild = 18 – 36; moderate = 38 – 56; severe = 58 – 100) and then the individual TFI global scores were stratified to the corresponding THI grade; (b) Perceived Problem rating category, in which patients rating of perceived problem were used to stratify the global TFI scores into one of the distinct categories (small problem, moderate problem, big problem or very big problem). The distribution of the global TFI scores across approaches were examined for similarities.