Psychometric Validity, Reliability, and Responsiveness of the Tinnitus Functional Index

Navshika Chandra*
Kevin Chang†
Arier Lee‡
Giriraj S. Shekhawat*§
Grant D. Searchfield*,**††‡‡

Abstract

Background: The effects of treatments on tinnitus have been difficult to quantify. The Tinnitus Functional Index (TFI) has been proposed as a standard questionnaire for measurement of tinnitus treatment outcomes. For a questionnaire to achieve wide acceptance, its psychometric properties need to be confirmed in different populations.

Objective: To determine if the TFI is a reliable and valid measure of tinnitus, and if its psychometric properties are suitable for use as an outcome measure.

Research Design: A psychometric evaluation of the TFI from secondary data obtained from a cross-sectional clinic survey and a clinical trial undertaken in New Zealand.

Study Sample: Confirmatory factor analysis and evaluation of internal consistency reliability were undertaken on a sample of 318 patients with the primary complaint of tinnitus. In a separate sample of 40 research volunteers, test–retest reliability, convergent and divergent validity were evaluated. Both samples consisted of predominantly older Caucasian male patients with tinnitus.

Results: The internal structure of the original US TFI was confirmed. The Cronbach's Alpha and Intra-class correlation coefficients were >0.7 for the TFI overall and each of its subscales, indicating high internal consistency and test–retest reliability. Strong Pearson correlations with the Tinnitus Handicap Questionnaire and tinnitus numerical rating scales indicated excellent convergent validity, and a moderate correlation with the Hearing Handicap Inventory, indicated moderate divergent validity. Evaluation of the clinical trial showed good test–retest reliability and agreement between no-treatment baselines with a smallest detectable change of 4.8 points.

Conclusions: The TFI is a reliable and valid measure of tinnitus severity in the population tested and is responsive to treatment-related change. Further research as to the TFI's responsiveness to treatment is needed across different populations.

Key Words: psychometric, questionnaire, severity, tinnitus, Tinnitus Functional Index

Abbreviations: CFA = confirmatory factor analysis; HHI = Hearing Handicap Inventory; ICC = intraclass correlation coefficient; SEM = standard error of measurement; tDCS = transcranial direct current stimulation; TFI = Tinnitus Functional Index; THI = Tinnitus Handicap Inventory; THQ = Tinnitus Handicap Questionnaire; TPFQ = Tinnitus Primary Function Questionnaire

*Audiology Section, The University of Auckland, Auckland, New Zealand; †Statistics, The University of Auckland, Auckland, New Zealand; ‡Epidemiology and Biostatistics, The University of Auckland, Auckland, New Zealand; §Health Systems, The University of Auckland, Auckland, New Zealand; **Centre for Brain Research, The University of Auckland, Auckland, New Zealand; ††Brain Research New Zealand, Auckland, New Zealand; ‡‡Eisdell Moore Centre, The University of Auckland, Auckland, New Zealand

Corresponding Author: Grant D. Searchfield, Audiology Section, University of Auckland Tamaki Campus, Auckland 1142, New Zealand; Email: g.searchfield@auckland.ac.nz

INTRODUCTION

Tinnitus is the perceived sensation of sound in the absence of an acoustic stimulus (Holmes and Padgham, 2009). Epidemiological studies conducted in different countries report the prevalence of chronic tinnitus to be between 10% and 15% of the general adult population (Hoffman and Reed, 2004). In New Zealand, the prevalence of tinnitus is estimated to be 6% of the total population, rising to 13.5% in people aged 65 yr and above (Wu et al, 2015). Chronic tinnitus (tinnitus of >6 mo duration) can be associated with a range of negative effects, including sleep interference and concentration difficulties (Tyler and Baker, 1983). Tinnitus can also have negative effects on emotional well-being resulting in irritability, frustration, and depression. Such disabling effects can lead to handicap at the societal level in the form of poor work performance, withdrawal from social activities, lessened interest in leisure activities, and ultimately reduced quality of life (Sanchez and Stephens, 1997; Meikle, 2002; Holmes and Padgham, 2011).

There is currently no cure for tinnitus (Holmes and Padgham, 2009; 2011; Tunkel et al, 2014). Although research to evaluate currently used methods and to develop new treatments is ongoing, evaluating the effectiveness of treatment methods has been impeded by a lack of standardized measures validated for both intake assessment and outcomes measurement (Meikle et al, 2012). The assessment of patients with tinnitus is a challenging task as there are currently no objective tests that can verify the presence of tinnitus or evaluate its severity (Meikle et al, 2007; Møller, 2011). Self-report questionnaires quantifying the negative effects and severity of tinnitus thus play an important role in the clinical evaluation of patients with tinnitus. Given the subjective nature of tinnitus, such scales are also vital for monitoring treatment outcomes in the clinical setting and testing the efficacy of treatment interventions in the research arena (Newman and Sandridge, 2004). Even though several of the tinnitus questionnaires have become widely used and translated into several languages, none were universally accepted (Meikle et al, 2007; Kamalski et al, 2010). The application of different questionnaires across clinical trials has made comparing treatment outcomes difficult as the questionnaires vary with respect to their format, scaling, wording, and number of items, and are likely to vary in their responsiveness to treatment-related change (Langguth et al, 2007; Landgrebe et al, 2012). Standardizing measures would improve comparability of treatment outcomes between different treatment interventions and centers, facilitate meta-analyses, allow a more consistent basis for defining selection criteria, and assigning participants to treatment groups, and therefore, allow conclusions to be reached regarding the effectiveness of different treatment options. This would significantly aid progress in finding effective treatments for tinnitus (Meikle, 2002; Meikle et al, 2007; 2012).

Between 1988 and 1999, at least nine English-language self-report questionnaires were devised by independent groups of authors (Meikle et al, 2008; 2012). However, none of these tinnitus questionnaires covered all of the dimensions of tinnitus-related impact. Excluding dimensions could lead to inaccurate evaluation of a patient's tinnitus severity level and major areas of tinnitus impact, as well as limit a questionnaire's ability to show treatment-induced changes (Meikle et al, 2007; 2008; Langguth, 2011). Furthermore, these scales were designed primarily for intake evaluation or discriminative purposes, that is, for characterizing individual differences between patients with regard to their perceived tinnitus severity and significant areas of impact and have been validated for this purpose (Meikle et al, 2007; 2008). However, a systematic review has found that despite being used widely as outcome measures, none have been assessed regarding their responsiveness to treatment-related change or have been validated for this purpose (Kamalski et al, 2010). Their ability to detect treatment-related change is thus unknown (Kamalski et al, 2010; Meikle et al, 2012). Newman et al (1998) investigated the psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcomes. Due to high retest reliability and small standard error of measurement (SEM), the THI was considered by the authors to be useful for detecting treatment-related changes in self-perceived tinnitus severity. For a treatment to be considered effective, it was suggested that pre- and postintervention scores would have to differ by at least 20 points (Newman et al, 1998). However, in the past 2 decades, advances in psychometric research have occurred, highlighting the importance of emphasizing content validity, maximizing effect sizes, reducing floor and ceiling effects, and including fine-grained measurement intervals, in developing responsive outcome measures (Meikle et al, 2012; Fackrell et al, 2016a).

Two relatively new questionnaires have been developed to address the perceived shortcomings of existing questionnaires (Meikle et al, 2012; Tyler et al, 2014). The Tinnitus Primary Function Questionnaire (TPFQ) is the most recent of these questionnaires (Tyler et al, 2014). It has demonstrated high reliability, construct validity, and sensitivity to treatment-related change (Tyler et al, 2014). Statistically significant differences between pre- and post-treatment scores (total and subscale) were demonstrated. The use of a 0–100 scale and inclusion of questions only evaluating the primary effects of tinnitus were considered by Tyler et al (2014) to enhance responsiveness. The authors mentioned four primary effects of tinnitus on daily life: thought and emotions,

hearing, sleep, and concentration. As it is a new questionnaire, little is known yet about its reliability in different populations.

The Tinnitus Functional Index (TFI) is the focus of this study. The TFI was also recently developed in the United States (Meikle et al, 2012) and has undergone evaluation in several different populations (Rabau et al, 2014; Fackrell et al, 2016a,b; Henry, Griest, et al, 2016). Its developers aimed to provide comprehensive coverage of the multiple dimensions of tinnitus-related impact and have documented validity for both scaling the negative impact and severity of tinnitus, and for measuring treatment-related changes (responsiveness) (Meikle et al, 2012). The TFI was developed over a period of 4 yr, in which successively shorter versions of the questionnaire were tested resulting in the final 25-item version (Meikle et al, 2012). Most of the items in the TFI were selected from a pool of 175 items that were included in the nine preexisting questionnaires. The 25 items in the final version of the questionnaire were extracted as the best-functioning items from the responses obtained on a 30-item prototype obtained on a sample of 347 patients. Factor analysis revealed eight underlying dimensions: (a) intrusiveness of tinnitus; (b) sense of control; (c) cognitive interference; (d) sleep disturbance; (e) auditory difficulties; (f) interference with relaxation; (g) quality of life; and (h) emotional distress. High internal consistency and test–retest reliability of the overall scale and subscales were demonstrated. Strong correlations with the Tinnitus Handicap Questionnaire (THQ) and Visual Analog Scale suggested good convergent validity and moderate correlations with the Beck Depression Inventory-Primary Care demonstrated good divergent validity (Meikle et al, 2012). Effect sizes at 3 and 6 mo indicated high responsiveness to treatment-related change. The authors stated that these results represent initial steps toward development of expert consensus regarding use of the TFI as one part of a core set of standardized measures for tinnitus research and clinical practice (Meikle et al, 2012). The final 25-item TFI has been criticized as having been validated using data derived from participants' responses to the 30-item prototype (Fackrell et al, 2016a). Tyler et al (2014) state that a potential limitation of the TFI was its inclusion of items measuring secondary effects of tinnitus, such as impact on enjoyment of life, relationships with people, and ability to work. These were considered to reduce the sensitivity to treatment-related change because of being just as likely to be influenced by everyday life events as a treatment intervention (Tyler et al, 2014). At this stage, it is unknown whether inclusion of questions measuring secondary effects of tinnitus reduces the responsiveness of a questionnaire to treatment-related change.

Psychometric scales developed and validated in one population do not necessarily measure identical underlying dimensions in another population. Both the TFI and TPFQ were developed and initially tested in the Unites States. Validating questionnaires in different settings is therefore important to optimize it for those populations (Langguth et al, 2011). A measure should retain most of its original items and internal structure when validated in different populations if it is to be used as a standard (Langguth et al, 2011). The TFI has been translated and validated for use in a Dutch-speaking population. High internal consistency reliability was demonstrated by a Cronbach's Alpha of 0.96, and statistically significant correlations with the Visual Analog Scales for maximum and mean tinnitus loudness, and with percentage of time aware of tinnitus, indicated good convergent validity. Exploratory factor analysis with eight fixed factors showed that the eight factors in the original version could be used in the Dutch population (Rabau et al, 2014). Rabau et al (2014) found statistically significant correlations between the TFI and rating scales for maximum loudness ($r = 0.59$) mean loudness ($r = 0.66$) and awareness of tinnitus ($r = 0.58$). The TFI has also been evaluated in a research volunteer population in the United Kingdom (Fackrell et al, 2016a). The TFI total scores were more evenly distributed across possible scores than the THI and THQ. Statistically significant, strong correlations between the overall score of the TFI and the overall THQ and THI scores, and moderate correlations with a rating of percentage annoyance, a Visual Analog Scale of Loudness, the Beck's Depression Inventory, Beck's Anxiety Inventory, and the World Health Organization Quality of Life Assessment-BREF demonstrated high construct validity (Fackrell et al, 2016a). An intraclass correlation coefficient (ICC) of 0.86 indicated high test–retest reliability for the TFI overall score, and ICCs ranging from 0.81 to 0.95 for the subscale scores. Acceptable test–retest agreement (93%) was also found. Cronbach's Alpha was 0.80 for the overall score and was high for seven of the subscales, indicating high internal consistency. Only the intrusiveness subscale had a low alpha estimate of 0.58. The 8-factor structure found in the United States was not fully confirmed in the UK sample. The *auditory difficulties* subscale demonstrated poor loading with the higher order factor-functional impact of tinnitus. An optimized model did not differ greatly from the original (Fackrell et al, 2016a). Floor effects were found for 50% of the items, which the authors stated indicated that the TFI would be somewhat limited in its ability to detect treatment-related change in this population. Fackrell et al (2016a) also found that the smallest detectable change was 23 points, considerably higher than 13 points recommended by Meikle et al (2012) as an interim indicator of clinically meaningful change. The smallest detectable change measure is an effort to identify true or meaningful change, it differs from statistical significance. The

statistical difference is used in hypothesis testing to reduce "chance" results but identifies if a difference exists, not if the difference is sufficient to affect the patient. A large sample size of a population will identify smaller differences as being significant than small samples. Although groups (treatments) may be statistically different, it does not mean that the size of difference is clinically important. Clinically important differences, such as the smallest detectable change in TFI, are derived from group data but may be applied to individuals, in an attempt to ensure that any change is not the result of measurement error. Caution in interpreting both the meaning of statistically significant and clinically meaningful change is needed when applying to individuals and populations different from the sample evaluated. Fackrell et al (2016a) conclusions have been challenged by several authors of the original TFI paper (Folmer, 2016; Henry, Thielman, et al, 2016). Folmer (2016) expressed concern at the absence of outcome data in the evaluation of degree of change whereas Henry, Griest, et al (2016) were critical of conclusions being drawn on a sample from a research population rather than a clinical population.

Variation in responses for different populations is to be expected; language, culture, and other psychosocial characteristics affect tinnitus (Searchfield, 2014). It is important for questionnaires to be validated in populations representative of those where they will be used. Tinnitus questionnaires developed in the United States (e.g., THQ and THI) and trialed in New Zealand in the past have required changes in their factor structure to be optimized for our population and cultural differences (Searchfield et al, 2007; Searchfield and Jerram, 2010). Due to differences in population and debate as to the merit and shortcomings of the TFI (Fackrell et al, 2016a,b; Folmer, 2016; Henry, Thielman, et al, 2016), validation in our clinical and research population was needed.

The aims of this study were as follows:

1. To trial the final version of the TFI as an assessment of severity in a tinnitus clinic population.
2. To ascertain if the TFI is a reliable and valid measure for scaling the negative impact and severity of tinnitus in our tinnitus clinical population.
3. To determine what modifications, if any, need to be made to the TFI to optimize it for use in New Zealand.
4. To determine the usefulness of the TFI as an outcome measure in a research context.

METHODS

This research used anonymous secondary data from two studies previously conducted at the University of Auckland. One of these was an unpublished cross-sectional study investigating drug use among patients with tinnitus, which was conducted during 2012 and 2013. The second study that provided data for this research was a clinical trial investigating the effect of hearing aids with and without transcranial direct current stimulation (tDCS) on tinnitus that included a test–retest baseline (Shekhawat et al, 2014). The University of Auckland Human Participants Ethics Committee approved the research.

The TFI

The TFI consists of eight subscales. A 0- to 10-point Likert scale is used to measure the response to each item. The overall TFI and each of the eight subscale scores can range from 0 to 100, and is calculated by summing the responses obtained from all questions, dividing by the number of questions answered, and multiplying by 10 (Meikle et al, 2012). Participants missing two or more items on a subscale were excluded when calculating subscale mean scores. Participants with seven or more missing items on the TFI were excluded when calculating the overall mean score.

Study 1. Survey

Participants

Participants were all patients with tinnitus attending the University of Auckland Hearing and Tinnitus Clinic. Potential participants (871) were sent, via e-mail or post, a number of questionnaires including the TFI and a short questionnaire for collecting demographic data. The response rate for this survey was 36.5% and the sample size was 318. Nonresponders could not be contacted under the terms of our ethical approval. Data from this study were used to carry out confirmatory factor analysis (CFA) and to evaluate the internal consistency reliability of the TFI and its subscales. Table 1 shows the demographic characteristics of the study participants. The sample consisted of more males than females (57.5% versus 40.9%). Few (1.6%) did not state their gender. Most of the participants were in the 50- to 79-yr age range and most (84.3%) identified as New Zealand European (Caucasian).

Analysis

The factor analysis and internal consistency reliability analysis were carried out using data from Study 1. A CFA replicating the evaluation of TFI in the United Kingdom was undertaken (Fackrell et al, 2016a). Due to list-wise deletion of cases with one or more missing values during factor analysis, mean substitution of the missing item responses was performed. In this method, each missing value is replaced with the sample mean score

Table 1. Demographic Characteristics of Study Participants, Number (Percentage)

Demographic Characteristics	N (%)	
	Part 1 (N = 318)	Part 2 (N = 40)
Gender		
Male	183 (57.5)	36 (90)
Female	130 (40.9)	4 (10)
Missing	5 (1.6)	
Age (years)		
<30	5 (1.6)	
30–39	11 (3.5)	
40–49	26 (8.2)	5 (12.5)
50–59	77 (24.2)	16 (40)
60–69	106 (33.3)	14 (35)
70–79	73 (23.0)	5 (12.5)
≥80	18 (5.7)	
Missing	2 (0.6)	
Ethnicity		
New Zealand European	268 (84.3)	
Other European	27 (8.5)	
Maori or Pacific Island	5 (1.6)	
Indian	4 (1.3)	
Chinese	1 (0.3)	
Other Asian	4 (1.3)	
Other	7 (2.2)	
Missing	2 (0.6)	

of the item. Overall, 41 participants had missed at least one item on the TFI. Ten participants had missing values for more than 10% of the TFI items (three or more items) and therefore were excluded before mean substitution was carried out. Mean substitution was thus only carried out on 10% of the participants (31 participants) and only on those participants with 10% or less of their items missing. Based on a study conducted by Downey and King (1998), item mean substitution is recommended when the number of respondents with missing items and the number of items missing for each respondent is <20%. The cutoff level of 10% used in this study is thus acceptable and was further chosen because this approach to dealing with missing data were used by Meikle et al (2012). The final dataset (n = 308) was used to carry out factor analysis and evaluate the internal consistency reliability of the TFI using correlation coefficients.

The CFA was performed using Lavaan R package (Rosseel, 2012) in R version 3.2.3 (R Core Team, 2015). The 8-factor model was defined based on four properties:

1. The latent constructs: eight first-order factors corresponding to the eight TFI subscales and one second-order factor corresponding to the overall measure, namely "Impact of tinnitus on function."
2. Each item only loaded on to the designated first-order factor.

3. Residual variance associated with each variable was assumed to be uncorrelated and random.
4. The variance of the second-order factor, "Impact of tinnitus on function," was fixed at 1 as it was assumed that the first-order factors are completely explained by the relationship to the second-order factor.

Due to the non-normality in the TFI data, the model was estimated using maximum likelihood parameter estimation adjusted with Satorra–Bentler scaled χ^2 (Satorra and Bentler, 1994) and standardized root mean square residual (Hu and Bentler, 1998) to ensure robust standard errors for parameter estimates and goodness of fit, and the comparative fit index (Bentler, 1990), root mean square error of approximation (Steiger and Lind, 1980), and Tucker–Lewis index (Tucker and Lewis, 1973) were used.

Although there is no consensus regarding the appropriate sample size required to conduct factor analysis, between 5 and 10 participants, per item has been recommended. This is also acceptable for evaluating internal consistency reliability (Young and Pearce, 2013). The sample size of 308 used in this study is thus adequate for evaluating the 25-item TFI.

Internal Consistency

Reliability, which refers to the degree to which an instrument is free from random error, was measured by evaluating internal consistency reliability of the TFI (Lohr, 2002). The internal consistency reliability of a questionnaire is the degree to which questionnaire items measure the same underlying concept (Terwee et al, 2007). Cronbach's Coefficient Alpha, which represents the degree to which items in a questionnaire are intercorrelated, was computed to examine the internal consistency reliability of the TFI and its subscales. A Cronbach's Alpha of >0.7 indicates acceptable internal consistency reliability (Yu, 2001).

Study 2. Trial

Participants

For the clinical trial of tDCS and hearing aids (Shekhawat et al, 2014), participants were recruited via the University of Auckland Hearing and Tinnitus Clinic and ResearchStudies, an online portal that connects research volunteers with research opportunities. The sample size in this study was 40. The population consisted of more males than females (90% versus 10%). Most of the participants were in the 50- to 69-yr age range (Table 1). Inclusion criteria for the study were chronic tinnitus (more than 2 yr), aidable hearing loss with no previous experience of hearing aid use, and a

minimum score of 25 on the TFI. Volunteers were excluded if they had any contraindications for undergoing tDCS (personal or family history of seizures, metal and electronic implants, pregnancy, heart conditions, brain surgery, and others) as screened by a neurologist. Twenty volunteers not meeting the inclusion criteria were excluded.

Measures Used for Convergent and Divergent Validity

The THQ is a 27-item questionnaire used to scale the negative impact and severity of tinnitus. Individuals responded with a number between 0 and 100 for each item to indicate how strongly he or she agreed (100) or disagreed (0) with the statement. The questionnaire was composed of three subscales: physical, social, and emotional effects of tinnitus, hearing difficulties, and the patient's view of their tinnitus (Kuk et al, 1990). Subscale scores were calculated by summing the item scores obtained in each subscale, multiplying this by the number of items in the subscale, and dividing this value by 27. The scores obtained on each of the three subscales were added to obtain the overall THQ score. Higher overall scores reflected greater tinnitus severity. Factor 3 of the THQ, which assesses the patient's view of their tinnitus, is not usually used as a separate subscale as it has poor psychometric reliability (Kuk et al, 1990).

Six numerical rating scales were used to measure how annoying, unpleasant, and strong/loud an individual's tinnitus was, as well as how uncomfortable it makes one feel, how difficult it was to ignore it, and how much of a problem an individual's tinnitus was. Each of these questions was measured on a 10-point Likert scale (1–10) except for the *how much of a problem is your tinnitus* question, which was measured on a 5-point scale. The scales were bounded by a statement indicating direction of response, the *how much of a problem is your tinnitus* question asked for selections from five statements from "not a problem" to a "very big problem"; the 10-point scale was bound by statements such as (*for the how difficult it was to ignore it question*): very easy to ignore (1) to impossible to ignore (10).

The Hearing Handicap Inventory (HHI) is a 25-item questionnaire used to assess the degree of psychosocial impact caused by hearing loss (Newman et al, 1990). It is composed of two subscales: emotional handicap and social/situational handicap. Individuals indicated their level of agreement to each statement using one of three response options: yes (4 points), sometimes (2 points), and no (0 points). The total score can range from 0 to 100 with higher scores indicating greater perceived handicap associated with hearing loss. Subscale scores were obtained by summing the item scores and the total

HHI score was obtained by adding the subscale scores (Newman et al, 1990).

Participants completed the TFI at two points in time (two weeks apart) before the treatment intervention. The Hearing Handicap Inventory (HHI), THQ, and six numerical rating scales were also administered during the second assessment. Data from the first and second baseline were used to evaluate test–retest reliability of the TFI and its subscales, and data collected during the second assessment were used to evaluate convergent and divergent validity. A time interval of one or two weeks between the repeated assessments is considered by Terwee et al (2007) to be sufficient in preventing recall, yet short enough to ensure no clinical changes have occurred. Shi (2008) further recommends a gap of at least two weeks between the two assessment points. The test–retest period of two weeks in this study was thus considered appropriate for evaluating test–retest reliability. Based on the recommendation of Terwee et al (2007) that a sample size of 50 is adequate for evaluating test–retest reliability and construct validity and on previous validations of the TFI that have used sample sizes of n = 37 and n = 44 (Meikle et al, 2012; Fackrell et al, 2016a) for test–retest reliability, the sample size of 40 in the current study was considered appropriate for evaluating test–retest reliability, and convergent and divergent validity of the TFI.

Analysis

Test–retest reliability, agreement, and convergent and divergent validity analyses were carried out using the Statistical Package for the Social Sciences version 21. Test–retest reliability refers to the degree to which repeated measurements in stable individuals provide similar answers (Terwee et al, 2007). The ICC of the TFI and its subscales were computed to evaluate test–retest reliability. An ICC of >0.7 indicates an acceptable level of test–retest reliability (Lohr, 2002). R version 3.2.3 (R Core Team, 2015) was used to calculate smallest detectable change. The TFI was normally distributed (Q-Q plot lying close to 0 line; Shapiro–Wilk test [$p = 0.147$]).

A t-test was undertaken comparing the two TFI measures two weeks apart with no intervention between measurement times, a Bland–Altman plot of difference between measures as a function of mean with 95% confidence limits was plotted, and a linear regression was undertaken. The smallest detectable change was derived from the SEM between the repeated measures ($1.96 \times \sqrt{2} \times$ SEM).

Validity refers to the degree to which a questionnaire is measuring what it purports to measure. Construct validity, which involves evaluating logical relations that should exist with other questionnaires, was assessed by computing Pearson product-moment correlation

coefficients (Lohr, 2002). The cutoff levels used to define weak, moderate, strong, and very strong correlations in this study were as follows: 0–0.29 = weak, 0.30–0.59 = moderate, 0.60–0.79 = strong, and 0.80–1 = very strong (Sheskin, 2004; Shi, 2008).

To evaluate convergent validity, the total and subscale scores of the TFI were correlated with those of the THQ and rating scales, which were assumed to measure the same underlying construct. The TFI and THQ total scores were expected to correlate strongly. Strong correlations were expected between the *auditory difficulties* subscale of the TFI and *hearing ability* subscale of the THQ, as well as between the total TFI score and the *social, physical, and emotional* subscale of the THQ, as this THQ subscale contains items included in most of the TFI dimensions (*intrusive, cognitive, sleep, relaxation, quality of life, and emotional*). For this reason, moderate to strong correlations were also expected between this THQ subscale and these TFI subscales. Because the rating scales used also measured negative effects of tinnitus, moderate correlations were expected between the total TFI score and the six rating scale scores. The *intrusiveness* subscale of the TFI was expected to show moderate correlations with the *how strong/loud is your tinnitus* as well as the *how annoying is your tinnitus* rating scale, as a similar construct was measured by this TFI subscale. Similarly, a moderate correlation between the *sense of control* TFI subscale and the *how easy is it for you to ignore your tinnitus* rating scale was expected.

The HHI measures degree of psychosocial impact caused by hearing loss and was used to evaluate divergent validity of the TFI. Given that the underlying constructs measured by the TFI and HHI differ, moderate correlations were expected between the total and subscale scores of these questionnaires, indicating acceptable divergent validity. Strong correlations could be expected between the *auditory difficulties* subscale of the TFI and the total HHI score.

RESULTS

Study 1 Survey

Table 2 shows the mean, standard deviation, and range of the TFI total and subscale scores for our sample within the US 8-factor TFI structure. The mean TFI total score was 36.3 (SD = 20.8). The *sense of control* and *intrusive subscales* had the highest means: 49.9 (SD = 25.0) and 51.4 (SD = 23.4), respectively. The large standard deviations indicate large variances in the degree to which participants were affected by their tinnitus.

The TFI score was fairly evenly distributed across possible scores, but with few responses above 85 (Appendix A).

Table 2. Mean, Standard Deviation and Range of Scores of the TFI and TFI Subscales

Scale/Subscale	N	Mean	Standard Deviation	Range
TFI total score	314.00	36.30	20.80	0–94
TFI subscale scores				
Auditory	314.00	36.40	26.30	0–100
Sleep	315.00	28.80	28.40	0–100
Emotional	315.00	26.90	25.80	0–100
Cognitive	317.00	30.80	24.90	0–100
Quality of life	314.00	26.90	24.50	0–100
Intrusive	315.00	51.40	23.40	0–100
Relaxation	312.00	42.90	29.90	0–100
Sense of control	316.00	49.90	25.00	0–100

CFA of the Eight-Factor Structure

Factor Intercorrelation

The correlation between the first-order factors ranged from 0.45 (moderate, for the auditory to sleep scale) to 0.75 (strong, for the cognition to quality of life) with the average of 0.64 (Table 3). All values were within the recommended criteria of 0.3 to 0.85.

Goodness of Fit Indices

Because the estimates were adjusted by Satorra–Bentler scaled χ^2 (S–B χ^2), caution is needed when interpreting its significance because it is strongly influenced by the sample size and variability in the data. S–B χ^2 was large and significant ($\chi^2 = 591.46$, $p < 0.001$), but the χ^2/degrees of freedom ratio was only marginally higher (2.22) than the critical ratio cutoff of 2. In addition, the comparative fit index (0.95) and Tucker–Lewis index (0.95) were both acceptable, whereas the root mean square error of approximation (0.063) indicated reasonable fit. Lastly, the standardized root mean square residual value (0.054) was also considered as reasonable. Therefore, the 8-factor structure was confirmed (Figure 1), no respecification of the model was necessary.

Factor Loading Estimates and R^2

The standardized and unstandardized parameter estimates, standard error, confidence limits, and R^2 are shown in Table 4. In New Zealand, the fourth item ("Over the past week... Did you feel IN CONTROL in regard to your tinnitus?") had the lowest factor loading (0.60), but the factor loading estimates for the other items are all very high (ranged from 0.72 to 0.98) (Table 4). The fourth item also had the lowest R^2 value (0.36), but the R^2 values for the other items are all reasonably high (ranged from 0.52 and 0.97).

Table 3. Correlations between First-Order Factors in the CFA

Factor	Intrusiveness	Sense of Control	Cognitive	Sleep	Auditory	Relaxation	Quality of Life	Emotional
Intrusiveness	1.00							
Sense of control	0.69	1.00						
Cognitive	0.69	0.74	1.00					
Sleep	0.56	0.60	0.60	1.00				
Auditory	0.52	0.56	0.56	0.45	1.00			
Relaxation	0.68	0.72	0.73	0.59	0.55	1.00		
Quality of life	0.69	0.74	0.75	0.60	0.56	0.73	1.00	
Emotional	0.68	0.72	0.73	0.59	0.55	0.71	0.73	1.00

Auditory (0.65) and sleep (0.70) factors had the lowest factor loading with the second-order factor, the other first-order factors had factor loading estimates that ranged from 0.80 to 0.87. Auditory (0.42) and sleep (0.48) factors also had the lowest R^2 values, the other first-order factors had R^2 values ranged from 0.64 to 0.75. These lower estimates were higher than the estimates from the UK TFI study (Fackrell et al, 2016a). All factors contributed to the second-order global functional impact of tinnitus score (Table 5).

Summary of Factor Analysis

The item composition of each of the eight factors in the present study was found to be the same as that found by Meikle et al (2012) in the United States.

Reliability

Internal Consistency Reliability

Table 6 presents the Cronbach's alpha for the overall TFI and subscales. A Cronbach's alpha of 0.97 was produced for the TFI scale overall, indicating a high degree of consistency in responses among the items and therefore excellent internal consistency reliability. The *auditory, sleep, emotional, cognitive, quality of life,* and *relaxation* subscales of the TFI also demonstrated excellent internal consistency reliability as a Cronbach's Alpha of >0.90 was found for each of these subscales. Although the Cronbach's alpha was shown to be slightly lower for the *intrusive* and *sense of control* subscales when compared to the rest of the subscales, these two subscales nevertheless demonstrated high internal consistency reliability with Cronbach's alpha of >0.80. These subscales were also shown to have slightly lower internal consistency reliability compared to the other subscales when evaluated by Meikle et al (2012) in the United States. As shown in Table 6, the results were very similar to the results obtained by Meikle et al (2012). In the UK sample, the factor "intrusiveness" had poor internal consistency (Fackrell et al, 2016a).

Study 2 Test–Retest Reliability and Agreement

The mean, standard deviation, and range of the TFI total and subscale scores at first assessment and after two weeks are presented in Table 7. At assessment 1, the mean TFI score for the sample was found to be 47.3 (SD = 17.0). At assessment 2, the mean score was slightly lower at 44.4 (SD = 17.1), and there was no statistically significant difference between time points. The mean difference between measures was 2.92 points, and the standard error of the measurement was 1.58; thus, the smallest detectable change was 4.38. For three participants, the global TFI score was outside 95% confidence limits; two above (>22.5, higher value in test) and one below (<−16.7, higher on retest) (Appendix B). A linear regression was used to test the hypothesis that the two test periods differed from 0, the hypothesis was rejected; there was no evidence of proportional bias.

The ICCs of the TFI scale overall and TFI subscales are presented in Table 7. All subscales were above acceptable, with intrusiveness and sleep demonstrating excellent test–retest reliability As shown in Table 7, our results show higher test–retest reliability compared with the results obtained by Meikle et al (2012) in the United States, and slightly poorer than Fackrell et al (2016a) in the United Kingdom. Whereas test–retest correlations of the TFI overall score and all eight subscale scores reach the 0.7 criterion in the United Kingdom, the US data show that the *cognitive, quality of life*, and *relaxation* subscales have correlations lower than 0.7.

Validity

Convergent Validity

A statistically significant strong correlation between the TFI overall score and the overall score of the THQ was found, $r(38) = 0.717$, $p < 0.01$ (Appendix B). The TFI overall score and the *emotional distress* subscale of the TFI both correlated strongly with *the social, physical, and emotional* subscale of the THQ, $r(38) = 0.701$, and 0.649, respectively, $p < 0.01$. The *quality of life*
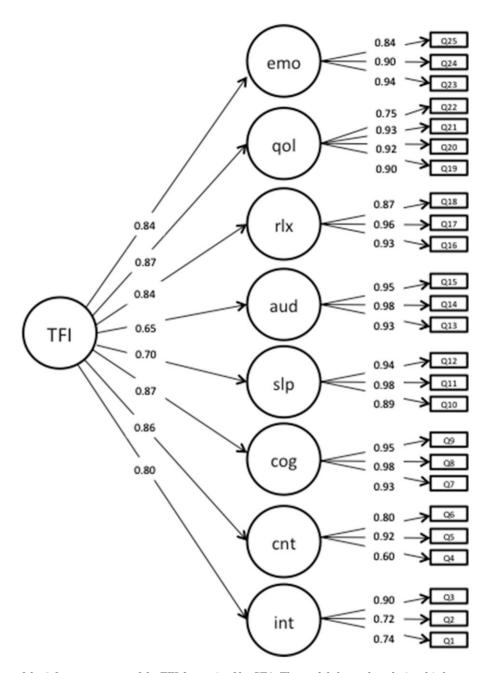
Figure 1. Diagram of the 8-factor structure of the TFI determined by CFA. The model shows the relationship between observed variables (Q1–Q25), the first-order factors (1–8, emo [emotion], qol [quality of life], rlx [relaxation], aud [auditory], slp [sleep], cog [cognition], cnt [control], int [intrusiveness]) and second-order factor (TFI, overall TFI score). The numbers are standardized parameter estimates relating items.

subscale of the TFI correlated strongly with both the *social, physical, and emotional* subscale of the THQ, $r(38) = 0.613$, $p < 0.01$, and the overall THQ score, $r(38) = 0.665$, $p < 0.01$. The *auditory difficulties* subscale of the TFI and the *hearing difficulties* subscale of the THQ were also shown to be strongly correlated, $r(38) = 0.778$, $p < 0.01$. Statistically significant moderate correlations were also observed between several of the other TFI and THQ overall and subscale scores. Each of the eight TFI subscales showed statistically signifi-

cant moderate or strong correlations with at-least two of the THQ scores (overall or subscale). The TFI overall score and the *intrusiveness, cognitive interference, auditory difficulties*, and *quality of life* subscales were shown to have statistically significant moderate or strong correlations with the THQ overall score, the *social, physical, and emotional* subscale, and the *hearing difficulties* subscale (Appendix C).

With regard to the six rating scales, a statistically significant strong correlation was only observed between

Table 4. The Standardized (Std) and Unstandardized (Unstd) Parameter Estimates, Standard Error (SE), Confidence Limits (CL), and R^2 for the CFA of Items Grouped in the First-Order Factors

Factor	Question	Std	Unstd	SE	CL Lower	CL Upper	R^2
Intrusiveness	Q1	0.74	1.00	0.00	1.00	1.00	0.55
Intrusiveness	Q2	0.72	0.61	0.05	0.51	0.71	0.52
Intrusiveness	Q3	0.90	1.10	0.08	0.95	1.24	0.81
Control	Q4	0.60	1.00	0.00	1.00	1.00	**0.36**
Control	Q5	0.92	1.11	0.10	0.92	1.30	0.85
Control	Q6	0.80	1.11	0.10	0.91	1.31	0.63
Cognitive	Q7	0.93	1.00	0.00	1.00	1.00	0.87
Cognitive	Q8	0.98	1.03	0.03	0.97	1.08	0.95
Cognitive	Q9	0.95	0.97	0.03	0.92	1.03	0.90
Sleep	Q10	0.89	1.00	0.00	1.00	1.00	0.80
Sleep	Q11	0.98	1.10	0.04	1.03	1.17	0.95
Sleep	Q12	0.94	1.03	0.04	0.95	1.10	0.88
Auditory	Q13	0.93	1.00	0.00	1.00	1.00	0.87
Auditory	Q14	0.98	1.04	0.03	0.99	1.09	0.97
Auditory	Q15	0.95	1.06	0.03	1.00	1.12	0.91
Relaxation	Q16	0.93	1.00	0.00	1.00	1.00	0.87
Relaxation	Q17	0.96	0.99	0.03	0.93	1.05	0.91
Relaxation	Q18	0.87	1.02	0.04	0.94	1.10	0.75
Quality of Life	Q19	0.90	1.00	0.00	1.00	1.00	0.82
Quality of Life	Q20	0.92	1.06	0.04	0.98	1.14	0.84
Quality of Life	Q21	0.93	1.06	0.04	0.98	1.13	0.87
Quality of Life	Q22	0.75	0.78	0.05	0.69	0.87	0.56
Emotional	Q23	0.94	1.00	0.00	1.00	1.00	0.88
Emotional	Q24	0.90	1.00	0.04	0.93	1.08	0.82
Emotional	Q25	0.84	0.87	0.04	0.79	0.95	0.70

Notes: The R^2 value in Bold is just below the recommended cutoff of <0.4 for association with the factor "control."

the *sense of control* subscale of the TFI and the *ignore* rating scale, $r(38) = 0.649, p < 0.01$. There were several statistically significant moderate correlations between the TFI and rating scale scores. The TFI overall score as well as the *intrusiveness* and *emotional distress* subscales of the TFI were shown to have moderate correlations with all six of the rating scales. The *quality of life* subscale of the TFI correlated moderately with all rating scale except for the *annoyed* scale, $r(38) = 0.370$ to $0.537, p < 0.05$. The *sense of control* subscale of the TFI correlated moderately with the *problem, annoyed,* and *unpleasant* scales of the rating scale, $r(38) = 0.319$ to $0.372, p < 0.05$. Moderate correlations were observed between the *relaxation* subscale of the TFI and the *uncomfortable, annoyed,* and *unpleasant* scales of the rating scale, $r(38) = 0.422$ to $0.462, p < 0.01$. The *sleep disturbance* subscale of the TFI correlated moderately with the *uncomfortable* and *unpleasant* scales of the rating scale, $r(38) = 0.351$ and 0.362, respectively, $p < 0.05$. The *cognitive interference* subscale of the TFI was shown to have a moderate correlation with only the *problem* scale of the rating scale, $r(38) = 0.449, p < 0.01$. The *auditory difficulties* subscale of the TFI was not shown to have statistically significant correlations with any of the rating scales.

Table 5. The Standardized (Std) and Unstandardized (Unstd) Parameter Estimates, SE, CL, and R^2 for the CFA for the First-Order Factors onto the Second-Order Factor (Total TFI Score)

Factor	Std	Unstd	SE	CL Lower	CL Upper	R^2
Intrusiveness	0.80	1.87	0.16	1.55	2.18	0.64
Control	0.86	1.77	0.17	1.43	2.11	0.74
Cognition	0.87	2.13	0.12	1.88	2.37	0.75
Sleep	0.70	1.86	0.15	1.57	2.15	0.48
Auditory	0.65	1.63	0.14	1.36	1.89	0.42
Relaxation	0.84	2.41	0.15	2.12	2.70	0.71
Quality of life	0.87	2.10	0.13	1.85	2.35	0.75
Emotion	0.84	2.18	0.13	1.92	2.44	0.71

Table 6. Cronbach's Alpha and ICCs of the TFI and its Subscales in New Zealand, the United States, and the United Kingdom

Scale/Subscale	Cronbach's Alpha			ICC		
	Present Study	United States (Meikle et al, 2012)	United Kingdom (Fackrell et al, 2016a)	Present Study	United States (Meikle et al, 2012)	United Kingdom (Fackrell et al, 2016a)
TFI	0.97	0.97	0.80	0.91	0.78	0.91
TFI subscales						
Auditory	0.97	0.97	0.95	0.88	0.90	0.95
Sleep	0.95	0.97	0.94	0.93	0.78	0.91
Emotional	0.93	0.94	0.91	0.84	0.76	0.87
Cognitive	0.97	0.96	0.95	0.83	0.66	0.89
Quality of life	0.93	0.93	0.90	0.86	0.63	0.86
Intrusive	0.82	0.85	**0.58**	0.90	0.83	0.92
Relaxation	0.94	0.96	0.93	0.76	0.67	0.83
Sense of control	0.80	0.82	0.75	0.77	0.75	0.81

Notes: Values in bold indicate poor internal consistency (alpha < 0.7).

Divergent Validity

No statistically significant strong correlations were found between the TFI and HHI scores. The TFI overall score correlated moderately with the overall score of the HHI, $r(38) = 0.394$, $p < 0.05$. The TFI overall score as well as the *auditory* and *quality of life* subscales of the TFI were each shown to correlate moderately with each of the two HHI subscales as well as the HHI overall score. However, no statistically significant correlations were observed between the remaining TFI and HHI subscale or overall scores.

Score Distribution

Response distributions for each subscale of the TFI for the clinic population (Study 1) and research group (Study 2) are shown in Figure 2. The distribution pattern across the subscales is similar for both study populations. The Study 2 population median is higher because of an inclusion criterion of a TFI overall score of at least 25. In the general clinic sample, the scores were more widely distributed in each subscale, reflect-

ing a more heterogeneous group. The percentage of persons with a score of 0 (minimal problem) ranged from 0.3% (intrusiveness) to 21% (sleep). Scores of 10 (maximum problem) ranged from 0% (emotion) to 2.5% (relaxation).

DISCUSSION

This study aimed to determine if the TFI was a reliable and valid measure of the negative impact and severity of tinnitus in a New Zealand clinic population, its responsiveness to treatment in a research trial, and to find out if any modifications needed to be made to the questionnaire. The CFA of the TFI with extraction of eight factors revealed the same underlying dimensions of tinnitus functional impact as those in the original questionnaire. Overall, the TFI appears a reliable and valid measure of tinnitus severity and negative impact, and does not need to be modified for use in our New Zealand patients with tinnitus. The distribution of results for our study and several other studies in the United States and the United Kingdom show that depending on the population in which the questionnaire

Table 7. Mean, Standard Deviation, and Range of Scores of the TFI and TFI Subscales at Assessments 1 and 2

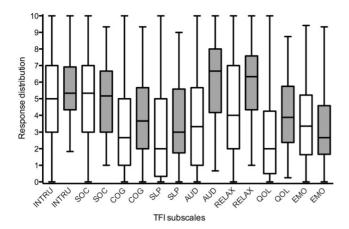| Scale/Subscale	Assessment 1			Assessment 2		
	Mean	Standard Deviation	Range	Mean	Standard Deviation	Range
TFI total score	47.30	17.00	25–86	44.40	17.10	15.6–84.8
TFI subscale scores						
Auditory	62.00	25.50	6.7–100	54.00	21.40	0–93.3
Sleep	37.90	25.60	0–90	37.70	26.50	0–90
Emotional	34.80	25.20	0–93.3	32.50	24.03	0–90
Cognitive	40.30	24.30	0–93.3	38.60	23.30	0–80
Quality of life	41.90	23.20	2.5–87.5	35.58	22.28	7.5–87.5
Intrusive	54.90	18.80	18.33–100	54.60	19.70	16.67–96.7
Relaxation	59.10	24.50	10–100	51.80	23.40	0–90
Sense of control	49.70	22.90	10–93.3	53.30	20.50	20–90

Figure 2. Response frequency distributions for each subscale of the TFI for Study 1 (white boxes) and Study 2 (gray boxes). Subscales of TFI: INTRU = intrusiveness; SOC = sense of control; COG = cognition; SLP = sleep; AUD = auditory; REL = relaxation; QOL = quality of life; EMO = emotional.

is used, score distribution will be different (Table 8); for example, our Study 2 trial participants were included only if their initial screening TFI was >25. Our clinic population may include people who have found ways of managing their tinnitus, accounting for the reasonably high number of survey respondents reporting low scores on the sleep subscale, or the sleep subscale may not be as responsive to change in this particular population.

The New Zealand results differ from those from the United Kingdom (Fackrell et al, 2016a) in that all first-order factor intercorrelations were within criteria. In the United Kingdom, the auditory and emotional factors were too weakly or too strongly correlated with other factors. In our population, question 4 "Over the past week... Did you feel IN CONTROL in regard to your tinnitus?" had the lowest factor loading and correlation with the first-order factor "Sense of Control." This was also found in the United Kingdom. None of the second-order factors relationship with the second-order factor were outside criteria in our results, whereas in the United Kingdom, factors 4 (sleep) and 5 (auditory) had low correlations (including in their respecified model). An exploratory factor analysis found the 8-factor struc-

ture suitable for the Dutch version of the TFI (Rabau et al, 2014).

The factor structure of the TFI in the New Zealand clinic potentially did not change from the original because the population sampled and culture of participants was similar to that of the United States (a high proportion of New Zealand Europeans [Caucasian] in the sample, over 90%). The demographic characteristics of the sample used for factor analysis in the present study adequately represented the New Zealand population with tinnitus as the prevalence of tinnitus is shown to be higher in men, older adults, and New Zealand Europeans (Wu et al, 2015); however, the results may have been different if the sample represented the New Zealand general population, with European 74.0%, Māori 14.9%, Asian 11.8%, and Pacific people 7.4% (Statistics New Zealand, 2015). A validation of a generic health-related quality of life measure, the Short Form Health Survey (SF-36; Scott et al, 2000) showed a similar factor structure in New Zealand European to the United States but a substantially different structure in both the Māori and Pacific groups when compared with the United States.

Excellent internal consistency reliability was found for the TFI, as indicated by a Cronbach alpha of 0.97. This signifies a high degree of consistency in responses among the items, indicating that the items all measured the same underlying construct. The subscales were also highly internally consistent. Some experts have suggested a Cronbach's alpha greater that 0.90 or 0.95 to indicate redundancy of items within a scale. Terwee et al (2007) proposed a criterion of 0.70–0.95 as a measure of good internal consistency. Based on this, the results of this study indicate redundancy in the structure of the TFI, specifically for the TFI overall score and the *auditory difficulties* and *cognitive* subscales. Fackrell et al (2016a) suggested that future studies examine the effect of removing the auditory subscale and using it separately.

For a test–retest period of two weeks, the TFI demonstrated excellent test–retest reliability with an ICC of 0.91 ($p < 0.01$). This shows that on average, there was little variance in the TFI scores within participants over the two administrations; thus, the TFI is able to

Table 8. Percentage of Respondents in Each Tinnitus Problem Category for Four Previous Evaluations of the TFI and the Two Datasets Evaluated in the Current Research

Study	Not Problem (0–17)	Small Problem (18–31)	Moderate Problem (32–53)	Big Problem (54–72)	Very Big Problem (73–100)
Meikle et al (2012)	9%	14%	22%	28%	28%
Henry, Griest, et al (2016)	0%	3%	14%	41%	42%
Henry, Griest, et al (2016)	5%	9%	29%	31%	27%
Fackrell et al (2016a)	12%	27%	31%	24%	5%
Study 1 (clinic survey)	20%	26%	32%	15%	6%
Study 2 (randomized controlled trial)	0%	20%	55%	12.50%	12.50%

produce stable scores over time. Test–retest reliability of the subscales was also good. Internal consistency reliability of the TFI and its subscales was similar to that found in the United States. Test–retest reliability was generally higher than the United States and slightly lower than the United Kingdom. The responsiveness of the TFI to treatment-related change was evaluated by examining change in baseline test–retest scores (Shekhawat et al, 2014). The smallest detectable change was determined from the variance in these measures and was found to be 4.8. We recommend that in a population similar to that tested here, a change of at least 4.8 points in the TFI be considered as a guide to whether change seen was a true change. The results are very different from a UK research sample recommendation indicating a 23-point change (Fackrell et al, 2016a). This difference may reflect the population included in the different samples (Henry, Thielman, et al, 2016). The TFI appears to be a responsive measure in our population with high test–retest reliability and good agreement of the TFI; the TFI is able to produce stable scores over time. Any documented changes in tinnitus scores after treatment are thus likely to be due to the treatment, rather than error.

Strong convergent and moderate divergent validity of the TFI is evident from the pattern of correlations reported here, a higher number of strong and moderate statistically significant correlations were observed with measures used for the same purpose (THQ and rating scales) compared with a measure used for a different purpose (HHI). This provides evidence that the TFI measures what it was intended to. Convergent validity of the TFI was evidenced by a strong correlation between the overall TFI and THQ scores, and between some of the subscale scores of these questionnaires. The overall TFI and THQ scores were also reported to correlate strongly by Fackrell et al (2016a) in the United Kingdom. Divergent validity of the TFI was evidenced by a small number of moderate statistically significant correlations with the HHI overall and subscale scores. The TFI overall score, and the *auditory difficulties* and *quality of life* subscale scores correlated moderately with the HHI overall and subscale scores. This is likely because the *auditory difficulties* subscale of the TFI includes questions that relate to hearing loss, such as patients' ability to hear clearly, understand people who are talking, and follow conversations in a group or at meetings. The *quality of life* subscale assesses the extent to which social activities, relationships with people, enjoyment of life, and work are affected by tinnitus; hearing loss can also affect these. Tyler et al (2014) suggested that a potential shortcoming with the TFI is that it includes questions related to "secondary" effects of tinnitus (e.g., "quality of life" and "relationships"). The TPFQ developed by Tyler et al (2014) includes hearing as a primary effect, whereas quality of life would be a secondary effect. Debate of what constitutes direct and indirect effects of tinnitus, and whether they should be included in questionnaires, needs to occur.

Tyler et al (2014) also believe that the 100-point scale used in the TPFQ offers greater resolution over the 10-point scale used in the TFI. Both Fackrell et al (2016a) and our study found that responses to the TFI were fairly evenly distributed across response options, although no one in our large sample responded with a score above 90 (Appendix A). Fackrell et al (2016a) found few high scores to the THQ (a 100 response option questionnaire); the TFI results were more evenly distributed than the THQ or THI. The relative merits of the TFI and TPFQ regarding both resolution and responsiveness need to be ascertained. Although it may result in some redundancy, the use of both questionnaires is an appropriate strategy in clinical trials. We believe that the TFI should be included, along with other measures, as part of a standard tinnitus assessment battery (Langguth et al, 2011).

Study 1 had a response rate to the survey of 36.5%; such a low sample can result in nonresponse bias and may mean the results are misleading. We used secondary data and ethical approval that did not allow follow-up contact of nonresponders. However, we have not found any evidence in the literature that indicates that individuals with tinnitus who respond to psychometric scales such as the TFI differ in any significant way to individuals who choose not to respond. Vernon et al (1992) sent follow-up questionnaires to patients with tinnitus to determine if recommendations had been followed. There were no statistically significant differences between those who returned these questionnaires without any additional inquiry (classified as responders) and those who did not respond until repeated efforts were made to reach them (classified as nonresponders). Although this study did not use a psychometric scale measuring tinnitus severity, it suggests that responders and nonresponders of questionnaires do not necessarily differ. Future surveys should employ methods facilitating a greater proportion of population being studied to be captured. Overall, we do not believe our response rate has impacted negatively on our results.

Although the primary aim of the present study was to evaluate the psychometric adequacy of the TFI specifically for use in our clinic and research, this study provides further evidence, that the 25-item TFI has adequate psychometric properties and is a reliable and valid measure of the negative impact and severity of tinnitus (Meikle et al, 2012 ; Henry, Griest, et al, 2016), and in most studies, it is sensitive to treatment effects (Meikle et al, 2012; Henry, Griest, et al, 2016).

CONCLUSION

The TFI is a reliable and valid measure of tinnitus severity in New Zealand, and has the same 8-factor structure as the original questionnaire developed in the United States. The questionnaire has excellent

test–retest reliability and acceptable agreement with no proportional bias. The TFI should, on the basis of test–retest variance, be sensitive to treatment. We recommend a criterion of 4.8-point change in TFI as indicating a true change in populations similar to ours. We believe that the use of a comprehensive tinnitus measure is essential. Given its psychometric properties, the TFI has the potential to be used as a standard measure for intake and outcome assessment. The TFI should be validated and its factor structure examined in other countries and cultures; this may aid progress in finding effective evidence-based treatments for tinnitus (Meikle et al, 2012).

REFERENCES

Bentler PM. (1990) Comparative fit indexes in structural models. *Psychol Bull* 107(2):238–246.

Downey RG, King C. (1998) Missing data in Likert ratings: a comparison of replacement methods. *J Gen Psychol* 125(2):175–191.

Fackrell K, Hall DA, Barry JG, Hoare DJ. (2016a) Psychometric properties of the Tinnitus Functional Index (TFI): assessment in a UK research volunteer population. *Hear Res* 335:220–235.

Fackrell K, Hall DA, Barry JG, Hoare DJ. (2016b) Response to letter: psychometric properties of the Tinnitus Functional Index (TFI): assessment in a UK research volunteer population. *Hear Res* 335:237–238.

Folmer RL. (2016) Reply to: psychometric properties of the Tinnitus Functional Index (TFI): assessment in a UK research volunteer population. *Hear Res* 335:236.

Henry JA, Griest S, Thielman E, McMillan G, Kaelin C, Carlson KF. (2016) Tinnitus functional index: development, validation, outcomes research, and clinical application. *Hear Res* 334:58–64.

Henry JA, Thielman E, Zaugg T. (2016) Reply to: Psychometric properties of the Tinnitus Functional Index (TFI): Assessment in a UK research volunteer population. *Hear Res* 335:220–235.

Hoffman HJ, Reed GW. (2004) Chapter three: epidemiology of tinnitus. In: Snow JB, ed. *Tinnitus: Theory and Management*. New York, United States: BC Decker Inc.

Holmes S, Padgham ND. (2009) Review paper: more than ringing in the ears: a review of tinnitus and its psychosocial impact. *J Clin Nurs* 18(21):2927–2937.

Holmes S, Padgham ND. (2011) "Ringing in the ears": narrative review of tinnitus and its impact. *Biol Res Nurs* 13(1):97–108.

Hu LT, Bentler PM. (1998) Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychol Methods* 3:424.

Kamalski DM, Hoekstra CE, van Zanten BG, Grolman W, Rovers MM. (2010) Measuring disease-specific health-related quality of life to evaluate treatment outcomes in tinnitus patients: a systematic review. *Otolaryngol Head Neck Surg* 143(2):181–185.

Kuk FK, Tyler RS, Russell D, Jordan H. (1990) The psychometric properties of a tinnitus handicap questionnaire. *Ear Hear* 11(6):434–445.

Landgrebe M, Azevedo A, Baguley D, et al. (2012) Methodological aspects of clinical trials in tinnitus: a proposal for an international standard. *J Psychosom Res* 73(2):112–121.

Langguth B. (2011) A review of tinnitus symptoms beyond 'ringing in the ears': a call to action. *Curr Med Res Opin* 27(8):1635–1643.

Langguth B, Goodey R, Azevedo A, et al. (2007) Consensus for tinnitus patient assessment and treatment outcome measurement: Tinnitus Research Initiative meeting, Regensburg, July 2006. *Prog Brain Res* 166:525–536.

Langguth B, Searchfield G, Biesinger E, Greimel K. (2011) History and Questionnaires. In: Moller A, Langguth B, Ridder DD, Kleinjung T, eds. *Textbook of Tinnitus*. New York: Springer, 387–404.

Lohr KN. (2002) Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research* 11(3):193–205.

Meikle MB. (2002) A conceptual framework to aid the diagnosis and treatment of severe tinnitus. *Aust NZ J Audiol* 24:59–67.

Meikle MB, Henry JA, Griest SE, et al. (2012) The tinnitus functional index: development of a new clinical measure for chronic, intrusive tinnitus. *Ear Hear* 33(2):153–176.

Meikle MB, Stewart BJ, Griest SE, Henry JA. (2008) Tinnitus outcomes assessment. *Trends Amplif* 12(3):223–235.

Meikle MB, Stewart BJ, Griest SE, Martin WH, Henry JA, Abrams HB, McArdle R, Newman CW, Sandridge SA. (2007) Assessment of tinnitus: measurement of treatment outcomes. *Prog Brain Res* 166:511–521.

Møller AR. (2011) Introduction. In: Moller A, Langguth B, DeRidder D, Kleinjung T, eds. *Textbook of Tinnitus*. New York, NY: Springer, 3–7.

Newman CW, Sandridge SA. (2004) Tinnitus Questionnaires. In: Snow JB, ed. *Tinnitus: Theory and Management*. New York: BC Decker Inc, 237–254.

Newman CW, Sandridge SA, Jacobson GP. (1998) Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcome. *J Am Acad Audiol* 9(2):153–160.

Newman CW, Weinstein BE, Jacobson GP, Hug GA. (1990) The hearing handicap inventory for adults: psychometric adequacy and audiometric correlates. *Ear Hear* 11(6):430–433.

Rabau S, Wouters K, Van de Heyning P. (2014) Validation and translation of the Dutch tinnitus functional index. *B-ENT* 10(4):251–258.

R Core Team. (2015) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosseel Y. (2012) lavaan: an R Package for Structural Equation Modeling. *J Stat Softw* 48(2):1–36 http://www.jstatsoft.org/v48/i02/.

Sanchez L, Stephens D. (1997) A tinnitus problem questionnaire in a clinic population. *Ear Hear* 18(3):210–217.

Satorra A, Bentler PM. (1994) Corrections in test statistics and standard errors in covariance structure analysis. In: von Eye AE, Clogg CC, eds. *Latent Variables Analysis: Applications for Developmental Research*. Thousand Oaks, CA: Sage Publications Inc., 399–419.

Scott KM, Sarfati D, Tobias MI, Haslett SJ. (2000) A challenge to the cross-cultural validity of the SF-36 health survey: factor

structure in Maori, Pacific and New Zealand European ethnic groups. *Soc Sci Med* 51(11):1655–1664.

Searchfield GD. (2014) Tinnitus what and where: an ecological framework. *Front Neurol* 5:271.

Searchfield GD, Jerram C. (2010) Tinnitus assessment. *N Z Med J* 123:101–111.

Searchfield GD, Jerram C, Wise K, Raymond S. (2007) The impact of hearing loss on tinnitus severity. *Aust NZ J Audiol* 29:67–76.

Shekhawat GS, Searchfield GD, Stinear CM. (2014) Randomized trial of transcranial direct current stimulation and hearing aids for tinnitus management. *Neurorehabil Neural Repair* 28(5): 410–419.

Sheskin DJ. (2004) *Test 28: the Pearson Product-Moment Correlation Coefficient. Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd ed. Boca Raton: CRC Press, 945–1025.

Shi L. (2008a) *Measurements in Health Services Research. Health Services Research Methods*. 2nd ed. New York: Delmar Learning, 288–322.

Statistics New Zealand. (2015) National Population Estimates. http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections.aspx. Accessed June 30, 2015.

Steiger JH, Lind JC. (1980) Statistically based tests for the number of common factors. *Annual meeting of the Psychometric Society, Iowa City, IA*.

Terwee CB, Bot SD, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 60(1):34–42.

Tucker LR, Lewis C. (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38:1–10.

Tunkel DE, Bauer CA, Sun GH, et al. (2014) Clinical practice guideline: tinnitus. *Otolaryngol Head Neck Surg* 151:S1–S40.

Tyler RS, Baker LJ. (1983) Difficulties experienced by tinnitus sufferers. *J Speech Hear Disord* 48(2):150–154.

Tyler R, Ji H, Perreau A, Witt S, Noble W, Coelho C. (2014) Development and validation of the tinnitus primary function questionnaire. *Am J Audiol* 23(3):260–272.
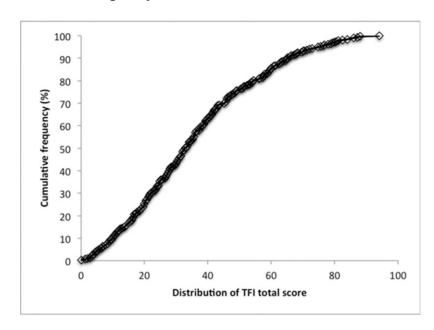
Vernon J, Griest S, Press L. (1992) Plight of unreturned tinnitus questionnaires. *Brit J Audiol* 26(2):137–138.

Yong AG, Pearce S. (2013) A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutor Quant Methods Psychol* 9:79–94.
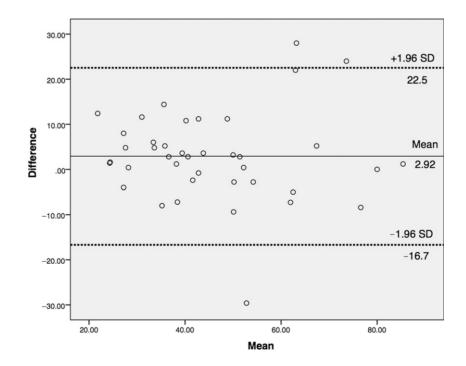
Yu CH. (2001) An introduction to computing and interpreting Cronbach Coefficient Alpha in SAS. *Proceedings of 26th SAS User Group International Conference*. 22–25.

Wu BP, Searchfield G, Exeter DJ, Lee A. (2015) Tinnitus prevalence in New Zealand. *N Z Med J* 128(1423):24–34.

APPENDIX A. Cumulative frequency distribution of the TFI for individual data from study 1.

.

APPENDIX B. Bland Altman Plot. The symbols represent individual results for the difference between test and retest measures as a function of the mean value of the TFI. The solid horizontal line is the mean difference and the dashed lines are 95% confidence intervals.

.

APPENDIX C. Correlation Matrix for Questionnaires and Scales Used in this Study

Shading legend:
- 0.0–0.29 Weak
- 0.3–0.59 Moderate
- 0.6–0.79 Strong
- 0.8–1 Extremely Strong

	problem	strong	uncomf	annoy	ignore	unpl	overall	intru	cont	cogn	sleep	aud	relax	QOL	emo	soc, phy, emo	hear	view	overall	emo	soc	overall
R Scales problem	1																					
strong	0.683**	1																				
uncomf	0.690**	0.757**	1																			
annoy	0.603**	0.535**	0.770**	1																		
ignore	0.368*	0.311	0.431**	0.417**	1																	
unpleas	0.710**	0.704**	0.894**	0.812**	0.547**	1																
TFI overall	0.542**	0.351	0.454**	0.441**	0.391*	0.508**	1															
intrusive	0.485**	0.393*	0.443*	0.380*	0.384**	0.488**	0.795**	1														
control	0.319*	0.157	0.248	0.326*	0.649**	0.372*	0.676**	0.415**	1													
cognitive	0.449*	0.175	0.178	0.243	0.164	0.242	0.818**	0.540**	0.528**	1												
sleep	0.298	0.146	0.351*	0.287	0.104	0.362*	0.597**	0.459**	0.315*	0.398*	1											
auditory	0.223	0.202	0.116	0.045	−0.049	0.066	0.556**	0.405**	0.163	0.493**	0.131	1										
relax	0.293	0.245	0.461**	0.462**	0.244	0.422**	0.687**	0.539**	0.421**	0.459**	0.432**	0.183	1									
QOL	0.537**	0.370*	0.382*	0.263	0.380*	0.477**	0.665**	0.665**	0.528**	0.634**	0.303	0.528**	0.423**	1								
emotion	0.495**	0.321**	0.431**	0.556**	0.410**	0.486**	0.780**	0.619**	0.515**	0.627**	0.396**	0.236	0.527**	0.567**	1							
THQ soc, phy, emo	0.777**	0.651**	0.586**	0.586**	0.461**	0.654**	0.701**	0.434**	0.434**	0.541**	0.434**	0.404**	0.387**	0.567**	0.567**	1						
hearing	0.398*	0.391*	0.450**	0.342*	0.031	0.346	0.515**	0.395	0.081	0.345	0.217	0.404**	0.25	0.532**	0.532**	0.633**	1					
view	0.24	0.169	0.24	0.121	0.285	0.284	0.513**	0.377	0.285	0.365*	0.201	0.328	0.335*	0.532**	0.297	0.432**	0.326*	1				
overall	0.699**	0.484**	0.631**	0.536**	0.357*	0.601**	0.717**	0.477**	0.433**	0.532**	0.395	0.587**	0.389	0.665**	0.581**	0.955**	0.822**	0.517**	1			
HHI emotion	0.295	0.118	0.126	0.108	0.115	0.159	0.435*	0.312	0.178	0.269	0.194	0.424**	0.224	0.528**	0.309	0.430**	0.495**	0.429**	0.515**	1		
social	0.209	0.04	0.045	−0.013	0.127	0.043	0.318*	0.221	0.11	0.153	0.082	0.419**	0.183	0.459**	0.141	0.373*	0.543**	0.428**	0.493**	0.850**	1	
overall	0.264	0.084	0.09	0.052	0.126	0.108	0.394*	0.279	0.151	0.222	0.146	0.438**	0.213	0.514**	0.237	0.419**	0.538**	0.446**	0.524**	0.965**	0.959**	1

Column groups (bottom axis): Rating scales — problem, strong, uncomf, annoy, ignore, unpl; TFI — overall, intru, cont, cogn, sleep, aud, relax, QOL, emo; THQ — soc/phy/emo, hear, view, overall; HHI — emo, soc, overall.

Notes: Stronger correlations are indicated with darker shading, $*p < 0.01$, $**p < 0.001$.