AJA AMERICAN JOURNAL OF AUDIOLOGY

How to choose between measures of tinnitus loudness for clinical research? A report on the reliability and validity of an investigator-administered test and a patient-reported measure using baseline data collected in a phase IIa drug trial.

Journal:	*The American Journal of Audiology*
Manuscript ID	AJA-16-0129.R2
Manuscript Type:	Research Note
Date Submitted by the Author:	n/a
Complete List of Authors:	Hall, Deborah; National Institute for Health Research, Nottingham Biomedical Research Centre; University of Nottingham, Otology and Hearing Group, Division of Clinical Neuroscience, School of Medicine Mehta, Rajnikant; National Institute for Health Research, Nottingham Biomedical Research Centre Fackrell, Kathryn; National Institute for Health Research, Nottingham Biomedical Research Centre; University of Nottingham, Otology and Hearing Group, Division of Clinical Neuroscience, School of Medicine
Keywords:	Tinnitus, Adults, Outcomes

Title: How to choose between measures of tinnitus loudness for clinical research? A report on the reliability and validity of an investigator-administered test and a patient-reported measure using baseline data collected in a phase IIa drug trial.

Running title: Properties of tinnitus loudness measures

*Deborah A. Hall [1,2], Rajnikant L. Mehta [1,2], Kathryn Fackrell [1,2]

1. National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre

2. Otology and Hearing Group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham

* Corresponding author

NIHR Nottingham Biomedical Research Centre

Ropewalk House, 113 The Ropewalk

Nottingham

NG1 5DU, UK

deborah.hall@nottingham.ac.uk

+44 (0) 115 8232600

ABSTRACT

Purpose: Loudness is a major auditory dimension of tinnitus, and is used to diagnose severity, counsel patients or as a measure of clinical efficacy in audiological research. There is no standard test for tinnitus loudness, but matching and rating methods are popular. This article provides important new knowledge about the reliability and validity of an audiologist-administered tinnitus loudness matching test and a patient-reported tinnitus loudness rating.

Method: Retrospective analysis of loudness data for 91 participants with stable subjective tinnitus enrolled in a randomised controlled trial of a novel drug for tinnitus. There were two baseline assessments (Screening, Day1) and a post-treatment assessment (Day28).

Results: About 66-70% of the variability from Screening to Day1 was attributable to the true score. But measurement error, indicated by the Smallest Detectable Change, was high for both tinnitus loudness matching (20 dB) and tinnitus loudness rating (3.5 units). Only loudness rating captured a sensation that was meaningful to people with the lived experience of tinnitus.

Conclusions: The tinnitus loudness rating performed better against acceptability criteria for reliability and validity than did the tinnitus loudness matching test administered by an audiologist. But the rating question is still limited because it is a single-item instrument and is probably able to detect only large changes (at least 3.5 points).

INTRODUCTION

Tinnitus has auditory, psychological and social dimensions. Psychoacoustic testing of tinnitus by otologists and audiologists typically includes tinnitus loudness and pitch matching, minimum masking levels, and residual inhibition testing (Tunkel et al., 2014). Tinnitus loudness is an important auditory dimension, notably for evaluating patient benefit from tinnitus treatment perhaps more so than for diagnosing tinnitus severity (Tunkel et al., 2014). For example, a systematic review of 228 tinnitus intervention trials found that 14% of clinical studies chose loudness as a primary outcome for assessing therapeutic benefit (Hall et al., 2016). Unfortunately currently there is no standard test for measuring tinnitus loudness. Indeed, the same review by Hall and colleagues (2016) identified a diverse range of instruments in use for measuring tinnitus loudness; not only those mentioned by Tunkel et al (2014) but also loudness discomfort level, Visual Analogue Scales and other numerical rating scales. These instruments are used by clinicians with the understanding that they all measure the same construct (i.e. that they have convergent validity).

The American Academy of Otolaryngology—Head and Neck Surgery Foundation (AAO-HNSF) clinical practice guideline recommended that further research is needed to determine a reliable assessment of perceived tinnitus loudness (Tunkel et al., 2014). Knowledge about the statistical properties of each method can be highly informative for helping clinicians to choose between different assessments (Prinsen et al., 2016), although these properties are not necessarily fixed across populations (Feldt & Brennan, 1989). Reliability and validity are two major properties that should be measured to improve the clinical relevance of instruments to be used for evaluative purposes. A number of studies do report reliability of tinnitus loudness matching, but in some studies the main question focussed on comparisons of procedural variation in tests (Henry et al., 1999) or on assessing a new instrumentation (Henry et al., 2006), and in other studies those measures were repeated across multiple test frequencies which is time-consuming (Hoare et al., 2014; Mitchell et al., 1993). Studies have not yet asked questions about reliability of tinnitus loudness measures that are feasible to conduct in a routine clinical setting or in a clinical trial. In those contexts, considerations of time, expertise and specialised instrumentation also inform the choice of measure.

This Research Note addresses that gap by presenting evidence on the reliability and validity of two tinnitus loudness measures that are both feasible to conduct in clinic. The two measures were a audiologist-administered test (a tinnitus loudness matching method) and a patient-reported measure (a numerical rating scale of loudness). Tinnitus loudness matching and a numerical rating scale from 0 to 10 are both very common in clinical research (Hall et al., 2016). A particular strength of our approach was that our statistical analysis was informed by an international standard for evaluating the psychometric properties of instruments for clinical research (Mokkink et al., 2010; Terwee et al., 2007).

In addition to this, we examine whether the tinnitus loudness matching expressed in units of dB Hearing Level (HL) or dB Sensation Level (SL) contributes any additional error in the measurement (see Henry et al., 1999). While dB HL reflects the level of the tone that is perceived as having the same loudness as the tinnitus (taken from the audiometer scale), dB SL is the difference between that measure and the hearing threshold at the same frequency. Loudness expressed in dB SL thus introduces another source of within-subject variability which may affect overall reliability of the method. For example, Henry and colleagues (1999) concluded that dB SL was less reliable, but their dependence on the Pearson's correlation coefficient warrants further investigation because this statistic does not take systematic error into account (Mokkink et al., 2010). It is clear from this brief summary of the literature, that there is little evidence available to help clinical researchers identify the most appropriate measure of tinnitus loudness that is valid and practical to use.

METHODS

This was a retrospective analysis of tinnitus loudness data collected as part of a randomised, placebo-controlled, blinded, phase IIa trial assessing 28-days 800mg once-daily dosing of a new medicine (AUT00063) (QUIET-1, ClinicalTrials.gov Identifier:NCT02315508). Fifteen NHS sites in England were involved in the trial, recruiting participants by direct physician referral, GP surgeries, ENT and Audiology clinics, pharmacies and features in newspapers, magazines, radio, websites and social media. Eligibility criteria relating to tinnitus were a stable subjective tinnitus that was consistent from day to day, duration of tinnitus between 6 and 18 months and a global score of 24 to 68 on the Tinnitus Functional Index (Meikle et al., 2012). Tinnitus loudness measures were completed at the Screening visit (up to 28 days before dosing), Day1 visit (first dose), and Day28 visit (study end-point). The Screening and Day1 visits occurred before any drug was taken and so provide the loudness data for reliability analyses. Data were considered suitable because the scores between test and retest were expected to be stable (interval: 3 - 35 days) and none of the participants received any interventions for their tinnitus between these visits. The type of administration, environment and instructions were the same for all measurements and all visits. The loudness data collected at the Screening and Day1 visits comprised 91 participants (71 men and 20 women) with an average age of 54 years (range = 27-76). At the Screening visit, hearing threshold at 1 kHz was 12.4 dB HL (SD 11.3). The sample size fulfils the recommended minimum requirement (> 50 participants) for reliability analyses (Altman, 1991; Terwee et al., 2007). Data were collected in accordance with the permissions granted by the Yorkshire & The Humber - Leeds East NHS Research Ethics Committee (Ref:14/YH/1090) and the Sponsor (Autifony Therapeutics Ltd).

This Research Note does not report any findings related to clinical efficacy. The reliability and validity analyses reported here were guided by an international standard for evaluating the psychometric properties of instruments for clinical research (Mokkink et al., 2010), that we have previously implemented for evaluating the Tinnitus Functional Index questionnaire (Fackrell et al., 2016). We used criteria for good measurement properties reported by Terwee et al. (2007).

Measures

Tinnitus loudness matching test. Matching of the tinnitus loudness to a presented sound was conducted for one single-frequency tone. We chose to use a 1-Hz tone instead of a frequency corresponding to the dominant tinnitus pitch for the following reason. Generally 1 kHz is within the normal hearing range, while the tinnitus pitch tends to fall within the hearing loss region (Sereda et al., 2011). Hence, loudness matching at the dominant tinnitus pitch introduces a risk of bias when interpreting the loudness estimate because it can be unduly affected by loudness recruitment (an abnormal growth in the perception of loudness associated with sensorineural hearing loss). In our procedure, an audiologist adjusted the level of a 1-kHz tone in the non-tinnitus ear (or ear where the tinnitus was least dominant) in 2-dB steps using a staircase (up-down) method until its loudness was "about the same" as the participant's tinnitus (see Vernon and Meikle, 1981). Although responses may be variable with 2-dB steps, two match responses were obtained at the lowest level for that level to be recorded as the loudness match. Loudness was expressed in units of dB HL and dB SL. The loudness matching test took about 15 minutes to complete with an audiometer.

Tinnitus loudness rating. Question 2 of the Tinnitus Functional Index provided the patient-reported measure of loudness rating. Patients rated how loud their tinnitus had been over the past week on an 11-point (0-10) Likert scale, with higher scores indicating greater levels of loudness (Meikle et al., 2012). For analysis purposes, the scale was treated as continuous which is reasonable because there are 11-points, the underlying concept is on a continuous scale and the interval between points are approximately equal (Johnson and Creech, 1983; Zumbo and Zimmerman, 1993). Moreover comparing measures of central tendency (i.e. mean and median) confirmed a symmetric distribution since these values were approximately equal. This question took minimal time to complete and required no expertise or specialised instrumentation.

Clinical Global Impression (CGI) scale. On Day28, participants were also asked to rate the extent to which they had experienced any treatment-related change on a single 7-point Likert scale with categories from 'much worse' to 'much improved', with 'no change' at the mid-point.

ANALYSES AND RESULTS

Reliability

Reliability refers to the consistency of the measurement in test-retest situations and all analyses used the data collected on Screening and Day1 visits (n = 91). Two measurement properties are relevant; reliability and measurement error (Mokkink et al., 2010; Terwee et al., 2007). It can be rather confusing that the same term 'reliability' refers to both the category and one of the measurement properties within that category. We try to avoid confusion by referring to the measurement property as 'test-retest reliability'. Test-rest reliability indicates how well participants can be distinguished from each other, despite measurement errors. Measurement error refers to the difference between a measured value of loudness and its true value.

Test-retest reliability using IntraClass Correlation (ICC). Test-retest reliability was computed using the measurement variance over two time points with the same participants (n=91). An ICC is the ratio of all variances ranging from 0 (no reliability) to 1 (perfect reliability), with 0.4 to 0.7 considered acceptable (Andresen, 2000). The ICC findings were obtained using a two-way random-effects model in SPSS and are shown in Table 1. Results indicate that 70% (dB HL) and 67% (dB SL) of the variability in the observed tinnitus loudness matching scores could be attributed to the true score. For tinnitus loudness rating, it was 66%. All findings are acceptable and we interpret this result as indicating that the way in which participants differed from one another at test was reasonably stable at retest.

** Table 1 **

Measurement error using Limits of Agreement (LoA) and Smallest Detectable Change (SDC). Both of these statistics are directly related to the Standard Error of Measurement (SEM) and were calculated using Stata software. It is generally regarded that the LoA and SDC values should be comparable since they both reflect the same statistical property. For completeness, we report both.

First, the SEM assessed the accuracy of the loudness measures by estimating the deviation between the scores on the test and retest visits. SEMs are expressed in same units of measurement as the original test score and are therefore are reasonably easy to interpret, with large values indicating low levels of “test” accuracy and a large degree of error. SEM$_{Consistency}$ does not account for any systematic bias in the measurement, while SEM$_{Agreement}$ does. Results for both SEM formulae are reported in Table 1. Values of SEM$_{Agreement}$ were greater than those for SEM$_{Consistency}$ consistent with a very small degree of systematic bias. SEM$_{Agreement}$ was 7.6 dB for tinnitus loudness matching in dB HL, 7.2 in dB SL, and 1.25 units for tinnitus loudness rating.

The Bland–Altman method (1986) for LoA calculates the mean difference in scores between two repeated visits (the 'bias') and 95% limits of agreement. 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. Plots of these distributions are shown in Figure 1 and findings are also tabulated (Table 1). First, the LoA data demonstrate that being enrolled in the clinical trial did not unduly bias the loudness data because the difference in loudness scores were evenly distributed around the mean zero position. Notably, one would expect any systematic bias from one testing session to another to be visible in these plots either in terms of a weighting in the difference in loudness scores above or below the zero line (e.g if re-test scores were consistently lower at one of the testing visits) or in terms of a funnel shape (e.g. if the difference scores were influenced by the magnitude of the initial loudness score). Neither pattern was observed. For tinnitus loudness matching (dB HL), a change score of 20.22 dB HL or smaller was likely to be due to measurement error, and 92.0% of the data points fell within the LoA indicating that scores greater or equal to 20.22 dB HL would represent real change in 92% of participants. For dB SL, results were 18.05 dB SL and 93.0%, respectively. For tinnitus loudness rating, a change score of 2.61 on the Likert scale (or smaller) was likely to be due to measurement error, and 94.5% of the differences fell within the LoA.

** Figure 1 **

SDC represents the smallest change in score that is beyond measurement error and is given by the formula: SDC_{ind} = 1.96 × √2 × $SE\mathrm{M}_{\mathrm{agreement}}$. Table 1 shows that SDC values for tinnitus loudness matching and loudness rating were always comparable to the LoA values; just one point or so greater.

Interpreting the measurement errors. Terwee's criteria for acceptable confidence in the observed estimates of psychometric reliability require that the LoA is higher than the reported SEM values and that the SDC values are broadly equivalent to their LoA counterparts (Terwee et al., 2007). In summary, these criteria were achieved for all loudness measures. Expressing tinnitus loudness matching in dB HL or SL made no material difference, for this group of participants. While Henry et al. (1999) had expressed a concern that loudness expressed in dB SL introduced another source of within-subject variability, we found no evidence to suggest that this affected overall reliability.

The effects of test-retest interval on reliability: All estimates were consistent with a rather high measurement error relative to the absolute mean score on each loudness measurement. Since patients were re-assessed across a rather wide interval (3-35 days), and since Terwee et al. (2007) recommend an interval of up to 2 weeks for such an analysis, we wondered whether longer test-retest intervals were associated with a higher measurement error than shorter intervals. To address this question, the dataset was split into two subgroups (3-19 days and 20-35 days, comprising n=47 and n=44 patients respectively) for a post hoc analysis. Results for tinnitus loudness matching (presented in Table 2), fail to show any substantive difference in reliability across the two subgroups. Thus our previous findings were confirmed.

** Table 2 **

Validity

Face validity using patient improvement ratings. Face validity means that the test 'looks like' it will work. If the loudness test captures a clinically relevant aspect of the tinnitus construct, then those participants experiencing an overall improvement should have a reduction in loudness on Day28 compared to Day1, and vice versa. In other words, those with a CGI rating of 'improved' should have a reduction in loudness score and those participants with a CGI rating of 'worsened' should have an increase in score. In both cases, the magnitude of the increases and decreases should exceed the measurement error.

** Figure 2 **

Fifteen participants did not complete Day28 visit and so the analysis of face validity was conducted on 76 participants. These data demonstrated that tinnitus loudness rating produced this expected monotonic function (Figure 2). The 'slightly improved' category had a 0.97 unit mean reduction compared to the 'no change' category, and the 'much improved' category had a 2.41 unit mean reduction. In contrast, tinnitus loudness matching was rather flat, and one of the change categories did not make intuitive sense (the 'slightly worsened' loudness (dB SL) corresponded to a decrease in the loudness estimate), but the confidence intervals around this estimate were large. Our findings are indicative rather than definitive, since the magnitude of loudness changes between adjacent categories fell within the measurement error.

Convergent validity using correlation. If the two loudness tests measure the same theoretical construct, then they should show excellent convergent validity. A correlation coefficient ≥ 0.60 indicates excellent convergence, and 0.30-0.59 is acceptable (Andresen, 2000). This question was investigated using the Day1 ('pre-dose') data (n=91). A Spearman's correlation was calculated because the tinnitus loudness matching data (dB SL) were not normally distributed (i.e. they were positively skewed). The tinnitus loudness rating had either poor or borderline acceptable convergence with the two tinnitus loudness matches (0.35 for dB HL, and 0.29 for dB SL). Tinnitus loudness matches in dB HL and dB SL showed acceptable convergent validity with one another (0.57). But even this estimate is surprisingly lower than expected given that these measurements points originate from the same test.

DISCUSSION

Our psychometric exploration of the reliability and validity of two tinnitus loudness measures generated a number of findings that are important for selecting a measure of tinnitus loudness for clinical research.

Test-retest reliability assessed by the ICC was acceptable, meaning that participants could be distinguished from each other, despite measurement error. However, measurement error was rather large for both tinnitus loudness tests that we investigated. The more conservative estimate of measurement error is given by the SDC which was 21 dB HL and 20 dB SL for the tinnitus loudness matching test, and 3.5 points for the single tinnitus loudness rating question. The measurement (im)precision is such that large changes exceeding the SDC would be needed in order for that change to be attributed to a tinnitus intervention, and not simply to measurement error. For tinnitus loudness matching, the loudness estimate would need to change by approximately 200% from its baseline value for it to indicate a reliable change. This makes the loudness matching test somewhat undesirable, irrespective of whether the measurement is in dB HL or dB SL (c.f. Henry et al., 1999). Furthermore, the too many low scores in the distribution for dB SL would violate the normality distribution assumption required for parametric statistical testing. For tinnitus loudness rating that change would need to be 50% of its baseline value. While the tinnitus loudness rating is appealing in terms of this property, single-item measurement instruments tend to be viewed as ‘psychometrically suspect’; (i) they are more vulnerable to random measurement errors which are more likely to be eliminated with multiple items, (ii) the reliability statistic 'internal consistency' cannot be computed, and (iii) they are more vulnerable to unknown biases in meaning and interpretation.

Any preference for tinnitus loudness rating over tinnitus loudness matching should take into account the degree to which each measure captured a clinically relevant aspect of tinnitus, as experienced by the participant. The tinnitus loudness rating had some degree of face validity, for example with an observed reduction of 0.97 ('slightly improved' – 'no change' CGI categories); a magnitude that is comparable to that reported by Adamchic at al. (2012). However, this needs to be tempered by the effect of measurement error; not considered by Adamchic at al. (2012). The magnitude of this minimum threshold for distinguishing between patient groups could not be distinguished from measurement error. What this means, in the context of clinical research, is a reduced ability to statistically detect the smallest differences between treatment and control groups that are meaningful to patients, one would have to rely on the tinnitus loudness ratings of 'much/very much' changed to have confidence that the change was above error.

Finally, the tinnitus loudness rating question seemed to be measuring a different theoretical concept from the loudness matching test because the two measurements had rather poor convergent validity. This observation is noteworthy given that investigators consider these instruments as equivalent and hence interchangeable. Other studies have reported a high correlation between tinnitus loudness rating and global scores on a multi-attribute, multi-item tinnitus questionnaire (e.g. Adamchic et al., 2012; Zenner and Maddalena, 2005); suggesting that they measure a similar construct. These findings would therefore suggest that the single question about loudness is interpreted by patients as synonymous with overall tinnitus impact.

In conclusion, the implication for clinicians is that while the tinnitus loudness matching test administered by an audiologist assesses perceptual attributes of tinnitus, the tinnitus loudness rating (a patient-reported measure) might instead be assessing its subjective impact. Our specific conclusions about tinnitus loudness matching are limited to a 1-kHz tone, and cannot be extrapolated to other tone stimuli (e.g. those corresponding to the dominant tinnitus pitch). In terms of the main research questions, tinnitus loudness rating performed better against acceptability criteria for reliability and validity than did the tinnitus loudness matching. But the rating question is still limited because it is a single-item instrument, is likely able to detect only large changes in tinnitus loudness after treatment or between patient groups that probably correspond to perceptions of 'much/very much' changed, and may be assessing overall tinnitus impact rather than loudness per se.

ACKNOWLEDGMENTS

This article presents independent research with all co-authors funded by the National Institute for Health Research (NIHR) under the Biomedical Research Unit programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The authors report no conflicts of interest. Data were collected as part a clinical trial funded by INNOVATE UK ref 35370-247243 (Charles Large, Principal Investigator, Autifony Therapeutics Ltd and Deborah Hall, Academic Partner). We thank Alice Sharman (Autifony Therapeutics Ltd) for facilitating access to the dataset.

REFERENCES

Adamchic, I., Langguth, B., Hauptmann, C., Tass, P.A. (2012). Psychometric evaluation of visual analog scale for the assessment of chronic tinnitus. *Am J Audiol.* 21(2), 215-25.

Altman, D.G. (1991). *Practical Statistics for Medical Research.* London: Chapman & Hall.

Andresen, E.M. (2000) Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation* 81(2), 15-20.

Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurements. *The Lancet* 327, 307-10.

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, 220-35.

Feldt, L., Brennan, R. (1989). Reliability. In R. Linn (Ed). *Educational Measurement* (3rd Ed), Washington DC, The American Council on Education and the National Council on Measurement in Education.

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(1), 270.

Henry, J.A., Flick, C.L., Gilbert, A., Ellingson, R.M., Fausti, S.A. (1999). Reliability of tinnitus loudness matches under procedural variation. *J Am Acad Audiol.* 10, 502-20.

Henry, J.A., Rheinsburg, B., Owens, K.K., Ellingson, R.M. (2006). New instrumentation for automated tinnitus psychoacoustic assessment. *Acta Otolaryngol Suppl.* 556, 34-8.

Hoare, D.J., Edmondson-Jones, M., Gander, P.E., Hall, D.A. (2014). Agreement and reliability of tinnitus loudness matching and pitch likeness rating. *PLoS One.* 9(12), e114553.

Johnson, D.R., & Creech, J.C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. American Sociological Review, 48, 398-407.

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(2), 153-76.

Mitchell, C.R., Vernon, J.A., Creedon, T.A. (1993). Measuring tinnitus parameters: loudness, pitch, and maskability. *J Am Acad Audiol.* 4, 139-51.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research.* 19(4), 539-49.

Prinsen, C.A.C., Vohra, S., Rose, M.R., Boers, M., Tugwell, P., Clarke, M., Williamson, P.R., Terwee, C.B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials* 17, 449.

Sereda, M., Hall, D.A., Bosnyak, D.J., Edmondson-Jones, M., Roberts, L.E., Adjamian, P., Palmer. A.R. (2011). Re-examining the relationship between audiometric profile and tinnitus pitch. Int J Audiol. 50:303-12.

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. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 60, 34-42.

Tunkel DE, Bauer CA, Sun GH, Rosenfeld RM, Chandrasekhar SS, Cunningham ER Jr, Archer SM, Blakley BW, Carter JM, Granieri EC, Henry JA, Hollingsworth D, Khan FA, Mitchell S, Monfared A, Newman CW, Omole FS, Phillips CD, Robinson SK, Taw MB, Tyler RS, Waguespack R, Whamond EJ. Clinical practice guideline: tinnitus. Otolaryngol Head Neck Surg. 2014 Oct;151(2 Suppl):S1-S40.

Vernon, J.A., Meikle, M.B. (1981). Tinnitus masking : unresolved problems. *Ciba foundation Symposium* 85, 239-56.

Zenner, H.P., De Maddalena, H. (2005). Validity and reliability study of three tinnitus self-assessment scales: loudness, annoyance and change. *Acta Otolaryngol* 125(11), 1184-8.

Zumbo, B.D., & Zimmerman, D.W. (1993). Is the selection of statistical methods governed by level of measurement? Canadian Psychology, 34, 390-400.

FIGURE LEGENDS

Figure 1. ‘Bland–Altman’ limits of agreement plots of measurement error for repeated measures of tinnitus loudness at (pre-dose) Screening and Day1 visits. The top panel presents loudness matching (dB HL), the middle panel loudness matching (dB SL), and the bottom panel loudness rating. Dashed line = mean difference (see Table 1 for actual values). Solid black lines = the 95% limits of agreement.

Figure 2. Change in loudness score (Day28 - Day1) as a function of CGI ratings. The top panel plots loudness matching (dB HL), the middle panel plots loudness matching (dB SL), and the bottom panel plots loudness rating. Error bars show the 95% confidence interval. Categories for ‘much improved’ and ‘moderately improved’ were pooled together (n=8), and so were categories for ‘moderately worse’ to ‘much worse’ (n=6). The number of participants in each of the other categories was: ‘slightly improved’ (n=8), ‘no change’ (n=47), and ‘slightly worse’ (n=7).

Table 1. Reliability evaluation of the two loudness tests for the Full Safety Population completed at (pre-dose) Screening and Day1 visits. Agree=Agreement; CI=Confidence Intervals; Con=Consistency; N=size of dataset at each visit; SD=standard deviation; SE=standard error; ICC=Intra Class Correlation; LoA=Limits of Agreement; SDC= Smallest Detectable Change; SEM=Standard Error of Measurement. ICC values are reported for the single measure which applies to individual scores. Two participants did not provide data for loudness matching at the Day1 visit which explains why n=89.

	Descriptive statistics						**Reliability**	**Measurement error**						
	Mean (SD)			**Difference**			**Reliability**	**SEM**		**SDC**	**LoA**			
Test	**N**	**Screening**	**Day1**	**Mean diff**	**SE**	**SD$_{diff}$**	**ICC (95% CI)**	**Con**	**Agree**	**SDC**	**LoA**	**LoA Lower limit (95%CI)**	**LoA Upper limit (95%CI)**	**%**
Loudness matching (dB HL)	91,89	28.81 (13.19)	28.33 (13.16)	0.37	1.08	10.16	0.70 (0.58, 0.79)	7.18	7.60	21.07	20.22	-19.95 (-23.65, -16.25)	20.69 (16.99, 24.39)	92.0
Loudness matching (dB SL)	91,89	17.85 (11.97)	17.52 (10.62)	0.46	0.98	9.21	0.67 (0.54, 0.77)	6.51	7.20	19.96	18.05	-17.96 (-21.30, -14.62)	18.88 (15.54, 22.22)	93.0
Loudness rating	91,91	6.31 (1.63)	6.19 (1.58)	0.12	0.14	1.33	0.66 (0.52, 0.76)	0.94	1.25	3.46	2.61	-2.54 (-3.02, -2.06)	2.78 (2.30, 3.26)	94.5

Table 2. Duration of the interval between test (Screening visit) and re-test (Day1 visit) does not unduly affect estimates of reliability and measurement error, for tinnitus loudness matching. Data for Table 1 were re-estimated for two sub-groups; short test-retest interval (3-19 days) and long test-retest interval (20-35 days). Two participants in the 'short interval' subgroup did not provide data for loudness matching at the Day1 visit which explains why n=45. Abbreviations are defined in Table 1.

	Descriptive statistics						**Reliability**	**Measurement error**						
	Mean (SD)			**Difference**			**Reliability**	**SEM**		**SDC**	**LoA**			
Test	**N**	**Screening**	**Day1**	**Mean diff**	**SE**	**SD_{diff}**	**ICC (95% CI)**	**Con**	**Agree**	**SDC**	**LoA**	**LoA Lower limit (95%CI)**	**LoA Upper limit (95%CI)**	**%**
Short test-retest interval (3-19 days)														
Loudness matching (dB HL)	47, 45	29.38 (13.72)	28.13 (13.14)	1.04	1.59	10.65	0.68 (0.49, 0.81)	7.53	8.98	24.89	21.30	-20.26 (-25.76, -14.76)	22.34 (16.84, 27.84)	91.0
Loudness matching (dB SL)	47,45	18.06 (11.08)	17.11 (11.06)	1.22	1.46	9.83	0.61 (0.39, 0.77)	6.95	9.05	25.08	19.66	-18.44 (-23.54, -13.34)	20.88 (15.78, 25.98)	91.0
Loudness rating	47, 47	6.40 (1.50)	6.18 (1.58)	0.22	0.23	1.60	0.46 (0.20, 0.66)	1.13	1.57	4.34	3.20	-2.98 (-3.78, -2.18)	3.42 (2.62, 4.22)	91.0
Long test-retest interval (20-35 days)														
Loudness matching (dB HL)	44, 44	28.20 (12.73)	28.52 (13.33)	-0.32	1.46	9.71	0.72 (0.54, 0.84)	6.87	7.03	19.48	19.42	-19.74 (-24.82, -14.66)	19.10 (14.02, 24.18)	91.0
Loudness matching (dB SL)	44,44	17.61 (12.98)	17.93 (10.27)	-0.32	1.29	8.58	0.73 (0.56, 0.84)	6.07	6.25	25.00	22.66	-17.48 (-21.94, -13.02)	16.84 (12.38, 21.3)	95.0
Loudness rating	44, 44	6.22 (1.76)	6.22 (1.61)	0.00	0.15	1.03	0.81 (0.68, 0.89)	0.73	0.73	2.02	2.06	-2.06 (-2.60, -1.52)	2.06 (1.52, 2.60)	97.0

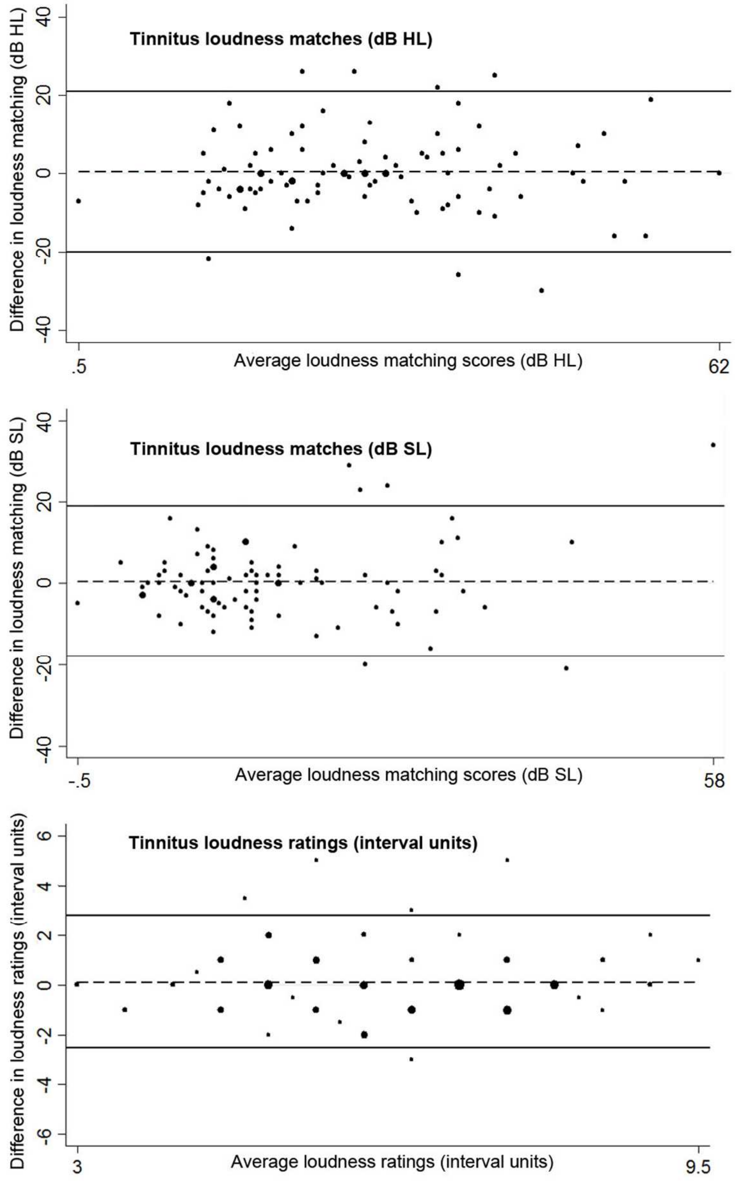

Figure 1. 'Bland–Altman' limits of agreement plots of measurement error for repeated measures of tinnitus loudness at (pre-dose) Screening and Day1 visits. The top panel presents loudness matching (dB HL), the middle panel loudness matching (dB SL), and the bottom panel loudness rating. Dashed line = mean difference (see Table 1 for actual values). Solid black lines = the 95% limits of agreement.

360x561mm (72 x 72 DPI)

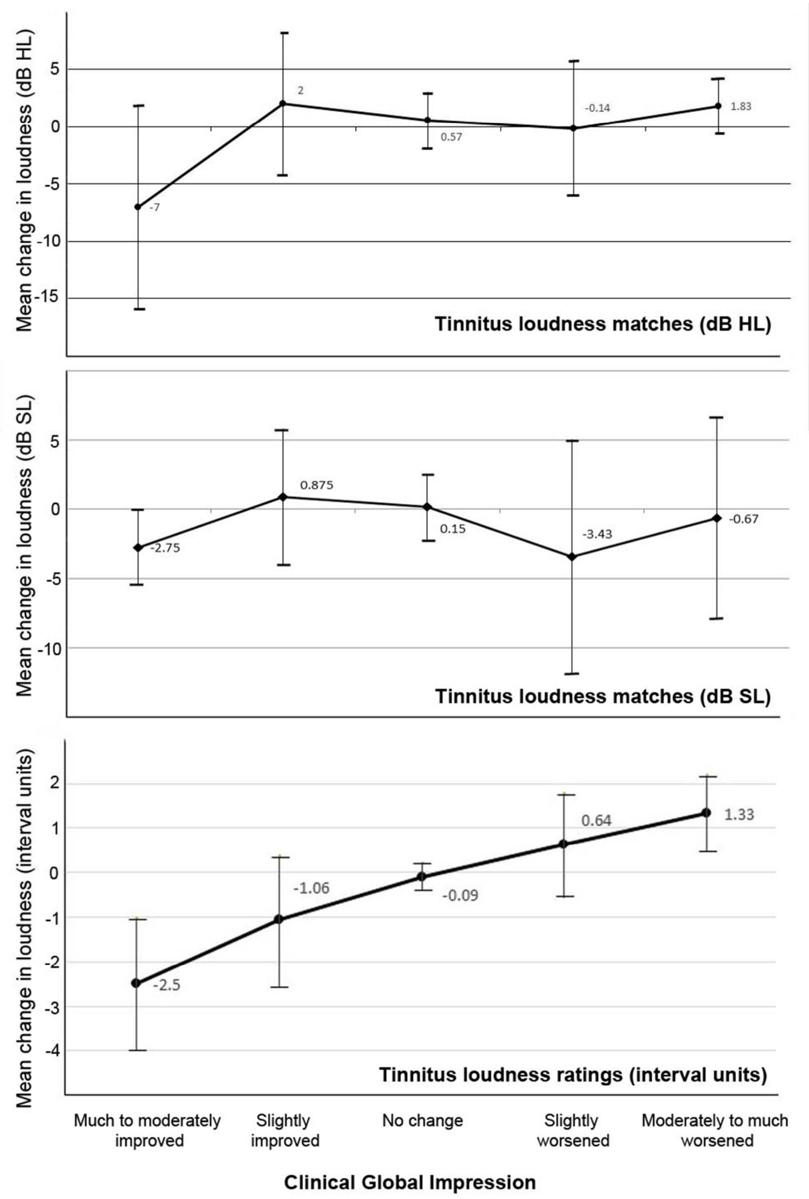

Figure 2. Change in loudness score (Day28 - Day1) as a function of CGI ratings. The top panel plots loudness matching (dB HL), the middle panel plots loudness matching (dB SL), and the bottom panel plots loudness rating. Error bars show the 95% confidence interval. Categories for 'much improved' and 'moderately improved' were pooled together (n=8), and so were categories for 'moderately worse' to 'much worse' (n=6). The number of participants in each of the other categories was: 'slightly improved' (n=8), 'no change' (n=47), and 'slightly worse' (n=7).

204x303mm (120 x 120 DPI)