Interpreting the Tinnitus Questionnaire (German version): what individual differences are clinically important?

Journal:	*International Journal of Audiology*
Manuscript ID	TIJA-2017-06-0191.R2
Manuscript Type:	Technical Report
Date Submitted by the Author:	n/a
Complete List of Authors:	Hall, Deborah; University of Nottingham, Otology and Hearing group, Division of Clinical Neuroscience, School of Medicine; National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre, Hearing theme Mehta, Rajnikant; University of Nottingham, Otology and Hearing group, Division of Clinical Neuroscience, School of Medicine; University of Nottingham, National Institute for Health Research Nottingham Hearing Biomedical Research Unit Argstatter, Heike; German Center of Music Therapy Research (Viktor Dulger Institute) DZM e.V., Deutsches Zentrum für Musiktherapieforschung (Viktor Dulger Institut) DZM e.V.
Keywords:	Tinnitus, Instrumentation, Psycho-social/Emotional, Adult or General Hearing Screening

Interpreting the Tinnitus Questionnaire (German version): what individual differences are clinically important?

Deborah A Hall [1,2], Rajnikant L. Mehta [1,2], Heike Argstatter [3]

1. National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre, Ropewalk House, 113 The Ropewalk, Nottingham UK NG1 5DU

2. Otology and Hearing group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham UK NG7 2UH

3. Deutsches Zentrum für Musiktherapieforschung (Viktor Dulger Institut) DZM e.V., 69123 Heidelberg, Germany

Corresponding Author: Email: deborah.hall@nottingham.ac.uk

Tinnitus

Instrumentation

Psycho-social/Emotional

Adult or General Hearing Screening

Interpreting the Tinnitus Questionnaire (German version): what individual differences are clinically important?

Deborah A Hall [1,2], Rajnikant L. Mehta [1,2], Heike Argstatter [3]

1. National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre, Ropewalk House, 113 The Ropewalk, Nottingham UK NG1 5DU

2. Otology and Hearing group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham UK NG7 2UH

3. Deutsches Zentrum für Musiktherapieforschung (Viktor Dulger Institut) DZM e.V., 69123 Heidelberg, Germany

Corresponding Author: Email: deborah.hall@nottingham.ac.uk

ABSTRACT

Objective Reporting of clinical significance is recommended because findings can be statistically significant without being relevant to patients. For aiding clinical interpretation of the Tinnitus Questionnaire (TQ), many investigators use a 5-point change cut off as a minimal clinically important difference (MCID). But there are shortcomings in how this value was originally determined.

Design The MCID was evaluated by analysing retrospective clinical data on the TQ (German-version). Following recommended standards, multiple estimates were computed using anchor- and distribution-based statistical methods. These took into account not only patients' experience of clinical improvement, but also measurement reliability.

Study sample Pre- and post-intervention scores were assessed for 202 patients.

Results Our six estimates ranged from 5 to 21 points in TQ change score from pre- to post-intervention. The 5-point TQ change score was obtained using a method that considered change *between* groups, and did not account for measurement error or bias. The size of the measurement error was considerable, and this comprises interpretation of individual patient change scores.

Conclusions To enhance confidence that a TQ change over time in individual patients is clinically meaningful, we advise at least the median MCID of 12 points.

INTRODUCTION

There is little evidence-based guidance to facilitate design decisions for definitive trials that evaluate treatment efficacy for adults with tinnitus. Eligibility criteria, characteristics of enrolled participants, tinnitus-related outcome instruments, and criteria for interpreting any observed treatment-related change are all highly diverse across clinical trials and this precludes direct comparison across findings (Hall et al., 2016a). With respect to outcome instruments, while trial findings report group-level statistical results on the treatment-related change they rarely seek to additionally interpret whether or not those observed improvements are clinically meaningful. Reporting of clinical significance is recommended because findings can be statistically significant, without being clinically significant (i.e. there is a change but it is not relevant to patients). A minimal clinically important difference (MCID) should be defined as a threshold for a change in global questionnaire score over which a patient or physician would consider that change to be meaningful and worthwhile. Hence, determining MCID is critical for conducting and interpreting meaningful clinical trials, as well as for facilitating the establishment of treatment recommendations for patients. If different intervention studies interpret their clinical efficacy results in different ways, then findings cannot easily be synthesised and conclusions cannot be compared.

Using a list of 228 clinical trials identified in a recent systematic review (Hall et al., 2016a), we found that the Tinnitus Handicap Inventory (THI, Newman et al., 1998) was the most popular choice of outcome instrument, with the Tinnitus Questionnaire (TQ) second, and the Tinnitus Functional Index (TFI, Meikle et al., 2012) third (Hall et al., 2016a). All three instruments yield a composite score reflecting the impact of tinnitus. The THI test-retest reliability indicates an MCID of 20 points or greater for interpreting individual patients (Newman et al., 1998). TFI development paid careful attention to its responsiveness to treatment-related change (more so than the THI or TQ) and the authors estimated an MCID of 13 points or greater (Meikle et al., 2012).

The TQ was originally developed by Hallam et al. (1988). It was modified and translated into German by Hiller and Goebel (1992). In Germany, therefore, the TQ tends to be chosen in preference to the THI or TFI (Hall et al., 2016a). This work concerns the German version of the TQ (henceforth referred to as TQ). The global TQ score is based on the 42 questions corresponding to the TQ subscales, with a maximum score of 84, with higher values indicating greater symptom severity. Goebel and Hiller (1998) proposed a four-category grading system for diagnostic assessment (mild, moderate, severe, very severe), but no recommendations about MCID. The review of 228 clinical trials found that 15 of those studies reporting the TQ then went on to interpret individual pre-post change scores using a threshold criterion for defining 'clinical improvement' ((Hall et al., 2016a, see Supplemental File 1). Eleven of those studies used a 5-point cut off, with four studies justifying that choice by citing an article by Kleinjung et al. (2007, p591). But this article contains insufficient information to explain how this threshold was determined, and the original data are no longer accessible [Kleinjung, personal communication]. A subsequent report, which reiterated a 5-point cut off, did not fully account for the test-retest reliability of the measurement (Adamchic et al. 2012).

This Technical Report responds to an appeal by Hall et al. (2016b) to reassess clinically important difference estimates taking account of measurement reliability as well as the patient experience of clinical improvement (Mokkink et al., 2010; Terwee et al., 2007).

METHODS

This was a retrospective analysis using a clinical dataset that comprised 202 patients whom received the Heidelberg Neuro-Music-Therapy intervention at the DZM e.V. (German Center for Music Therapy Research) in Germany, from 2011 to 2016. This intervention is targeted at patients with a clinical diagnosis of chronic tonal tinnitus persisting for a minimum of 6 months. Median duration was 7 years. Patients were not randomised and were not blinded to the treatment. Inclusion criteria were representative for patients receiving this form of treatment in routine clinic:

- age > 18
- no severe psychiatric disorders (such as major depression, psychosis, dementia)
- no ongoing medication interfering with therapy (such as high dosages of tranquilizers) or possibly aggravating tinnitus (such as cisplatin)
- no substance abuse
- no treatable, ongoing organic causes for tinnitus (apart from hearing loss; ear disorders in the past possibly causing tinnitus were okay if patients had recovered from the disease)
- baseline TQ > 30.

All patients were treated according to the compact model of the Heidelberg Neuro-Music Therapy, which was nine sessions of individual music therapy over five consecutive days, combined with psychoeducation, relaxation training, tinnitus habituation, and stress management (Argstatter et al., 2015). The Heidelberg Model aims to restore emotional well-being. The TQ is a reasonable choice of outcome measure since it has a large number of items assessing emotional well-being associated with tinnitus. The mean age of the sample was 52 years (SD 13), with 134 men and 68 women. The TQ was completed both before and after therapy. In addition, patients were asked an anchor question about whether their tinnitus symptoms were changed by the music therapy. There were pre-defined response options (worsened, unaffected or improved) and those whose tinnitus worsened or improved were asked a follow-up question about its severity. This enabled us to classify responses into five categories: ‘much better’, ‘slightly better’, ‘no change’, ‘slightly worse’ and ‘much worse’. We refer to this question as the Clinical Global Impression (CGI) scale.

MCID analyses

There is no single method for determining what the MCID should be. But methodologists generally recommend triangulating the results of multiple methods for determining clinically important difference (Revicki et al., 2006, see also Adamchic et al., 2012). Here we report findings from anchor-based *and* distribution-based methods. First, anchor-based methods

assess what change on the TQ corresponds with a minimal important change defined on the anchor which is an external criterion used to operationalise a relevant or important difference. Typically the anchor is the CGI. A within-patients estimate corresponds to the change in global TQ scores for patients who responded 'slightly better' and can be a good indicator of the smallest change that is important from the individual patients' perspective to identify 'responders'. A between-patients estimate corresponds to the degree of change in global TQ scores between the anchor response categories 'slightly better' and 'no change' and can be a good indicator of the smallest change for determining differences between treatment groups.

Second, distribution-based methods convey information about reliability (i.e. the consistency of the TQ measurement in test-retest situations). These were the IntraClass Correlation (ICC) used to assess reliability over test-retest situations, and Limits of Agreement (LoA) and Smallest Detectable Change (SDC) used to estimate measurement error. The sample size fulfils the recommended minimum requirement (> 50 participants) for these distribution-based reliability analyses (Terwee et al., 2007). The ICC partials out variance across participants from other sources of variance such as measurement error. Furthermore, our calculated ICC values were interpreted as follows: <0.40 (poor), 0.40-0.75 (good) and >0.75 (excellent) (Fleiss, 1986). A two-way mixed model was chosen in which patients were random but time points were fixed. The LoA estimates the interval between which lies 95% of the difference in scores between 'pre' and 'post'. The assumption is that if the mean difference between the scores was zero (i.e. no bias), then 95% of the data points would be within ±2 standard deviations of the mean difference. The SDC is related to the Standard Error of Measurement (SEM) ($SDC=1.96*\sqrt{2}*SEM_{Con}$) and it expresses the smallest change that would have to occur for that change to be considered a 'real' change not solely due to measurement error. We used SEM_{Con}, which is less susceptible to systematic differences in time points than SEM_{Agree}. We also used the 1/2 standard deviation rule since this magnitude of difference between groups corresponds to Cohen *d* = 0.5 and

represents a medium effect size (Cohen, 1988; Norman et al., 2003; Sawilowsky, 2009). Finally, we calculated the Reliable Change Index (RCI, i.e. the pre- versus post-intervention difference divided by the standard error of the difference), used widely in psychological research (Jacobson and Truax, 1991).

RESULTS

For 20 patients we did not have data to categorise the 5-point CGI, and so we estimated MCID using the remaining n=182. Individual anchor response categories were screened for their associated TQ change scores. TQ change was determined by subtracting the global score at pre-treatment from the post-treatment baseline ('post-pre'). Thus, a negative TQ change score indicates a tinnitus improvement. The 'slightly better' CGI group had a 12.0 point 'post-pre' reduction on the TQ (SD 9.8), and for the 'no change' CGI group it was 7.4 points (SD 10.6) (Table 1). Global TQ change scores plotted as a function of CGI groups, revealed the expected positive monotonic function (Figure 1). If the TQ measures the same theoretical construct as the CGI, then the two scales should show excellent convergent validity. The degree of convergence was assessed by the correlation coefficient, where at least 0.30 is acceptable for inclusion in MCID calculations (Andresen, 2000). Spearman's rank correlation was borderline acceptable ($r=0.28$, $p<0.001$). However, the 95% confidence intervals around each category mean score were very broad (Figure 1, Table 1) such that the difference between 'slightly better' and 'no change' CGI groups for the TQ change was not statistically reliable according to post-hoc testing ($p=0.12$). Revicki et al. (2006) note that in cases of variability in the 'no change' CGI group, then the MCID may be based on the TQ change difference between the 'no change' and the 'slightly better' CGI groups (i.e. the between-patients estimate). This was 4.6 mean score points in our dataset.

** Tables 1 and 2 **

For the distribution-based results, we first assessed reliability using the pre- and post-treatment scores for the 'no change' CGI-I group (n=52), since these patients should

show stable TQ 'pre' and 'post' scores. Results are given in Table 2. The observed ICC value of 0.61 (95% CI 0.41/0.76) indicated that 61% of the variability in the TQ could be attributed to the true score. This value was acceptable (Andresen, 2000; Fleiss, 1986) and demonstrated the way in which patients differed from one another was reasonably stable at retest. Measurement error is the other form of reliability and here this refers to the difference between the TQ score and its true value. Terwee et al's (2007) criteria for acceptable confidence in the observed estimates of psychometric reliability are that the LoA is higher than the reported SEM_{Con}, and that the SDC and LoA values are broadly equivalent. Values obtained for LoA, SEM_{Con} and SDC demonstrate that these criteria were met (see Table 2 and Figure 1, inset panel). The LoA indicates that a change score of 21 points or smaller was likely to be due to measurement error, and that 94% of the data points fell between -28.54 and +13.70. The 1/2 standard deviation was 6.46, and the RCI was 12.89.

The clinical relevance of these different MCID criteria was explored by comparing the classification of responders and non-responders, according to the CGI, with the classification according to each of the above statistical methods. Supplemental File 2 reports these findings. No single classification using TQ change scores fully agreed with the patients' own report, indicating the added value of using the CGI.

Finally we noted a tendency to deviate towards a tinnitus improvement and away from the expected value of zero because the 'no change' CGI group still scored a mean reduction of 7.4 points on the TQ, corresponding to a medium effect size (Cohen *d* = 0.62, Cohen, 1988; Sawilowsky, 2009). This pattern is suggestive of a systematic bias in patient reporting.

DISCUSSION

We examined the clinically important difference using a range of anchor- and distribution-based methods for a large retrospective sample of clinical data. From the present data, a numerical change in TQ score of 5 points did not reliably identify which patients in routine

clinic perceived an improvement in tinnitus-related emotional well-being, and it did not reliably distinguish patients reporting an improvement from those reporting no change in symptoms. Instead, a pragmatic interpretation of the data would indicate using the most central MCID value of 12 points (the median) for interpreting individual patient change scores on the TQ. This accounts for the effect of patient reporting bias but may not obviate measurement error in the clinical interpretation.

Cautions about a minimal clinically important difference (MCID) of 5 change points

Taken in isolation, the anchor-based method suggested an MCID of 4.6 change points on the German TQ for interpreting a treatment-related difference in improvement *between* patient groups. This MCID agree with the 5 integer change points reported by others (e.g. Kleinjung et al.; 2007; Adamchic et al., 2012). However, three noteworthy observations lead us to caution against using a small magnitude MCID when interpreting change in individual patients and when interpreting data collected from non-randomised, or unblinded clinical studies.

First, we observed a wide variability across patients, which meant that the 4.6 point difference in TQ change scores between the 'slightly better' and 'no change' CGI groups was not statistically significant. Second, we observed that those patients perceiving 'no change' still reported a mean TQ decrease in 7.4 points on post-treatment assessment compared with pre-treatment baseline (Table 1). A potential source of this systematic bias could be an *ascertainment bias* since the data were not collected in the context of a randomised, blinded, controlled trial. Each patient was told to expect some effects to occur during the 5 days of music therapy, but that they should continue exercises afterwards, and each was aware that his/her clinician was reviewing the post-intervention data. Third, when we considered the distribution-based estimates, they were all greater than 5 points (Table 2). Previous estimates of 5 points in global TQ change score are probably too conservative because they have not accounted for measurement error or such bias.

Implications for clinical trials and clinical practice

In conclusion, the findings add to a growing body of evidence that MCID values on these multi-attribute questionnaire instruments for tinnitus can be quite large. Based on the observed data, at least the median MCID of 12 points.is advised when interpreting TQ change scores. For clinical trials, an MCID of 12 points provides an evidence-based parameter in the power calculation needed to estimate sample size, it provides an evidence-based threshold for interpreting clinical significance of treatment-related change in 'before and after' study designs or of group mean differences in controlled study designs, and it provides an evidence-based numerical *a priori* criterion for classifying 'responders' and 'non-responders'. For clinical practice, TQ change scores can be informative, but clinicians should primarily remain sympathetic towards the patient's perception of any treatment-related change.

FUNDING DETAILS

DAH and RLM are funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

DISCLOSURE STATEMENT

The authors report no conflicts of interest.

BIOGRAPHICAL NOTE

Deborah Hall is an Experimental Psychologist and Professor of Hearing Sciences at the University of Nottingham. She has worked in hearing for the past 20 years and her current research interests include the design and conduct of clinical trials for evaluating the therapeutic benefit of interventions for tinnitus and outcome instruments for assessing hearing-related conditions.

Rajnikant Mehta has a masters degree in Medical Statistics. He currently works across projects in the School of Medicine at the University of Nottingham and provides statistics advice related to clinical trial design, use of robust statistical techniques/methods and appropriate funding streams/grant applications to clinical staff through the NIHR Research Design Service East Midlands. He currently is writing up his PhD, which is related to the epidemiology of diabetes and associated chronic disease co-morbidities in South Asians and White Europeans.

Heike Argstatter is a Clinical Psychologist and Senior Researcher at the German Center for Music Therapy Research (Deutsches Zentrum für Musiktherapieforschung DZM e.V.). In 2008, she received the Sigrid and Viktor Dulger Prize for the study on the impact of music therapy in chronic tinnitus. Her current research interests include the long-term effects of music therapy in tinnitus and music therapy in early rehabilitation after cochlear implantation.

ACKNOWLEDGEMENTS

Thanks to Tobias Kleinjung and Thomas Steffens who confirmed the statistical approach supporting their 5-point estimate for the MCID (Kleinjung et al., 2007).

REFERENCES

Adamchic, I., Tass, P.A., Langguth, B., Hauptmann, C., Koller, M., Schecklmann, M., Zeman, F., & Landgrebe, M. (2012). Linking the Tinnitus Questionnaire and the subjective Clinical Global Impression: which differences are clinically important? *Health Qual Life Outcomes, 10*, 79.

Andresen, E.M. (2000). Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil, 81*(12 Suppl 2), S15-20.

Argstatter, H., Grapp, M., Hutter, E., Plinkert, P.K., Bolay, H.V. (2015). The effectiveness of neuro-music therapy according to the Heidelberg model compared to a single session

of educational counseling as treatment for tinnitus: a controlled trial. *J Psychosom Res*, 78(3):285-92.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed). Hillsdale, NJ; Erlbaum.

Fleiss, J. (1986). Design and analysis of clinical experiments. New York; John Wiley and Sons.

Goebel, G., Hiller, W. (1998). Tinnitus-Fragebogen (TF). Ein Instrument zur Erfassung von Belastung und Schweregrad bei Tinnitus. Handanweisung. Goettingen: Hogrefe.

Hall, D. A., Haider, H., Szczepek, A. J., Lau, P., Rabau, S., Jones-Diette, J., & Fuller, T. (2016a). Systematic review of outcome domains and instruments used in clinical trials of tinnitus treatments in adults. *Trials*, *17*(1), 1.

Hall, D.A. (2016b). Interpreting treatment-related changes using the Tinnitus Questionnaire in Argstatter, H., Grapp, M., Plinkert, P. K., & Bolay, H. V. (2012). Heidelberg Neuro-Music Therapy" for chronic-tonal tinnitus-treatment outline and psychometric evaluation. Int Tinnitus J, 17(1), 31-41. *Int Tinnitus J* 20(2), 73-5.

Hallam, R.S., Jakes, S.C., & Hinchcliffe, R. (1988). Cognitive variables in tinnitus annoyance. *Br J Clin Psychol*, 27(3), 213–222.

Hiller, W., Goebel, G., (1992). A psychometric study of complaints in chronic tinnitus. *J Psychosom Res*, 36(4), 337-348.

Jacobson, N.S., Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol.* 59(1):12-9.

Kleinjung, T., Steffens, T., Sand, P., Murthum, T., Hajak, G., Strutz, J., & Eichhammer, P. (2007). Which tinnitus patients benefit from transcranial magnetic stimulation?. *Otolaryngology--Head and Neck Surgery*, *137*(4), 589-595.

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.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., & 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), 539-549.

Newman, C.W., Sandridge, S.A., Jacobson, G.P. (1998) Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcome. *J Am Acad Audiol.* 9(2), 153-60.

Norman, G. R., Sloan, J. A. Wyrwich, K. W. (2003) Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Med Care* 41:582-592.

Revicki, D. A., Cella., D., Hays., R. D., Sloan, J. A., Lenderking, W. R., Aaronson, N. K. (2006). Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes*, 4, 70.

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 467–474.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42.

Figure 1. The main panel presented the global TQ score changes from baseline, categorised by the Clinical Global Impression scale. Error bars indicate 95% confidence interval around the mean. The inset panel on the right presents a summary of the anchor- and distribution-based estimates of the MCID for the TQ.

Hall Running title: Interpreting the Tinnitus Questionnaire (German version)

Table 1. TQ characteristics of Clinical Global Impression (CGI) scale categories. CI= Confidence Interval.

CGI	Number of patients	ΔTQ, Mean (SD)	95% CI of ΔTQ	Cohen d	95% CI of d
Much better	20	-14.7 (12.6)	-27/-2.4	1.28	-3.63/6.71
Slightly better	92	-12.0 (9.8)	-21.7/-2.3	1.09	-0.97/3.52
No change	52	-7.4 (10.6)	-17.7/2.9	0.62	-2.48/4.05
Slightly worse	12	-7.6 (12.3)	-19.7/4.5	0.57	-5.59/9.62
Much worse	6	2.5 (20.6)	-18.1/23.1	-0.17	-8.09/16.72
Not coded for severity	*20*				

Table 2. Reliability evaluation of the TQ for the n=52 participants in the "no change" category of the CGI. 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. CI= Confidence Interval.

Descriptive statistics					Reliability	Measurement error						
Mean (SD)		Difference			Reliability	SEM		SDC		LoA		
Screening	Day1	Mean diff	SE	SD_{diff}	ICC (95% CI)	Con	Agree	SDC	LoA	LoA Lower limit (95%CI)	LoA Upper limit (95%CI)	%
43.60 (11.36)	36.17 (12.59)	-7.42	1.46	10.56	0.61 (0.41, 0.76)	7.47	38.58	20.71	21.12	-28.54 (-33.58, -23.50)	13.70 (8.66, 18.74)	94.0

Supplemental File 1. Summary of numerical criteria for what constitutes a minimal clinically important difference (MCID) on versions of the TQ using investigator-reported data extracted from 228 trials (Hall et al., 2016a). NR = not reported

Hall Running title: Interpreting the Tinnitus Questionnaire (German version)

Supplemental File 2. This table illustrates how many patients in our sample meet the different numerical criteria for what constitutes a minimal clinically important difference (MCID). Above the bold line are all those numerical criteria reported in the Technical report. Below the line are two additional criteria that were used only once in our review of 228 clinical trials (Hall et al., 2016a).

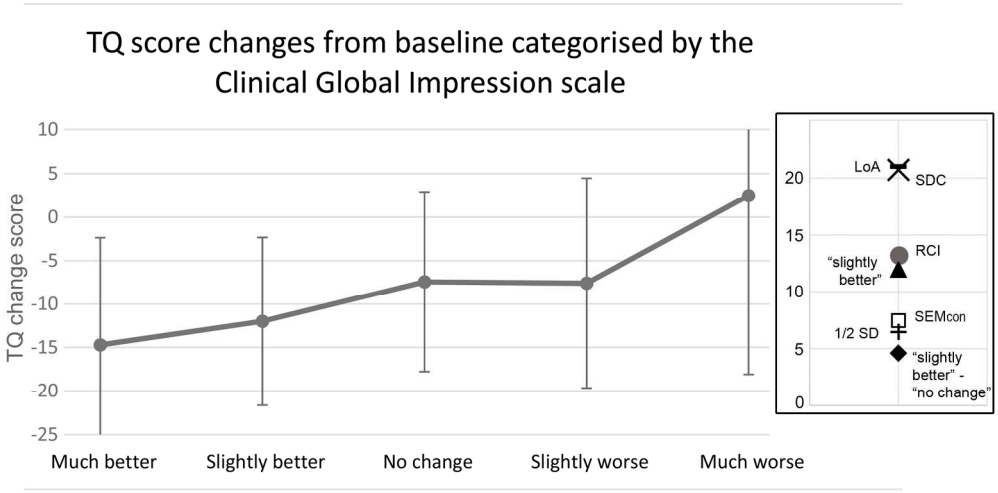

Figure 1. The main panel presented the global TQ score changes from baseline, categorised by the Clinical Global Impression scale. Error bars indicate 95% confidence interval around the mean. The inset panel on the right presents a summary of the anchor- and distribution-based estimates of the MCID for the TQ.

191x93mm (300 x 300 DPI)

Hall Running title: Minimal clinically important difference on the Tinnitus Questionnaire

Supplemental File 1. Summary of numerical criteria for what constitutes a minimal clinically important difference (MCID) on versions of the TQ using investigator-reported data extracted from 228 trials (Hall et al., 2016a). NR = not reported

TQ version	Investigator-reported MCID	Reference	Source of the MCID value
Criterion of improvement is a global TQ score reduction >5 points			
Version of translation unknown	Quote: "'therapeutic success' (TQ reduction >5)"	Chung et al. (2012)	Kleinjung et al. (2007)
German version	Quote: "treatment responders (TQ reduction > 5)"	Langguth (2012a)	NR
German version	Quote: "treatment responders (TQ reduction > 5)"	Langguth (2013a)	NR
Criterion of improvement is a global TQ score reduction >=5 points			
German version	Quote: "benefit from treatment, which was reflected by a reduction of the TQ score of five points or more"	Kleinjung et al. (2007)	Justification presented in the article
German version	Quote: "treatment response which was defined as amelioration of at least 5 points in the TQ"	Kreuzer et al. (2011)	NR
German version	Quote: "treatment response, which was defined as amelioration of at least 5 points in the TQ"	Kreuzer et al. (2012)	Kleinjung et al. (2008)
German version	Quote: "a clinical relevant change of tinnitus severity (i.e. 5 points on the (TQ) questionnaire of Goebel and Hiller"	Landgrebe et al. (2008)	Kleinjung et al. (2007)
German version	Quote: "treatment responders (TQ reduction ≥5)"	Langguth (2012b)	NR
German version	Quote: "treatment responders … total score reduction ≥ 5"	Langguth (2013b)	NR
German version	Quote:"treatment responders, defined as a minimum difference of five points"	Langguth et al. (2014)	Kleinjung et al. (2007)
German version	Quote: "treatment responders as defined by a reduction in the Tinnitus Questionnaire score of ≥5 points"	Lehner et al. (2013)	Goebel and Hiller (1994)
Criterion of improvement is a global TQ score reduction >6.1 points			
German version	Quote: "a critical difference of 6.1 points"	Argstatter et al. (2015)	Goebel and Hiller (1998)
Criterion of improvement is a global TQ score reduction >10 points			
German version	Quote: "clinically relevant tinnitus improvement … patients who demonstrated reduction of >10	Kleinjung et al. (2008)	Kleinjung et al. (2007)

	points"		
Criterion of improvement is a global TQ score reduction of >= 25% points			
Version of translation unknown	Quote: "when looking at the predefined clinically relevant effect of at least 25% improvement (10 points)"	Hoekstra et al. (2013)	NR
Criterion of improvement is a global TQ score that crosses the boundary from 47 to 46 points			
German version	Quote: "Tinnitus is considered to be 'compensated' at a TQ level of = 46 (no secondary symptoms) and 'decompensated' at a TQ level of = 47 (permanent annoyance and psychological strain)"	Mazurek et al. (2009)	Goebel and Hiller (1999)

Additional references cited in the Supplemental Table, but not cited in the main body of the article:

Argstatter, H., Grapp, M., Hutter, E., Plinkert, P.K., & Bolay, H.V. (2015). The effectiveness of neuro-music therapy according to the Heidelberg model compared to a single session of educational counseling as treatment for tinnitus: a controlled trial. *J Psychosom Res, 78*(3), 285-92.

Chung, H.K., Tsai, C.H., Lin, Y.C., Chen, J.M., Tsou, Y.A., Wang, C.Y., Lin, C.D., Jeng, F.C., Chung, J.G., & Tsai, M.H. (2012). Effectiveness of theta-burst repetitive transcranial magnetic stimulation for treating chronic tinnitus. *Audiol Neurootol, 17*(2), 112-20.

Goebel, G., Hiller, W. (1998). *Tinnitus-Fragebogen:(TF); ein Instrument zur Erfassung von Belastung und Schweregrad bei Tinnitus; Handanweisung*. Hogrefe, Verlag für Psychologie.

Goebel, G., Hiller, W. (1999). Quality management in the therapy of chronic tinnitus. In *Proceedings of the 6th International Tinnitus Seminar*. Hazell JW (ed). London: The Tinnitus and Hyperacusis centre. pp 357-63.

Hoekstra, C. E., Versnel, H., Neggers, S. F., Niesten, M. E., & Van Zanten, G. A. (2013). Bilateral low-frequency repetitive transcranial magnetic stimulation of the auditory

cortex in tinnitus patients is not effective: a randomised controlled trial. *Audiology and Neurotology*, *18*(6), 362-373.

Kleinjung, T., Eichhammer, P., Landgrebe, M., Sand, P., Hajak, G., Steffens, T., & Langguth, B. (2008). Combined temporal and prefrontal transcranial magnetic stimulation for tinnitus treatment: a pilot study. *Otolaryngology--Head and Neck Surgery*, *138*(4), 497-501.

Kreuzer, P. M., Landgrebe, M., Schecklmann, M., Poeppl, T. B., Vielsmeier, V., Hajak, G., & Langguth, B. (2011). Can temporal repetitive transcranial magnetic stimulation be enhanced by targeting affective components of tinnitus with frontal rTMS? A randomized controlled pilot trial. *Frontiers in Systems Neuroscience*, 5, 88.

Kreuzer, P. M., Goetz, M., Holl, M., Schecklmann, M., Landgrebe, M., Staudinger, S., & Langguth, B. (2012). Mindfulness-and body-psychotherapy-based group treatment of chronic tinnitus: a randomized controlled pilot study. *BMC Complementary and Alternative Medicine*, *12*(1), 1.

Landgrebe, M., Binder, H., Koller, M., Eberl, Y., Kleinjung, T., Eichhammer, P., & Langguth, B. (2008). Design of a placebo-controlled, randomized study of the efficacy of repetitive transcranial magnetic stimulation for the treatment of chronic tinnitus. *BMC Psychiatry*, *8*(1), 1.

Langguth, B. (2012a). NCT01663311 Repetitive magnetic stimulation with double cone coil in chronic tinnitus (Ti-CDC).

Langguth, B. (2012b). NCT01663324 rTMS for the treatment of chronic tinnitus: optimisation by stimulation of the cortical tinnitus network (multisite rTMS).

Langguth, B. (2013a). NCT01907022 Combined rTMS and relaxation in chronic tinnitus.

Langguth B. (2013b). NCT01965028 Daily bi-temporal transcranial random noise stimulation in tinnitus (tRNS-tin).

Langguth, B., Landgrebe, M., Frank, E., Schecklmann, M., Sand, P. G., Vielsmeier, V., & Kleinjung, T. (2014). Efficacy of different protocols of transcranial magnetic stimulation for the treatment of tinnitus: pooled analysis of two randomized controlled studies. *The World Journal of Biological Psychiatry*, *15*(4), 276-285.

Lehner, A., Schecklmann, M., Kreuzer, P. M., Poeppl, T. B., Rupprecht, R., & Langguth, B. (2013). Comparing single-site with multisite rTMS for the treatment of chronic tinnitus–clinical effects and neuroscientific insights: study protocol for a randomized controlled trial. *Trials*, *14*(1), 1.

Mazurek, B., Haupt, H., Szczepek, A. J., Sandmann, J., Gross, J., Klapp, B. F., & Caffier, P. P. (2009). Evaluation of vardenafil for the treatment of subjective tinnitus: a controlled pilot study. *Journal of Negative Results in Biomedicine*, *8*(1), 1.

Supplemental File 2. This table illustrates how many patients in our sample meet the different numerical criteria for what constitutes a minimal clinically important difference (MCID). Above the bold line are all those numerical criteria reported in the Technical report. Below the line are two additional criteria that were used only once in our review of 228 clinical trials (Hall et al., 2016a).

Responders are categorised as individuals whose change pre- versus post-treatment change score either meets or exceeds the MCID defined by that statistical method. Non-responders are those individuals whose change pre- versus post-treatment change score is less than the MCID defined by that statistical method. The three Clinical Global Impression (CGI) categories (better, unchanged, worse) form the benchmark for comparing each statistical method with the patients' subjective personal impression. Three further statistical metrics assist interpretation of these data.

Cohen's kappa coefficient measures the degree of agreement between the responder classification based on CGI and that based on the statistical method. Kappa coefficients for some of the methods fell between 0.2 and 0.4 indicated no better than 'fair' agreement (Landis and Koch, 1977). Kappa coefficients for LoA and the boundary method indicated poor agreement.

Sensitivity refers to the percent of responders (classification based on CGI) that were correctly diagnosed by the statistical method, while specificity refers to the percent of non-responders (classification based on CGI) that were correctly diagnosed by the statistical method.

Taken in combination and over the whole group, the SEMCon method showed the highest value of kappa and the best trade-off between sensitivity and specificity. However, the important caveat to this interpretation is that the CGI may not be the 'true' classification of treatment-related improvement because self-reports can be prone to bias.

Method of estimating individual responders	Responders	Non-responders	Kappa	Sensitivity %	Specificity %
‡ CGI	115	85			
LoA & SDC	32	170	0.08	20.0	89.4
slightly better & RCI	90	112	0.19	53.0	67.1
SEM_{Con}	124	78	0.30	73.9	55.3
½ SD	132	70	0.22	74.8	47.1
slightly better – no change	138	64	0.20	76.5	42.4
* boundary for compensated vs decompensated	46	156	0.03	24.3	78.8
percent improvement (at least 25%)	99	103	0.27	60.9	67.1

‡ CGI data available for 200 participants
* In the German version of the TQ, a total score of 47 and above (range = 0-84 points) has been regarded as a clinically significant level of distress (Goebel and Hiller, 1998; Hallam, 2008).

Additional references cited in the Supplemental Table, but not cited in the main body of the article:

Hallam, R.S. (2008). Manual of the Tinnitus Questionnaire (TQ). Revised and updated. London: Polpresa Press.

Landis, J.R. Koch, G.G. (1977). The measurement of observer agreement for categorical data". Biometrics, 33 (1): 159–174.