**Trends in Amplification**
Volume 12 Number 3
September 2008 223-235

10.1177/1084713808319943
http://tia.sagepub.com
hosted at
http://online.sagepub.com

# Tinnitus Outcomes Assessment

Mary B. Meikle, PhD, Barbara J. Stewart, PhD,
Susan E. Griest, MPH, and James A. Henry, PhD

Over the past two decades, recognition has grown that measures for evaluating treatment outcomes must be designed specifically to have high responsiveness. With that in mind, four major types of tinnitus measures are reviewed, including psychoacoustic measures, self-report questionnaires concerning functional effects of tinnitus, various rating scales, and global outcome measures. Nine commonly used tinnitus questionnaires, developed in the period 1980-2000, are reviewed. Because of many similarities between tinnitus and pain, comparisons between pain and tinnitus measures are discussed, and recommendations that have been made for developing a core set of measures to evaluate treatment-related changes in pain are presented as providing a fruitful path for developing a core set of measures for tinnitus. Finally, the importance of having both immediately obtainable outcome measures (psychoacoustic, rating scales, or single global measures) and longer term measures (questionnaires covering the negative effects of tinnitus) is emphasized for further work in tinnitus outcomes assessment.

**Keywords:** tinnitus outcomes; core set of measures; responsiveness of measures; tinnitus questionnaires; tinnitus rating scales

Assessment of tinnitus treatment outcomes has become an important concern because of the relatively high prevalence of "problem" tinnitus. Estimates of the prevalence of bothersome tinnitus in the adult population of the United States have varied from 4.5% (Brown, 1990) to 8.4% (Hoffman & Reed, 2004). This work has demonstrated that such estimates are quite sensitive to the specific wording of questions asking people whether or not they experience tinnitus, but in either case, the U.S. Census data show that significant tinnitus affects many millions of people.

When such people seek treatment, practitioners who offer treatment for tinnitus must demonstrate that there is evidence that the treatments they offer are efficacious, that is, that there is statistical evidence, obtained in appropriate clinical trials, that the offered treatment is capable of providing benefit. The increasing emphasis on evidence-based medicine testifies to the need for such demonstrations in general medical practice (Feussner, 1998; Relman, 1988), in otologic and audiologic contexts (Axelsson, Coles, Erlandsson, Meikle, & Vernon, 1993; Cox et al., 2000; Gatehouse, 2000; McArdle, Chisolm, Abrams, Wilson, & Doyle, 2005; Piccirillo, 1994), and in tinnitus clinical practice (Dobie, 2002; Turk, 2002).

Not all tinnitus treatments are equally efficacious, nor is it necessarily the case that a given treatment will be equally effective for all segments of the clinical population. Chronic, often intractable disorders such as tinnitus (as well as a host of chronic diseases such as arthritis, obstructive pulmonary disease, and treatment-resistant syndromes such as fibromyalgia and other pain disorders) require that clinicians and researchers pay attention even to small clinical improvements, as these may yield insights that can lead to an accumulating repertoire of techniques to provide relief, if not cure. To detect small treatment-related improvements, however, it is essential to have measures of treatment-related change that are sensitive enough to detect small changes. That realization has led to groundbreaking

From Oregon Health & Science University, Portland, Oregon (MBM, BJS, SEG, JAH), and VA National Center for Rehabilitative Auditory Research, Portland, Oregon (JAH).

This work was partially supported by the Tinnitus Research Consortium.

Address correspondence to: Mary B. Meikle, PhD, Department of Otolaryngology/Hearing Research, NRC 04, 3181 SW Sam Jackson Park Road, Portland, OR 97239-3098; e-mail: meiklem@ohsu.edu.

insights into "treatment effectiveness research" (Lipsey, 1990). Within that body of work, a key concept is *responsiveness* of outcome measures.

## Growing Recognition of Responsiveness as a Key Aspect of Outcome Measures

In recent years, there has been growing recognition that there are important differences between measures designed for intake evaluation and screening purposes versus measures designed for assessing treatment outcomes. The differences were summarized in a seminal article dealing with treatment evaluation in chronic disease (Kirshner & Guyatt, 1985), in which the authors used the term *discriminative* to refer to measures designed to evaluate differences between individual patients at a single point in time (e.g., differences in regard to the perceived severity of the condition or to the particular functional problems affecting them). They introduced the term *evaluative* to refer to measures designed for evaluating longitudinal changes over time, either in regard to severity or to relevant functional problems. They emphasized the point that outcome measures must be designed specifically to have high sensitivity to treatment-related change, which is termed *responsiveness*. Since then, the importance of developing outcome measures with specific attention to their responsiveness has received detailed discussion in a variety of contexts including, among others, psychotherapy (Vermeersch, Lambert, & Burlingame, 2000), studies of caregiver role strain and effectiveness (Stewart & Archbold, 1992, 1993), and chronic pain (S. F. Dworkin & Sherman, 2001).

The importance of responsiveness as the essential aspect of outcome measures has received especially thorough discussion in a valuable monograph dedicated to that topic (Lipsey, 1990). The responsiveness of a measure can be expressed in terms of *effect size*, which is obtained empirically. To measure its effect size, the measure in question must be tested in one or more appropriate samples of respondents who undergo treatment. Effect sizes are computed as a ratio in which a mean difference score (e.g., the mean for a treatment group minus the mean for a placebo group) forms the numerator (or the numerator could be the difference between pre- and posttreatment means for a single group). The denominator of the ratio is the average standard deviation[1] for the two groups (or the standard deviation of the difference scores for pre- vs. posttreatment comparisons). Expressing treatment effects in standard deviation units provides a more unbiased comparison between treatment results obtained in different places or at different times. Because effect sizes are obtained empirically, they may be affected by circumstances such as the length of the posttreatment interval, the effectiveness of the treatment, and other variables affecting treatment outcomes. Therefore, effect sizes obtained in a given study must be viewed as estimates of the responsiveness of the relevant outcome measures.

Effect sizes can be larger than 1.0 (i.e., exceed one standard deviation or more), but in most cases, they tend to be smaller. A rule-of-thumb estimate for interpreting effect sizes was provided by Cohen (1988), whose categories have been widely quoted and used: Effect sizes less than 0.2 would be considered inconsequential; those between 0.2 and 0.5 considered small; between 0.5 and 0.8, moderate; and above 0.8, effect sizes are considered large. To maximize the responsiveness of outcome measures, their effect sizes should be as large as possible.

## Measurement Methods for Tinnitus

Turning now to techniques used to evaluate tinnitus treatment outcomes, in broad outlines, the literature reveals at least four different approaches that have traditionally been used to evaluate proposed tinnitus treatments (summarized in Table 1). Formal studies of tinnitus treatments tend to rely on the use of rating scales and/or questionnaires describing the functional and emotional effects of tinnitus, probably because both types of measures can be acquired relatively rapidly, with good reliability and validity, and may require little or no examiner involvement (Newman & Sandridge, 2004). Psychoacoustic measures of the sensory aspects of tinnitus (mainly its loudness and maskability) are less commonly used to quantify treatment-related changes in tinnitus, in part because they are more time consuming and require specific acoustic equipment and protocols (Henry & Meikle, 2000; Tyler, 2000). Patients' global ratings of perceived improvement following treatment have also been used in a number of studies, although the rating methods are far from uniform, and there is little information on reliability and validity of such ratings.

It is helpful to consider the strengths and weaknesses of each type of measure listed in Table 1. An important caveat is that information about effect

**Table 1.** Techniques for Evaluating Outcomes of Tinnitus Treatment

| Measurement Technique | Example | Use for Rapidly Acting Treatments? | Use for Slowly Acting Treatments? |
|---|---|---|---|
| Psychoacoustic tests of tinnitus | Pitch match | X | X |
| | Loudness matches | X | X |
| | Maskability (minimum masking levels) | X | X |
| | Residual inhibition | X | X |
| Rating scales | Verbal Rating Scale | X | X |
| | Numerical Rating Scale | X | X |
| | Visual Analog Scale | X | X |
| | Other (e.g., "poster" style with faces showing increasing unhappiness/ discomfort, mechanical devices requiring patient to manipulate indicator, etc.) | X | X |
| Questionnaires describing functional effect(s) of tinnitus | Self-administered | — | X |
| | Examiner administered | — | X |
| Patients' global perception of treatment-related change | Self-administered | X | X |
| | Examiner administered | X | X |

sizes has rarely been available for most of the tinnitus measures used to date.

## Psychoacoustic Measures of Treatment-Related Change

Psychoacoustic measurement techniques for describing patients' tinnitus at intake have been used since the 1940s (Fowler, 1943), and a formal recommendation to use a standard battery of four such measures was published as an appendix to the CIBA symposium on tinnitus in 1981 (Evered & Lawrenson, 1981). The four measures, including pitch matches, loudness matches (LMs), minimum masking levels (MMLs), and residual inhibition (temporary reduction or suppression of tinnitus), are shown in Table 1. All four measures remain in use today for intake evaluation. Although there are other psychoacoustic tests for tinnitus, they are not commonly implemented in clinical settings. For evaluating treatment effects, tinnitus LMs and MMLs have proven to be most useful.

Several recent reviews of psychoacoustic measurement techniques have been published, providing useful detail as well as historical background and current clinical recommendations (Henry & Meikle, 2000; Tyler, 2000; Vernon & Meikle, 2003). LMs are considered most useful if performed at 1000 Hz, that is, patients are asked to match an external 1000 Hz tone to the loudness of their tinnitus. The LM is usually stated in dB Sensation Level (dB SL, or dB above threshold for the tone). Although most tinnitus patients match the pitch of their tinnitus to high-frequency tones (generally above 4000 Hz; see Meikle, Creedon, & Griest, 2004), it is not a difficult task for patients to provide LMs at 1000 Hz. When tested repeatedly to evaluate test–retest reliability, the large majority of tinnitus patients are very reliable at producing LMs, probably because they have an "internal standard" to compare with the external tone (Vernon, 1996). Most LMs are not large—two thirds of the clinic population select LMs that are no greater than 15 dB SL (Meikle et al., 2004).

MMLs are typically obtained using a broadband noise as an external stimulus and adjusting it in small steps (1-2 dB) until patients report that they can no longer hear their tinnitus. Not all patients experience complete masking (some report only partial masking regardless of the level of the external noise); however, most patients do experience complete masking, and for nearly 80% of patients, the MMLs are at or below 15 dB (Meikle et al., 2004). The test–retest reliability of MMLs has been shown to be good, if care is taken to ensure that stimuli prior to masking have not caused long-lasting

changes in the tinnitus (e.g., residual inhibition or, in rare cases, exacerbation of the tinnitus).

In summary, with appropriate attention to careful psychoacoustic technique, the reliability of both LMs and MMLs has been found to be good (Vernon & Meikle, 2003). Both LMs and MMLs have been found to be significantly reduced following effective treatment for tinnitus (P. Jastreboff, Hazell, & Graham, 1994; Johnson, Brummett, & Schleuning, 1993; McKinney, Hazell, & Graham, 1999a, 1999b; Saito et al., 1999). To date, however, information about effect sizes for LMs and MMLs is not routinely presented, and in some reports, the relevant data are presented in a metric that makes it difficult to calculate effect sizes (e.g., standard deviations shown only as error bars in a graph). It is to be hoped that further research using these measures in tinnitus outcome studies will provide better information on their responsiveness.

Practical difficulties in using psychoacoustic measures include the following: (a) an audiometric or comparable test booth needs to be set up with equipment for that purpose, (b) appropriate staff need to be trained to conduct the measures reliably and rapidly, (c) the measures take significant time even when obtained by a well-trained clinician, and (d) not all clinic patients can provide the desired data. In addition, because information about the responsiveness of psychoacoustic measures is hard to find, the ability to calculate statistical power and the requisite sample size for a proposed test using such measures is currently limited. Despite those difficulties, psychoacoustic techniques offer an appealing method to quantify treatment effects in terms of a well-established measurement metric (decibels). An innovative computerized method for obtaining psychoacoustic measures of tinnitus has been developed (Henry, Flick, Gilbert, Ellingson, & Fausti, 1999).[2]

Even without computer-assisted methods, LMs are considered particularly reliable measures (Vernon, 1996) and lend themselves well to clinical trials where patients can be tested several times to establish the reliability of their responses prior to initiating any potentially tinnitus-reducing treatment (Johnson et al., 1993). Furthermore, LMs can be obtained immediately or within a few minutes of administering rapidly acting treatments such as intravenous lidocaine, transcutaneous electrical stimulation, transcranial magnetic stimulation, and others. That fact makes them an ideal choice for use in such studies, where questionnaires concerning functional effects of tinnitus (such as sleep disturbance or difficulties at work) cannot be used because they require longer time intervals for adequate observation. Thus, for a variety of reasons, psychoacoustic measures seem well suited to certain types of tinnitus outcomes research.

## Rating Scales for Numerical Estimates of Tinnitus Severity and Negative Effect

The use of rating scales seems simple and straightforward at face value, as exemplified in Figure 1, which shows typical examples, including (a) a Verbal Rating Scale (VRS), (b) a Numeric Rating Scale (NRS), and (c) a Visual Analog Scale (VAS).

*Verbal Rating Scales*. Scales such as the VRS in Figure 1a may include any number of levels or response options, listed in ascending or descending order of severity. When entered into a database or other method of performing statistical calculations, the response levels are assigned progressively increasing or decreasing numerical values corresponding with their physical location in the list of response options. The problem with relatively short scales (e.g., with only 3 or 5 response levels) when used as outcome measures is that the measures achieved do not have very high resolution and, therefore, are lacking in precision, which can limit their ability to detect small changes following treatment. Verbal scales with as many as 15 different response levels have been devised (cf. Jensen & Karoly, 2001), but they take up considerable space on a printed page, which makes it difficult to compose a questionnaire that provides adequate coverage of the material without being overly bulky and unwieldy. They also require a higher level of reading skill and comprehension, with the resulting danger that some respondents may tire or lose focus before reading the complete range of response options for a given question and, thus, select a response that doesn't accurately reflect their status. The advantage of VRSs is that they don't require fluency with numbers so they may be more appropriate for less sophisticated respondents (e.g., children or individuals with cognitive impairment).

*Numeric Rating Scales*. NRSs (see Figure 1b) typically include a linear array of numbers, either increasing or decreasing from left to right (or in a vertical direction), with a verbal anchor at each end of the scale to specify the intended range of values

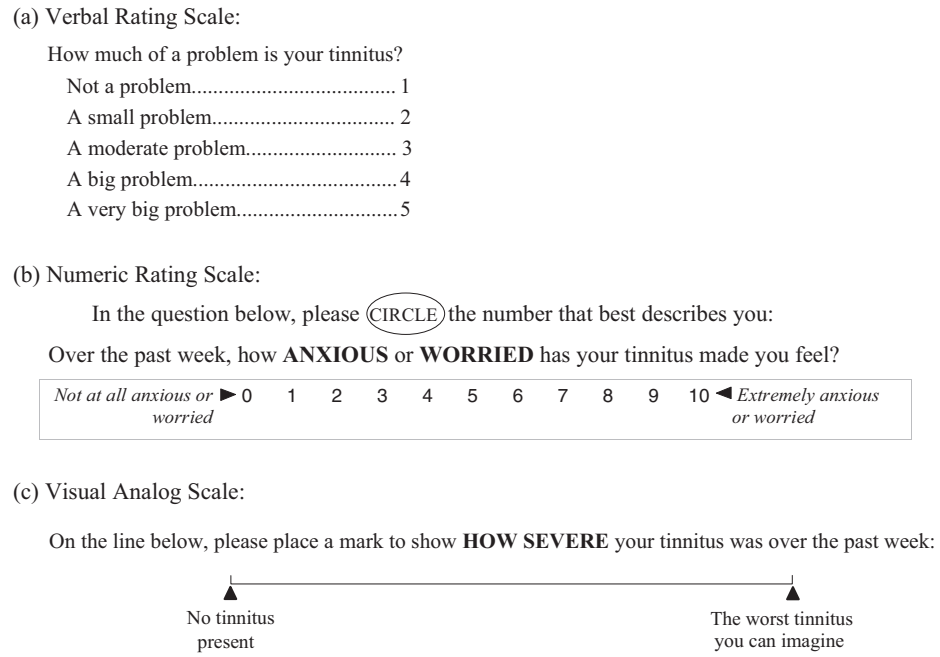


**Figure 1.** Three commonly used types of rating scales.

and explain what the extreme values are (in Figure 1b, "Not at all anxious or worried" is represented by 0; "Extremely anxious or worried" is represented by the maximum value of 10). Common ranges for NRSs include 0 to 10 (an 11-point scale) and 0 to 100 (a 101-point scale), but in theory, the numeric range can be any convenient set of numbers. Instructions to the respondent should emphasize the need to make the mark very clear (e.g., circling a number or placing an X over the chosen number) and also what to do if the respondent has variable tinnitus necessitating the choice of more than one number. Advantages of using the NRS approach to scaling tinnitus are that it is very economical of space, can provide high measurement resolution, is easy to score, and can be a very rapid method for eliciting responses. Those advantages assume that the respondent is numerically adept and has no difficulty deciding about where on the scale to mark the appropriate response. Not all respondents match that description, and it is up to the examiner to determine whether the respondent understands the task and is able to perform it with assurance. Therefore, the NRS approach may not be suitable for situations where there is little prior opportunity to evaluate the respondents' ability to complete the task.

*Visual Analog Scales*. The VAS approach to scaling tinnitus employs a straight line (the visual analog for such continua as the loudness or the severity of tinnitus), without any numbers. The standard length of the line is 10 cm, and there should be verbal anchors at each end of the line to indicate what the intended range of values is. The respondent receives instructions to mark the line clearly at the point that best corresponds with his or her tinnitus (the instructions specifying whether it is loudness, for example, or severity that is being scaled). After the respondent has completed the response, the examiner measures the position of the mark in millimeters, which provides a scale of 0 to 100 in 1 mm steps. It has traditionally been felt that VASs provide maximum precision compared with other types of scales, however, there is considerable evidence that they are difficult for some patients to respond to, in particular the elderly or those with some degree of physical or visual impairment (Gagliese, 2001; Guyatt, Townsend, Berman, & Keller, 1987; Jensen & Karoly, 2001). It has been reported that patients prefer VRSs or NRSs to the VAS type of scale. In their detailed review of such scales, Jensen and Karoly (2001) emphasize that "careful explanation and patient practice with the scale may decrease the

failure rate," suggesting that where a substantial proportion of the respondents are likely to be in the older decades of life, use of VASs may entail a greater investment of examiner and respondent effort than other types of scales.

There are some less commonly used scales that might be used for evaluating the severity as well as other aspects of tinnitus (its loudness, pitch, aversiveness, etc.), such as graphic scales depicting faces representing different levels of distress (Jensen & Karoly, 2001). Although graphic scales are frequently used for scaling the intensity of pain, there is little information available on their use with tinnitus. The disadvantage to scales using pictures, faces, or other graphic symbols is that it becomes difficult to render a sufficient number of discernibly different response levels as the number of levels exceeds about eight. Such scales, therefore, limit the scale resolution to less than 10, thus reducing precision.

Like psychoacoustic measures, rating scales can be used to evaluate rapidly acting tinnitus treatments as they require very little time for the respondent to register a response. And, like psychoacoustic measures, there is little or no information concerning their responsiveness to treatment-related changes in tinnitus.

## Global Measures of Treatment-Related Improvement

It is possible to question patients directly concerning the occurrence of improvement in their tinnitus, and a number of tinnitus treatment studies have done so. For example, in a study of tinnitus masking techniques, respondents were asked, "What is your overall feeling about the effect of tinnitus upon your life since you first came to the clinic?" and response options were *much better*, *considerably better*, *slightly better*, *much the same*, *slightly worse*, and *considerably worse* (Hazell et al., 1985). Other examples from a study of a potential tinnitus-reducing drug included, "Has the medication helped you in any way?" and "Has your tinnitus improved?" (the response options were not described; Dobie, Sakai, Sullivan, Katon, & Russo, 1993). The use of a single global question such as these for assessing treatment outcomes presents two potential problems: First, an outcome measure made up of a single item or question is not as statistically reliable as a measure involving a number of items (such as a multi-item questionnaire). Second, if the global measure provides a small number of response levels, the resolution for measuring treatment effects will be lower than optimal and the result may be to generate lower effect sizes than would have been obtained using a higher resolution measure.

Since the last time you filled out our questionnaire, which was about 3 months ago, how would you describe your overall tinnitus status:

*(Please CIRCLE only ONE answer below)*

| | |
|---|---|
| Very much improved | 1 |
| Much improved | 2 |
| Moderately improved | 3 |
| Slightly improved | 4 |
| No change | 5 |
| Slightly worse | 6 |
| Moderately worse | 7 |
| Much worse | 8 |
| Very much worse | 9 |

**Figure 2.** Global measure of patient's perception of change following treatment.

An example from the tinnitus clinic at Oregon Health & Science University illustrates the use of a global measure having nine response levels (see Figure 2), which provides quite good resolution. In this particular case, the global question was not used as the primary outcome measure but instead was used as a "grouping" factor to divide the respondents into three different groups based on their answers to the global item. The Improved group were those who chose responses from 1 to 4; the Same group chose the response of 5; and the Worse group were those who chose responses from 6 to 9. Respondents' scores on a tinnitus questionnaire were then compared for the three different groups to determine whether the effect sizes for the tinnitus questionnaires reflected the expected differences between groups with regard to both magnitude and sign, which they did.

Although use of a single global measure as the primary outcome measure to assess treatment outcomes has appeal because of its rapidity and simplicity, systematic information about how well such measures perform in clinical trials seems to be scarce. Not only is it difficult to find information about effect sizes in many published studies (because presentations of data have omitted standard deviations from which effect sizes could be computed) but, in addition, information about the reliability and validity of global measures seems to be scarce. Lacking such information, clinicians and investigators have little guidance in identifying

**Table 2.** Nine Principal English-Language Questionnaires for Evaluating the Clinical Status of Tinnitus

| Questionnaire Title | Authors and Year | Number of Items | Response Format |
|---|---|---|---|
| Tinnitus Questionnaire | Hallam, Jakes, and Hinchcliffe (1988) | 33 | 3 levels: *true, partly true, not true* |
| Tinnitus Handicap Questionnaire | Kuk, Tyler, Russell, and Jordan (1990) | 15 | 0–100 (0 = *strongly disagree,* 100 = *strongly agree*) |
| Tinnitus Severity Scale | Sweetow and Levy (1990) | 27 | 4 levels; response choices vary between questions |
| Subjective Tinnitus Severity Scale | Halford and Anderson (1991) | 16 | yes/no |
| Tinnitus Reaction Questionnaire | Wilson, Henry, Bowen, and Haralambous (1991) | 5 | 5 levels: *not at all, a little of the time, some of the time, a good deal of the time, almost all the time* |
| Tinnitus Severity Grading | Coles, Lutman, Axelsson, and Hazell (1992) | 9 | 5 levels; response choices vary between questions |
| Tinnitus Severity Index | Meikle (1992); Meikle, Griest, Stewart, and Press (1995) | 12 | 3-4 levels; response choices varied between questions (original version); changed to 5 levels in later version |
| Tinnitus Handicap Inventory | Newman, Jacobson, and Spitzer (1996) | 25 | 3 levels: yes, sometimes, no |
| Intake Interview for Tinnitus Retraining Therapy | M. M. Jastreboff and Jastreboff (1999) | 12 | response formats and levels vary between questions |

reliable, valid, and responsive examples of such measures.

## Self-Report Questionnaires for Scaling the Severity and Negative Effect of Tinnitus

Tinnitus that has reached the level of being a clinical problem is a multidimensional disorder in that many aspects of life are likely to be negatively affected. Although the functional and emotional effects of tinnitus are complex, they can be reliably quantified using self-report questionnaires that allow patients to rate the extent of tinnitus-related dysfunction separately for each such domain. The fact that the questionnaires include multiple items all addressing the underlying construct of tinnitus severity increases their reliability compared with single-item measures (Nunnally & Bernstein, 1994).

During the 1980s and 1990s, a number of questionnaires were designed to evaluate functional, emotional, and other effects of tinnitus, as summarized in Table 2 (cf. more detailed discussions in Meikle & Griest, 2002; Newman & Sandridge, 2004). Together, these nine questionnaires encompass a total of 175 separate questions or items. Each questionnaire was developed through interviews and preliminary questionnaires administered to relatively large numbers of tinnitus patients. This set of questionnaires represents the combined clinical experience of a number of different investigators at widely separated geographic locations (including the United Kingdom and Australia, in addition to the states of California, Georgia, Iowa, Oregon, and Michigan). The nine questionnaires thus embody a large and valuable store of information about the effects of tinnitus in a heterogeneous group of patients.

A number of the questionnaires in Table 2 meet high standards for validity and reliability for discriminative purposes, that is, for intake assessment of tinnitus. Test–retest reliability and internal consistency reliability were evaluated in many cases. Convergent validity was established through comparisons between the questionnaires and other measures considered to be reliable indicators of the severity of tinnitus, such as whether patients complained or did not complain of "problem" tinnitus (Hallam, Jakes, & Hinchcliffe, 1988); negative correlations with "life satisfaction" (Kuk, Tyler, Russell, & Jordan, 1990); positive correlations with subjectively perceived loudness of tinnitus (Meikle, 1992), depression (Newman, Jacobson, & Spitzer, 1996), and/or anxiety (Wilson, Henry, Bowen, & Haralambous, 1991); and independent clinical ratings of patients' negative reactions to tinnitus (Halford & Anderson, 1991).

For some of the later questionnaires, convergent validity was also established by demonstrating high correlations with previous tinnitus questionnaires such as the Tinnitus Handicap Questionnaire (Kuk et al., 1990) or the Tinnitus Questionnaire (Hallam et al., 1988).

It has traditionally been assumed that questionnaires such as those in Table 2 can serve as effective measures of treatment-related changes in tinnitus. However, because the nine existing tinnitus questionnaires were developed prior to the current recognition that responsiveness is an essential design goal, they were not specifically designed to maximize responsiveness to treatment-related change. As Table 2 shows, most of the questionnaires used relatively coarse response scales (i.e., involving five or fewer response levels) and, therefore, were less than optimal for achieving high sensitivity to treatment effects. Some of the questionnaires include one or more items that have been shown to be nonresponsive in carefully controlled clinical trials (Henry, 2008), and some address only a limited selection of the functional domains affected by tinnitus, thus restricting their ability to show treatment-related changes with regard to functional domains that are omitted.

Published information with regard to effect sizes in this group of questionnaires is almost nonexistent. Although some of them have been demonstrated to be capable of detecting treatment-related changes (e.g., Newman, Sandridge, & Jacobson, 1998), there is little systematic evidence concerning effect sizes at the item, subscale, or overall scale level for any of the questionnaires. It is likely that improvements in responsiveness for most of these questionnaires could be achieved by increases in scale resolution coupled with careful selection of items to maximize sensitivity to treatment-related change (eliminating those that are less likely to change). It has become increasingly evident that newer approaches to tinnitus outcomes evaluation, by including greater emphasis on responsiveness in their design, can probably provide substantially better performance for outcomes assessment.

Because the criteria for selecting well-functioning questionnaire items are different for intake or diagnostic purposes than for evaluating treatment outcomes, it is a challenge to develop questionnaires that have high validity for both purposes. Items included in intake or diagnostic questionnaires are typically designed to emphasize differences in tinnitus severity between people or to identify specific treatment needs that may vary across the clinical population. In contrast, items selected for outcome questionnaires must maintain a primary emphasis on responsiveness to treatment-related change. Despite those differences, experts in measurement science have stated that with careful development techniques, it should be possible to achieve both objectives in a single questionnaire (Kirshner & Guyatt, 1985; Nunnally, 1978). In confirmation of that statement, recent work has provided preliminary evidence that a single self-report questionnaire can, in fact, achieve high validity for both measurement objectives (Meikle et al., 2008).

## What We Can Learn From Pain Outcomes Assessment

An important recent trend in outcomes assessment is the concept of a standardized core set of variables that could be adopted widely to promote greater standardization of outcomes evaluation techniques (Tugwell & Boers, 1993). A *core set* is a set of variables that (a) are identified as being important for characterizing a chronic disease or condition and (b) are demonstrated to be responsive to change. The motivation for standardizing a core set of measures for outcomes evaluation is that there is a multiplicity of different ways to measure treatment outcomes, and as a result, it is difficult to compare results obtained at different clinics or by different investigators. For example, the lack of standardization of pain outcome measures has impeded progress by making it difficult or impossible to compare results obtained at different places (Turk et al., 2003). To address that problem, specialists in pain research and pain clinics have formed a group entitled Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) to develop consensus concerning identification of core domains for evaluating pain outcomes (R. H. Dworkin et al., 2005).

A detailed discussion of the need for core sets of measures was presented in a 1997 article dealing with the need for more standardized measures in treatment of ankylosing spondylitis (van der Heijde et al., 1997). They asked, "Why do we need a core set of suitable endpoint measures?" and their answers can be summarized in the following list:

1. When there are many different candidate variables used as outcome measures, statistically significant changes may be reported that are not true differences but result from chance alone.

2. Investigators may choose to present only those variables that show impressive results.
3. Variables may be used because they are traditional, even though they may be insensitive to treatment-related change.
4. Use of different endpoint measures in different studies makes it difficult to pool results in meta-analyses; as a result, potentially useful therapies may be difficult or impossible to compare directly.
5. Use of a core set of measures would, in effect, achieve standardization between clinics, allowing direct comparison of results in different demographic groups and enhancing the generalizability of results.

Work with pain outcome measures is particularly relevant for tinnitus, because both are positive neurological symptoms that add new sensations (as opposed to negative neurological symptoms, such as impairment of vision or of hearing, that represent reductions or loss of normal sensation). There have been many discussions of the similarities between tinnitus and pain (Meikle, 1995; Moller, 2007; Vernon & Meikle, 1985). As was succinctly stated by Moller (2007), "Since much more is known about pain than about tinnitus, it is valuable to take advantage of the knowledge about pain in efforts to understand the pathophysiology of tinnitus and find treatments for tinnitus" (p. 47).

To underscore that statement, Table 3 compares negative effects of chronic tinnitus (shown in Column 1) with the negative effects of chronic pain (shown in Column 2). It is clear that both chronic conditions share very similar patterns of negative effect, including difficulties sleeping, concentrating, participating in social activities, and others.

Work on pain measurement has generated many insights that could be helpful for research and clinical work on tinnitus (Turk & Melzack, 2001). For example, Table 4 presents a summary of the steps recommended by the IMMPACT group for developing new outcome measures.

If, as we hope, the community of tinnitus clinicians and investigators can agree to work toward a standardized core set of measures for evaluating tinnitus before and after treatment, the systematic development procedure described in Table 4 could serve the very useful purpose of guiding such efforts.

## Summary and Conclusions

Among the various outcome measures for tinnitus that can be found in the literature, it is clear that questionnaires describing the functional effects of tinnitus, as well as psychoacoustic measures of tinnitus loudness and maskability, have both received considerable attention intended to systematize the relevant measurement methods. However, neither type of measure has yet received much attention to quantifying their responsiveness to treatment-related change. Instead, so far, they have been developed primarily as measures of differences between individual patients at a single point in time, to permit scaling of the presenting characteristics of tinnitus.

In further work, it is to be hoped that investigators will address the need for information about the responsiveness of all the various types of tinnitus measures that were listed in Table 1. It is clear that the psychoacoustic measures of tinnitus, besides providing valuable data on the sensory aspects of tinnitus using a well-established metric with values on an interval scale (decibels), also deal most directly with that aspect of tinnitus that can be described as *impairment*, using the World Health Organization (1980) approach to classification of impairment, disability, and handicap. The psychoacoustic measures, as well as the various types of rating scales and single-item global measures of tinnitus severity, all share the virtue that they can be obtained quickly in studies of treatments that produce immediate reduction or relief of tinnitus (i.e., diminution of the impairment represented by tinnitus). For that reason alone, it is important to develop a knowledge base concerning their relative strengths and weaknesses with regard to reliability, validity, and responsiveness to treatment effects.

Although multiple-item questionnaires require much longer time periods over which to evaluate the many potential functional and emotional effects of tinnitus, they offer an important quantitative method for evaluating differences between individuals with regard to the effects of tinnitus on their daily life—on sleeping, cognitive performance, emotional status, and other reactions to the distress caused by tinnitus. The negative functional effects of tinnitus can be subsumed under the headings of *disability* or *handicap* (again, using the 1980 World Health Organization classification) and, thus, form a method of scaling the severity of tinnitus that is complementary to that obtained using psychoacoustic measures.

Clinically, there appear to be significant differences between patients with regard to the degree to which they focus on the sensory versus the functional effects of their tinnitus. The fact that measures of

**Table 3.** Comparison of Negative Effects of Tinnitus and Pain

| Negative Effects of Tinnitus[a] | Negative Effects of Pain[b] |
|---|---|
| 1. Sleep disturbance | 1. Falling asleep at night |
| 2. Difficulty concentrating | 2. Staying asleep at night |
| 3. Difficulty ignoring tinnitus | 3. Sex life |
| 4. Irritability, nervousness | 4. Caring for family |
| 5. Tension, stress | 5. Relations with family, significant others |
| 6. Reduced quality of life | 6. Relations with friends |
| 7. Interference with relaxing | 7. Work |
| 8. Interference with quiet leisure activities | 8. Household responsibilities |
| 9. Interference with social activities (family, friends, going out) | 9. Planning activities |
| 10. Depression | 10. Participating in family activities |
| 11. Anxiety | 11. Participating in recreation & social activities |
| 12. Interference with work | 12. Physical activities (walking, bending, lifting, etc.) |
| 13. Anger, frustration, annoyance | 13. Hobbies |
| 14. Discomfort in quiet | 14. Enjoyment of life |
| 15. Reduced sense of control | 15. Emotional well-being (feeling sad, depressed, less motivated) |
| 16. Inability to cope | 16. Fatigue, feeling tired |
| 17. Interference with hearing | 17. Weakness |
| 18. Feeling tired, ill, fatigued | 18. Difficulty concentrating |
| 19. Unhappiness, distress | 19. Difficulty remembering things |

[a]Listed in descending order according to frequency of mention in questionnaires. Items included in only one questionnaire were omitted.
[b]Listed in descending order according to importance of interference as rated by patients. (Adapted from Turk et al., in press.)

**Table 4.** Recommended Procedures for Developing Outcome Measures in Pain Assessment

1. Identify overall question being addressed and scope of assessment.
2. Determine the target population and specific goal(s) of assessment.
3. Establish factors or concepts to be covered in the instrument.
4. Develop the pool of candidate items (from literature search; patient interviews; focus groups with patients; focus groups with experts).
5. Determine item formats and appropriate scaling.
6. Include planning for data collection, scoring, and analysis.
7. Select items for inclusion; attempt to minimize respondent and examiner burdens (discuss number of items, verbal and/or numerical skill levels required).
8. Test the selected items in representative small sample—evaluate ease of use, clarity of instructions, potential for misunderstanding, or other errors.
9. Revise and retest as needed to finalize format and item wording.
10. Field-test items in larger group within target population (assess scoring performance, identify missing or incomplete data, review poorly working items); revise as necessary.
11. Instrument evaluation—evaluate psychometric properties (reliability, validity, responsiveness) in various target populations to evaluate generalizability.
12. Complete the instrument development—provide further revisions as necessary.

NOTE: Adapted from Turk et al., 2006.

those two aspects of tinnitus experience—sensory impairment versus functional disability and handicap—each provide unique insights into treatment-related changes in tinnitus reinforces the notion that both approaches are needed for insightful assessment of tinnitus treatment outcomes.

## Notes

1. More precisely, the denominator should be the pooled standard deviation for the two groups (cf. Lipsey, 1990); however, use of a simple average of the group standard deviations provides an acceptable "quick" estimate that tends to understate the observed effect size.

2. Validation was done by comparison with standard manual methods for psychoacoustic measures of tinnitus, and the method is now routinely used at a number of other clinical centers.

## References


Axelsson, A., Coles, R.R.A., Erlandsson, S. I., Meikle, M., & Vernon, J. (1993). Evaluation of tinnitus treatment: Methodological aspects. *Journal of Audiological Medicine*, *2*, 141-150.

Brown, S. C. (1990). *Older Americans and tinnitus: A demographic study and chartbook*. Washington, DC: Gallaudet Research Institute.

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

Coles, R.R.A., Lutman, M. E., Axelsson, A., & Hazell, J.W.P. (1992). Tinnitus severity gradings: Cross-sectional studies. In J.-M. Aran & R. Dauman (Eds.), *Tinnitus 91: Proceedings of the Fourth International Tinnitus Seminar* (pp. 453-455). New York: Kugler.

Cox, R., Hyde, M., Gatehouse, S., Noble, W., Dillon, H., Bentler, R., et al. (2000). Optimal outcome measures, research priorities, and international cooperation. *Ear & Hearing*, *21*, 106S-115S.

Dobie, R. A. (2002). Randomized clinical trials for tinnitus: Not the last word? In R. Patuzzi (Ed.), *Proceedings of the Seventh International Tinnitus Seminar* (pp. 3-6). Perth: University of Western Australia.

Dobie, R. A., Sakai, C. S., Sullivan, M. D., Katon, W. J., & Russo, J. (1993). Antidepressant treatment of tinnitus patients: Report of a randomized clinical trial and clinical prediction of benefit. *American Journal of Otolaryngology*, *14*, 18-23.

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*, *113*, 9-19.

Dworkin, S. F., & Sherman, J. J. (2001). Relying on objective and subjective measures of chronic pain: Guidelines for use and interpretation. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (2nd ed., pp. 619-638). New York: Guilford.

Evered, D., & Lawrenson, G. (1981). Appendix II. Guidelines for recommended procedures in tinnitus testing. In D. Evered & G. Lawrenson (Eds.), *CIBA Foundation Symposium 85: Tinnitus* (pp. 303-306). London: Pitman Books.

Feussner, J. (1998). Clinical research in the Department of Veterans Affairs: Using research to improve patient outcomes. *Journal of Investigative Medicine*, *46*, 264-267.

Fowler, E. P. (1943). Control of head noises: Their illusions of loudness and timbre. *Archives of Otolaryngology*, *37*, 391-398.

Gagliese, L. (2001). Assessment of pain in elderly people. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (2nd ed., pp. 119-133). New York: Guilford.

Gatehouse, S. (2000). The impact of measurement goals on the design specification of outcome measures. *Ear & Hearing*, *21*, 100S-105S.

Guyatt, G. H., Townsend, M., Berman, L. B., & Keller, J. L. (1987). A comparison of Likert and visual analogue scales for measuring changes in function. *Journal of Chronic Diseases*, *40*, 1129-1133.

Halford, J.B.S., & Anderson, S. D. (1991). Tinnitus severity measured by a subjective scale, audiometry and clinical judgement. *Journal of Laryngology and Otology*, *105*, 89-93.

Hallam, R. S., Jakes, S. C., & Hinchcliffe, R. (1988). Cognitive variables in tinnitus annoyance. *British Journal of Clinical Psychology*, *27*, 213-222.

Hazell, J.W.P., Wood, S. M., Cooper, H. R., Stephens, S.D.G., Corcoran, A. L., Coles, R.R.A., et al. (1985). A clinical study of tinnitus maskers. *British Journal of Audiology*, *19*, 65-146.

Henry, J. A. (2008). [Observations concerning item analysis of previous tinnitus questionnaires]. Unpublished raw data.

Henry, J. A., Flick, C. L., Gilbert, A. M., Ellingson, R. M., & Fausti, S. A. (1999). Fully-automated system for tinnitus loudness and pitch matching. In J. Hazell (Ed.), *Proceedings of the Sixth International Tinnitus Seminar* (pp. 520-521). London: Tinnitus and Hyperacusis Centre.

Henry, J. A., & Meikle, M. B. (2000). Psychoacoustical measures of tinnitus. *Journal of the American Academy of Audiology*, *11*, 138-155.

Hoffman, H. J., & Reed, G. W. (2004). Epidemiology of tinnitus. In J. B. Snow (Ed.), *Tinnitus: Theory and management* (pp. 16-41). London: BC Decker.

Jastreboff, M. M., & Jastreboff, P. J. (1999). Questionnaires for assessment of the patients and treatment outcome. In J. Hazell (Ed.), *Proceedings of the Sixth International Tinnitus Seminar* (pp. 487-490). London: Tinnitus and Hyperacusis Centre.

Jastreboff, P., Hazell, J.W.P., & Graham, R. L. (1994). Neurophysiological model of tinnitus: Dependence of the minimal masking level on treatment outcome. *Hearing Research*, *80*, 216-232.

Jensen, M. P., & Karoly, P. (2001). Self-report scales and procedures for assessing pain in adults. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (2nd ed., pp. 15-34). New York: Guilford.

Johnson, R. M., Brummett, R., & Schleuning, A. (1993). Use of alprazolam for relief of tinnitus: A double blind study. *Archives of Otolaryngology—Head and Neck Surgery*, *119*, 842-845.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases*, *38*, 27-86.

Kuk, F. K., Tyler, R. S., Russell, D., & Jordan, H. (1990). The psychometric properties of a Tinnitus Handicap Questionnaire. *Ear & Hearing*, *11*, 434-445.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

McArdle, R., Chisolm, T. H., Abrams, H. B., Wilson, R. H., & Doyle, P. J. (2005). The WHO-DAS II: Measuring

outcomes of hearing aid intervention for adults. *Trends in Amplification*, *9*, 127-143.

McKinney, C. J., Hazell, J.W.P., & Graham, R. L. (1999a). An evaluation of the TRT method. In J.W.P. Hazell (Ed.), *Proceedings of the Sixth International Tinnitus Seminar* (pp. 99-105). London: Tinnitus and Hyperacusis Centre.

McKinney, C. J., Hazell, J.W.P., & Graham, R. L. (1999b). The effects of hearing loss on tinnitus. In J.W.P. Hazell (Ed.), *Proceedings of the Sixth International Tinnitus Seminar* (pp. 407-414). London: Tinnitus and Hyperacusis Centre.

Meikle, M. B. (1992). Methods for evaluation of tinnitus relief procedures. In J.-M. Aran & R. Dauman (Eds.), *Tinnitus 91: Proceedings of the Fourth International Tinnitus Seminar* (pp. 555-562). New York: Kugler.

Meikle, M. B. (1995). The interaction of central and peripheral mechanisms in tinnitus. In A. Moller & J. Vernon (Eds.), *Mechanisms of tinnitus* (pp. 181-206). Boston: Allyn & Bacon.

Meikle, M. B., Creedon, T. A., & Griest, S. E. (2004). *Tinnitus archive* (2nd ed.). Retrieved from http://www.tinnitusarchive.org/

Meikle, M. B., & Griest, S. E. (2002). Tinnitus severity and disability: Prospective efforts to develop a core set of measures. In R. Patuzzi (Ed.), *Proceedings of the Seventh International Tinnitus Seminar* (pp. 157-161). Perth: University of Western Australia.

Meikle, M. B., Griest, S. E., Stewart, B. J., & Press, L. S. (1995). Measuring the negative impact of tinnitus: A brief severity index. *Abstracts of the Association for Research in Otolaryngology*, p. 167.

Meikle, M., Henry, J., Abrams, H., Frederick, E., Martin, W., McArdle, R., et al. (2008). Development of the Tinnitus Functional Index (TFI): Part 1. Assembling and testing a preliminary prototype. *Abstracts of the Association for Research in Otolaryngology*, p. 138.

Moller, A. (2007). Tinnitus and pain. *Progress in Brain Research*, *166*, 47-53.

Newman, C. W., Jacobson, G. P., & Spitzer, J. B. (1996). Development of the Tinnitus Handicap Inventory. *Archives of Otolaryngology—Head and Neck Surgery*, *122*, 143-148.

Newman, C. W., & Sandridge, S. A. (2004). Tinnitus questionnaires. In J. B. Snow, Jr. (Ed.), *Tinnitus: Theory and management* (pp. 237-254). Hamilton, Ontario: BC Decker.

Newman, C. W., Sandridge, S. A., & Jacobson, G. P. (1998). Psychometric adequacy of the Tinnitus Handicap Inventory (THI) for evaluating treatment outcome. *Journal of the American Academy of Audiology*, *9*, 153-160.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Piccirillo, J. F. (1994). Outcomes research and otolaryngology. *Otolaryngology—Head and Neck Surgery*, *111*, 764-769.

Relman, A. S. (1988). Assessment and accountability: The third revolution in medical care. *New England Journal of Medicine*, *319*(18), 1220-1222.

Saito, T., Manabe, Y., Shibamori, Y., Noda, I., Yamamoto, T., & Saito, H. (1999). Comparison between matched and self-reported change in tinnitus loudness before and after tinnitus treatment. In J. Hazell (Ed.), *Proceedings of the Sixth International Tinnitus Seminar* (pp. 522-524). London: Tinnitus and Hyperacusis Centre.

Stewart, B. J., & Archbold, P. G. (1992). Nursing intervention studies require outcome measures that are sensitive to change: Part one. *Research in Nursing & Health*, *15*, 477-481.

Stewart, B. J., & Archbold, P. G. (1993). Nursing intervention studies require outcome measures that are sensitive to change: Part two. *Research in Nursing & Health*, *16*, 77-81.

Sweetow, R. W., & Levy, M. C. (1990). Tinnitus severity scaling for diagnostic/therapeutic usage. *Hearing Instruments*, *41*, 44-46.

Tugwell, P., & Boers, M. (1993). Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. *Journal of Rheumatology*, *20*, 555-556.

Turk, D. C. (2002). *The three most important words in health care: Outcomes, OUTCOMES, OUTCOMES*. In R. Patuzzi (Ed.), *Proceedings of the Seventh International Tinnitus Seminar*. Perth: University of Western Australia.

Turk, D. C., Dworkin, R. H., Allen, R. R., Bellamy, N., Brandenburg, N., Carr, D. B., et al. (2003). Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain*, *106*, 337-345.

Turk, D. C., Dworkin, R. H., Burke, L. B., Gershon, R., Rothman, M., & Scott, J. (2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain*, *125*, 208-215.

Turk, D. C., Dworkin, R. H., Revicki, D., Harding, G., Burke, L. B., Cella, D., et al. (in press). Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. *Pain*.

Turk, D. C., & Melzack, R. (Eds.). (2001). *Handbook of pain assessment* (2nd ed.). New York: Guilford.

Tyler, R. S. (2000). Psychoacoustical measurement. In R. S. Tyler (Ed.), *Tinnitus handbook* (pp. 149-179). San Diego: Singular.

van der Heijde, D., Bellamy, N., Calin, A., Dougados, M., Khan, M. A., & van der Linden, S. (1997). Preliminary core sets for endpoints in ankylosing spondylitis. *Journal of Rheumatology*, *24*, 2225-2229.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome questionnaire: Item sensitivity to change. *Journal of Personality Assessment*, *74*, 242-261.

Vernon, J. (1996). Is the claimed tinnitus real and is the claimed cause correct? In G. E. Reich & J. A. Vernon (Eds.), *Proceedings of the Fifth International Tinnitus Seminar* (pp. 395-396). Portland, OR: American Tinnitus Association.

Vernon, J. A., & Meikle, M. B. (1985). Clinical insights into possible physiological mechanisms of tinnitus. In E. Myers (Ed.), *New dimensions in otorhinolaryngology—head & neck surgery* (Vol. 1, pp. 439-446). New York: Elsevier Science Publishers.

Vernon, J. A., & Meikle, M. B. (2003). Tinnitus: Clinical measurement. *Otolaryngology Clinics of North America*, *36*, 293-305.

Wilson, P. H., Henry, J., Bowen, M., & Haralambous, G. (1991). Tinnitus Reaction Questionnaire: Psychometric properties of a measure of distress associated with tinnitus. *Journal of Speech and Hearing Research*, *34*, 197-201.

World Health Organization. (1980). *International classification of impairments, disabilities and handicaps: A manual of classification relating to the consequences of disease*. Geneva: Author.